PSYCHOLOGICAL REPORTS

LABORATORY OF EXPERIMENTAL PSYCHOLOGY

University of Leuven
Department of Psychology
B-3000 Leuven (Belgium)
tel.: +32.16.325965 fax: +32.16.326099
<http://www.psy.kuleuven.ac.be/labexppsy>

W. De Neys, W. Schaeken, & G. d'Ydewalle

Working memory and everyday conditional reasoning: Retrieval
and inhibition of stored counterexamples

January 2003, N° 295

Working Memory and Everyday Conditional Reasoning: Retrieval and Inhibition of Stored Counterexamples

Four experiments examined the contribution of Working Memory (WM) to the retrieval and inhibition of background knowledge about counterexamples (alternatives and disablers, Cummins, 1995) during conditional reasoning. Experiment 1 showed that higher WM-capacity was associated with more efficient counterexample retrieval. Experiment 2 indicated that retrieval was less efficient when WM was burdened by an attention demanding secondary task. Experiment 3 presented a conditional reasoning task with everyday causal conditionals to a group of high and low WM-spans. High spans rejected the logically invalid AC and DA inferences to a larger extent than low spans, while low spans accepted the logically valid MP and MT inferences less frequently than high spans. In Experiment 4, an attention demanding secondary finger tapping task was imposed during the reasoning task. Findings corroborate that WM-resources are used for retrieval of stored counterexamples and that high spans will use WM-resources to inhibit the counterexample activation when the type of counterexample conflicts with the logical validity of the reasoning problem.

The ability to think conditional, if-then, thoughts is considered as one of the cornerstones of our mental equipment. As Edgington (1995, p. 235) puts it "there would not be much point in recognizing that there is a predator in your path unless you also realize that if you don't change direction pretty quickly you will be eaten". Similarly, when someone warns you 'If you don't stop bugging me, I'll beat you' and you want to avoid being beaten up by an angry person, you need to draw a conditional inference.

Given the central role conditional reasoning plays in our causal knowledge system and social interactions it is not surprising that it has become one of the most intensely studied topics in human reasoning research (Evans, Newstead, & Byrne, 1993). Many reasoning theories appeal to the notion of a limited capacity working memory in their explanation of reasoning performance (e.g., Braine & O'Brien, 1998; Johnson-Laird & Byrne, 1991; Rips, 1994). While the proposed reasoning mechanisms differ, the central assumption is that reasoning errors may occur when the capacity of working memory is overburdened.

There is evidence for a general link between working memory capacity and performance in a range of reasoning task (e.g., Barrouillet, 1996; Gilhooly, Logie, & Wynn,

1999; Kyllonen & Christal, 1990): People with higher scores on standard working memory tests tend to draw more logically correct conclusions. A few studies have established this link in the specific case of conditional reasoning (e.g., Barrouillet & Lecas, 1999; Markovits, Doyon, & Simoneau, 2002). Some researchers have even moved beyond a merely correlational approach and showed that burdening working memory with a secondary task, gives rise to conditional reasoning errors (Toms, Morris, & Ward, 1993; Meiser, Klauer, & Naumer, 2001).

While there is some evidence for the involvement of working memory in conditional reasoning, there is an important caveat in the current studies. These studies have almost exclusively (with Markovits et al., 2002 as exception) adopted 'abstract' conditionals of the form 'If square, then circle'. We label these conditionals 'abstract' because people have no prior knowledge about the relation the conditional expresses. In everyday life we typically reason with meaningful and content-rich conditionals (e.g., 'If you put fertilizer on plants, then they grow well'). Here our long-term memory contains prior knowledge about the conditional (e.g., you might think of the fact that in order to grow well the plants also need sunlight) and it is well established that this knowledge has a

massive impact on the inferences people draw (e.g., Staudenmayer, 1975). It was precisely to sidestep this background knowledge effect that studies on the role of working memory in conditional reasoning have explicitly preferred content-lean conditionals (Barrouillet & Lecas, 1999; Meiser et al., 2001). However, the ultimate goal of a psychological reasoning theory is to account for peoples daily life reasoning (e.g., Galotti, 1989; Johnson-Laird, 1983; Johnson-Laird & Byrne, 2002; Oaksford & Chater, 1998). If working memory is assumed to be involved in reasoning, it is crucial to examine its role in reasoning with the typical content-rich conditionals we use in everyday life. The present article starts this examination. We focus on two important functions: The retrieval and inhibition of background knowledge about counterexamples from long-term memory.

Investigations of conditional reasoning typically ask people to asses arguments of the following four kinds (in their abstract form):

Modus Ponens (MP)

If p then q, p, therefore q

Modus Tollens (MT)

If p then q, not q, therefore not p

Denial of the Antecedent (DA)

If p then q, not p, therefore not q

Affirmation of the Consequent (AC)

If p then q, q, therefore p

In standard logic MP and MT are considered valid inferences, while AC and DA inferences are considered fallacies. So, when you are told 'If fertilizer is put on plants, then they grow well' and you receive the information that fertilizer was indeed put on the plants, then logic tells you to accept the conclusion that the plants grow well (an MP inference). Likewise, if you receive the information that the plants do not grow well, you should infer that the plants were not fertilized (an MT inference). On the other hand, logically speaking, from the information that the plants grow well, you should not infer that the plants were fertilized (an AC inference). Likewise, upon knowing that the plants were not fertilized you should reject the conclusion that therefore the plants will not grow well (a DA inference).

Research on the impact of background knowledge about the conditional relation has showed that there are at least two important kinds of information stored in long-term

memory that affect the inferences people draw: Alternative causes and disabling conditions. Both are referred to as 'counterexamples' (Byrne, 1989). Consider for example the conditional

If the brake is depressed, then the car slows down.

This conditional expresses a causal relation between a cause, depressing the brake, specified in the first (the antecedent) part of the conditional and an effect, slowing down, specified in the second (the consequent) part of the conditional. An alternative cause (alternative) is a possible cause that can produce the effect mentioned in the conditional while a disabling condition (disabler) prevents the effect from occurring despite the presence of the cause. For example, possible alternative causes for the conditional are:

Running out of gas, having a flat tire, shifting the gear down, ...

The alternatives make it clear that it is not necessary to depress the brake in order to slow the car down. Other causes are also possible.

Possible disabling conditions are:

A broken brake, accelerating at same time, skid due to road conditions, ...

If such disablers are present, depressing the brake will not result in the slowing down of the car. The disablers make it clear that depressing the brake is not sufficient for the slowing down of the car. Additional conditions have to be fulfilled.

Pioneering studies have examined the impact of counterexample retrieval on conditional reasoning by the explicit presentation of counterexamples (e.g., Byrne, 1989; Romain, Connell, & Braine, 1983). For example, Byrne (1989) found that explicitly mentioning a possible disabler like 'If the library is open, then Ann studies late in the library' for the conditional 'If she has an essay to write, then she studies late in the library' decreased acceptance of MP and MT. Further studies established the importance of the outcome of the search for stored counterexamples. Cummins (1995; see also Cummins, Lubart, Alksnis, & Rist, 1991; Thompson, 1994) manipulated the availability

of possible counterexamples. She adopted conditionals for which pilot work indicated that people could retrieve many (e.g., a conditional with many possible alternatives 'If you study hard, then you pass the exam') or few counterexamples (e.g., a conditional with few possible alternatives 'If you grasp the glass with your bare hands, then your fingerprints are on it'). For conditionals with many alternatives, where successful retrieval was very likely, AC and DA were less accepted than for conditionals with only few possible alternatives. Likewise, MP and MT were less accepted when a conditional had many disablers than when only few were available.

In addition, reasoning performance has been related to individual differences in the efficiency of the counterexample retrieval process (e.g., Janveau-Brennan & Markovits, 1999; De Neys, Schaeken, and d'Ydewalle, 2002). In these studies participants were first presented a generation task where they were asked to generate, in a limited time, as many counterexamples as possible for a set of conditionals. The same participants then received a conditional reasoning task with different conditionals. Janveau-Brennan and Markovits found that the more alternatives one could generate in the generation task, the more AC and DA were rejected in the reasoning task. Likewise De Neys et al. observed that better disabler generation capacity resulted in lower MP and MT acceptance ratings.

In the first two experiments of the present study we test the hypothesis that working memory is crucial for the retrieval of stored counterexamples. Working memory (WM) is often characterized as a hierarchically organized system in which specific storage and maintenance components subserve a central component responsible for the control of information processing (e.g., Baddeley and Hitch, 1974; Cowan, 1995; Engle & Oransky, 1999). The controlling component or 'central executive' is conceived as a limited-capacity system that regulates the allocation of attentional resources. The executive functioning is mediated by the prefrontal cortex (Wickelgren, 1997).

Performance on standard working memory tests is assumed to primarily reflect central executive capacity (Engle, Tuholski, Laughlin, & Conway, 1999; Kane, Bleckley, Conway, & Engle, 2001). In line with this

view, the present investigation of working memory and everyday conditional reasoning focuses on the role of the central executive and not on the storage or slave systems. We also note that in our discussion we treat WM-resources as a domain-free attentional capacity. Thus, WM-resources are hypothesized to be modality a-specific (e.g., we do not distinguish a separate verbal and visual working memory, see Engle, 2002; Engle & Oransky, 1999).

We hypothesized that working memory capacity is important in everyday conditional reasoning because the counterexample retrieval process would require working memory resources. While some forms of memory retrieval are rather automatic and effortless, other forms demand controlled, effortful attention for their proper functioning (Kane & Engle, 2000; Moscovitch, 1995; Rosen & Engle, 1997).

If working memory is involved in counterexample retrieval, we expect to find a relation between the efficiency of the counterexample retrieval process and working memory capacity. This prediction is tested in Experiment 1. Experiment 2 further examines whether working memory is crucial for counterexample retrieval by examining the effect of a concurrent working memory load on counterexample retrieval.

Experiment 3 compares the everyday conditional reasoning performance of individuals with a high (high spans) and low score (low spans) on a measure of working memory capacity. People with a more efficient retrieval process will be more successful at retrieving counterexamples. We know that successful alternative retrieval leads to lower acceptance ratings of the AC and DA inferences. If high spans are more efficient at retrieving alternatives we expect that they will be less inclined to accept AC and DA compared to low spans. Since disabler retrieval results in lower MP and MT acceptance ratings one could also expect that because of the more efficient disabler search high spans will more frequently reject the MP and MT inferences. However, while AC and DA are logical fallacies, MP and MT are logically valid. Rejecting AC and DA is in line with standard, first-order logic, while rejecting MP and MT is not. For people who would approach the logical standards, there would be a dissonance between the tendency to reject MP and MT and the valid status of these inferences.

In the reasoning literature there is some debate about whether people are able to adhere to normative standards such as standard logic in reasoning (e.g., Evans & Over, 1996; Stanovich & West, 2000). Individual difference studies indicate that at least people of high cognitive capacity appear to have a logical, 'decontextualisation' tendency: A basic ability to put background knowledge aside when it conflicts with the logical standards (e.g., Stanovich & West, 2000). In these studies cognitive ability is typically operationalized in terms of scores on general intelligence test which show a strong connection to working memory test performance (see Engle et al., 1999). If high spans would have an elementary notion of logical validity, this should conflict with the retrieval of disablers. In this case we hypothesize that high spans will use their working memory resources to block or inhibit the disabler retrieval.

The inhibition of responses deemed inappropriate is considered as one of the key executive functions (e.g., Baddeley, 1996; Engle et al., 1999; Shallice & Burgess, 1993; Miyake & Shah, 1999). It has also been demonstrated that these inhibition mechanisms can be targeted at memory traces to control retrieval (e.g., Anderson & Bell, 2001; Conway & Engle, 1994; Radvansky, 1999; for a review see Levy & Anderson, 2002).

Markovits and Barrouillet (2002) already put forward the possibility of a disabler inhibition process in conditional reasoning. More general evidence for an inhibitory mechanism in reasoning comes from a neuroimaging study with highly educated participants (mostly graduate students) on syllogistic reasoning (Goel, Buchel, Frith, & Dolan, 2000). In conditions where the logical status of the conclusion conflicted with background knowledge (e.g., a valid but unbelievable conclusion like 'Some of the communists are golfers. All of the golfers are capitalists. Therefore, some of the communists are capitalists.') two regions of the right prefrontal cortex (Brodmann areas 8 and 46/45) were specifically activated. Goel et al. argued that this activation would reflect an inhibitory mechanism that is blocking the impact of background knowledge.

Granted that such a disabler inhibition mechanism operates we can expect to see higher MP and MT acceptance ratings for the

high spans, as compared to the low spans. This prediction was tested in Experiment 3. Experiment 4 presents additional, more direct evidence for the role of working memory in the retrieval and inhibition of counterexamples by examining the effects of a secondary working memory load on reasoning performance.

With reference to the everyday nature of our reasoning task we want to stress two crucial distinctions with more formal reasoning tasks. First, the study adopts meaningful, causal conditionals so that participants have access to relevant background knowledge about alternatives and disablers. In addition, contrary to most conditional reasoning studies participants are not specifically instructed to reason logically (e.g., participants are not instructed to accept the premises as always true or to derive only conclusions that follow necessarily). Although participants are still situated in a laboratory setting this should allow and encourage people to reason as they would in everyday life (Cummins, 1995; see also Galotti, 1989).

Experiment 1

If working memory capacity plays a role in the retrieval of counterexamples from long-term memory, we expect that people with more working memory resources will be better at retrieving counterexamples. In Experiment 1 we therefore tested whether participants' performance on a standard working memory test (Operation span task, see La Pointe & Engle, 1990) was related to the efficiency of the retrieval process as measured by the number of counterexamples they could generate for a set of conditionals in 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

One hundred and four 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.

Material

Counterexample generation task.

Participants were requested to generate counterexamples (alternatives or disablers) for a set of eight causal conditionals (see Appendix A). Item format and instructions were adopted from De Neys et al. (2002) 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 write down as many factors as you can that could make this situation possible.

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 black background. The fixed headers ("Rule:", "Fact:", and "Please write ...") appeared in gray letters, the remaining text in yellow ones. After 30 s the background colored 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., 2002; 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 few possible counterexamples. Within 1.5 min people typically generate about two counterexamples for a 'few' conditional and four counterexamples for a 'many' conditional (see De Neys et al.).

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 shirt', 'taking off sweater', 'taking off coat') would be scored as a single item and needed to be avoided. We also told participants they could give brief responses that only mentioned 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.

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 and De Neys, d'Ydewalle, Schaeken, & Vos, in press for details). Participants saw individual operation-word strings on the monitor of the computer. They read aloud and solved the math 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 $(5 \times 1) - 4 = 2$? BALL
IS $(3 \times 4) - 5 = 8$? 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 then 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 was 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. (2002) was provided to clarify the rating task. To get a grasp of the reliability of the rater's scoring criteria we first asked to indicate whether or not the rater agreed with the previous judgments (e.g., should 'taking shirt off' and 'taking sweater off' be scored as a single variation of 'taking clothes off'?). Judgments agreed on more than 93% of the classifications.

Results and discussion

For all the statistical analyses reported in this study rejection probability was .09. For completeness, we always report the individual estimated p-values for the effects that reached this liberal significance criterion.

Overall, 6.4% of the generated counterexamples were disallowed by the rater. On average participants generated a total number of 18.29 counterexamples ($SD = 4.55$) for the eight conditionals in the generation task (disablers = 18.62, $SD = 5.26$ and alternatives = 17.91, $SD = 3.72$). All participants solved at least 85% of the Ospan operations correctly. Mean Ospan score was 15.94 words recalled correctly ($SD = 8.4$).

A positive correlation between the number of generated counterexamples and working memory capacity, $r = .25$, $n = 104$, $p < .015$, indicated that, as expected, higher

working memory capacity was associated with a more efficient counterexample retrieval process. The pattern was clear for both types of counterexamples (disablers, $r = .24$, $n = 52$, $p < .09$; alternatives, $r = .27$, $n = 52$, $p < .06$).

In order to get a more specific picture we compared generation performance of participants in the upper (high spans) and bottom quartile (low spans) of the working memory capacity distribution. Thirty participants were classified as low spans (Ospan score 10 or lower), while 25 participants were classified as high spans (Ospan score 20 or higher). Approximately half of the low and high spans had generated disablers ($n_{low} = 14$; $n_{high} = 12$) while the other half had generated alternatives ($n_{low} = 16$; $n_{high} = 13$).

Working memory capacity (high or low) and type of generated counterexample (disabler or alternative) were entered as between-subject 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 that generated disablers and alternatives, $F(1, 51) < 1$.

The same ANOVA (span group $\times$ counterexample type) was run on the number of discarded counterexamples. While low spans ($M = 1.73$ errors or 9 % of total generations) tended to make somewhat more errors compared to high spans ($M = 1.32$ or 6% of total generations) none of the factors reached significance, all $F < 1$.

These results establish that people with higher WM-capacity are more efficient at retrieving alternatives and disablers from memory.

Experiment 2

Experiment 1 showed a positive relation between counterexample retrieval and WM-capacity. Higher WM span was associated with the capacity to retrieve more counterexamples (alternatives or disablers). The present experiment examines the causal

nature of this relation. In order to test the crucial role of WM-capacity in counterexample retrieval we studied the impact of a secondary task load on the generation efficiency. If WM-capacity is crucial for the counterexample retrieval, then burdening WM with a secondary task should reduce the efficacy of the retrieval process.

As the secondary task, participants were requested to tap a finger pattern with their non-dominant hand while generating counterexamples. This secondary tapping task was adopted from Kane and Engle (2000) and Moscovitch (1994).

The studies of Kane and Engle (2000) and Moscovitch (1994) indicated that tapping a complex, novel tapping 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 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 control 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 hypothesized 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 explicitly tested.

Furthermore, we also wanted to test whether the expected WM-load effects were different for participants with high and low WM-span. Rosen and Engle (1997) already observed that a dual task affected retrieval performance in a category generation task (e.g., generating instances of the category

'animals') only for high spans. They concluded that in their category generation task low spans relied on an automatic, associative retrieval mechanism. If this is also the case for counterexample retrieval 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) x 2 (counterexample type, between-subjects) x 2 (tapping type, between-subjects) design.

Participants

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

Material

Counterexample generation task.

Besides the fact that we used a different set of conditionals (see Appendix A), the generation task was identical to the one used in Experiment 1. Both for the disablers and alternatives generation task participants generated counterexamples for a set of eight causal conditionals. Half of the conditionals had many possible counterexamples, while the other half had only few. The conditionals were once more selected from the pilot generation studies of De Neys et al. (2002) and Dieussaert et al. (2002). We made sure that the mean number of counterexamples generated for the selected sets of eight 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 QUERTY-keyboard of the second computer.

Procedure

We tested each participant individually after a 2 min break that followed Experiment 1. The first 40 disabler and 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 cutoff 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 cutoff for the 60 s practice trial. Thus, if any one intertap interval was more than 150 ms slower than the established cutoff from the prior practice trial, the computer immediately emitted a 600 ms tone.

Note that contrary to the present procedure, Kane and Engle's (2000) participants did not longer receive 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 effortless and without errors. Therefore, as 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 cutoff 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 center 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 (only for complex tapping). The feedback cutoff time was calculated from the immediately preceding 30 s trial. Thus, mean preceding intertap interval + 150 ms for the complex tapping and mean preceding sequence interval + 600 ms for the cascade tapping. From this point onward participants always tapped with this feedback cutoff 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. We tested the effects of WM-load on retrieval performance by comparing participants retrieval performance in Experiment 1 (no load) and 2 (load).

Participants were assigned to one of the four groups (tapping type x counterexample type) in the experiment. A first control analysis established that the mean Ospan score of the participants in the different groups did not significantly differ, $F(3, 100) = 2.01$, $MSE = 68.47$. We ran a separate ANOVA on the number of retrieved counterexamples for participants that tapped the cascade and complex sequence with the type of generated counterexample as between subject factor. Results are shown in Figure 1.

As expected tapping the habitual cascade sequence did not affect the number of

retrieved alternatives or disablers (effect of load, type of counterexample, and interaction all $F(1,38) < 1$). This control condition established that any 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.

As predicted, tapping the complex sequence did result in a significant decrease in the number of retrieved counterexamples, $F(1, 62) = 79.72$, $MSE = 7.91$, $p < .0001$ ¹. As with the cascade-tapping, complex-tapping participants generated a similar number of alternatives and disablers and the type of counterexample did not affect the WM-load effect, both $F(1, 62) < 1$. Thus, we can conclude that putting a load on WM had a similar impact on the number of generated alternatives and disablers.

Insert Figure 1 Here

The next analysis checked whether the complex tapping effects on counterexample generation differed for high and low spans. Following the criteria of Experiment 1, 19 out of the 64 participants that 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 subject factor showed a main effect of span group, $F(1, 33) = 3.23$, $MSE = 37.58$, $p < .085$. As Figure 2 indicates, both under no load and load conditions high spans generated more counterexamples than low spans. However, when WM was burdened by the complex tapping task the retrieval performance of both span groups declined. Indeed, there

¹ In order to check whether the load findings were not affected by the higher power in the complex (vs. control) 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 effect of complex tapping, $F(1, 38) = 66.06$, $MSE = 6.69$, $p < .001$, and the other effects and interactions were not significant, all $F(1, 38) < 1$.

was no sign of an interaction between span group and WM-load, $F(1, 33) < 1$.

Insert Figure 2 Here

The same 2 (WM-load) x 2 (span group) ANOVA was run on the number of discarded counterexamples. Overall, the working memory load decreased the number of generation errors ($M = 1.31$ errors or 7.8% of total generations discarded under no load vs. $M = .8$ errors or 5.1% of total under load, $F(1, 33) = 8.52$, $MSE = 1.32$, $p < .01$), but this effect was similar for high and low spans, Load x Span interaction, $F(1, 33) = 1.93$, $MSE = 1.32$. As observed in Experiment 1, low spans ($M = 1.41$ errors or 7.8 % of total generations) also tended to make somewhat more errors compared to high spans ($M = 1.32$ or 6% of total generations) but the effect was not significant, $F(1, 33) < 1$.

Tapping task

For the tapping task we analyzed the mean number of correct taps per second across two relevant tapping periods: The "baseline" period present 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 presents the average tapping performance during the 240 s (8 items x 30 s) that participants tapped while generating counterexamples.

Insert Table 1 here

The two top lines in Table 1 present the tapping performance of participants that tapped the cascade or complex tapping sequence. A 2 (period, within-subjects) x 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 in the baseline and counterexample period further showed that concurrent counterexample generation declined complex-tapping performance, $F(1, 102) = 80.19$, $MSE = .15$, $p < .0001$, while

cascade tapping was not affected (period x tapping type interaction, $F(1, 102) = 19.78$, $MSE = .15$, $p < .0001$).

The two bottom lines in Table 1 compare the complex tapping performance of high and low span participants. A 2 (period, within-subjects) x 2 (span group, between-subjects) ANOVA indicated that high spans tapped better than low spans, $F(1, 33) = 3.43$, $MSE = .94$, $p < .075$. The WM-load caused by the concurrent generation of counterexamples resulted in a similar decrease in tapping performance for both high and low spans, $F(1, 33) = 33.39$, $MSE = .15$, $p < .0001$ (Period x Span interaction, $F(1, 33) = 2.01$, $MSE = .15$, $p > .16$). 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 establishes that the retrieval efficiency declines when WM is burdened with an attention demanding task. This implies that WM-capacity is crucial for the process of 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, based on the present findings we conclude 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. According to them 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 both people with low and high working memory spans. Both span groups would use their working memory resources to monitor the automatic retrieval to prevent errors and re-access of previously retrieved category instances. When additional WM-resources are available these will be used for

an active generation of cues to access new instances. The active cue generation will allow a much 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.

We propose a similar mechanism to model the counterexample retrieval. The finding that WM-load also affects low spans' counterexample retrieval is not surprising if one takes into account the different demands of the monitor component in both tasks. Rosen and Engle's participants generated category instances for 10 min, 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 four counterexamples for a conditional. Thus, the monitoring component will be much less demanding in the counterexample retrieval case (e.g., keeping track of four vs. 20 items). Therefore, both high and low spans will be able to use WM 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) 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. While 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.

Experiment 3

In Experiment 3 we compared the performance of people with low and high WM-span in an everyday conditional reasoning task. Retrieving alternatives is known to decrease acceptance of the logical fallacious AC and DA inferences (e.g., Byrne, 1989; Cummins, 1995; De Neys et al., 2002; Janveau-Brennan & Markovits, 1999; Quinn & Markovits, 1998). Our first experiments showed that high spans are better at retrieving alternatives. Therefore, we expect that high spans will be less inclined to accept AC and DA compared to low spans.

Retrieving disablers is known to decrease acceptance of the valid MP and MT inferences (e.g., Byrne, 1989; Cummins, 1995; De Neys et al., 2002; De Neys, Schaeken, & d'Ydewalle, in press). If the assumption that high spans use their WM-resources to inhibit disabler access is correct, we should observe that high spans accept MP and MT more than the low spans who allocate their WM-resources primarily to retrieval. Thus, the central prediction is an interaction between WM-capacity and inference type: While high spans should tend to accept MP and MT to a larger extent, low spans should show higher AC and DA acceptance ratings.

As in Cummins (1995) the number of available counterexamples for the conditionals in our reasoning task varied systematically (i.e., half of the conditionals had many/few possible alternatives/disablers). This manipulation is important because under the assumption that high spans have some kind of basic logic notion it might be suggested that they will use this notion to reject the fallacious AC and DA inferences. Thus, one might argue that a lower AC and DA acceptance rating for high spans does not result from counterexample retrieval but rather from a purely abstract, content-free reasoning ability. If high spans' inference acceptance would be solely determined by their logical knowledge, the availability of alternatives should not affect the conclusions. We expect that the alternative search process is crucial for both high and low spans. Successful retrieval will be more likely when many (vs. few) counterexamples are stored. Thus, if retrieval of alternatives determines the inferences people draw, inference acceptance should be affected by the number of available alternatives. Contrary to

the abstract reasoning ability hypothesis, we therefore predict that both low and high spans will be affected by the number of alternatives factor. We also predict that both span groups will be affected by the number of disablers of the conditionals. Under the assumption that high spans' disabler inhibition is an attention demanding process we can expect that the more instances that need to be inhibited, the less successful the inhibition will be. Therefore, although high spans should overall show higher MP and MT acceptance (vs. low spans), even the high span group is expected to show an impact of the number of disablers.

One should note that our reasoning experiments (see De Neys et al., in press, 2002; see also Cummins, 1995) use a graded scale to measure inference acceptance. Thus, when we refer to acceptance and rejection of an inference this should be interpreted relative to the rating scale (i.e., rejection points to lower acceptance ratings).

Method

Participants screening for working memory capacity

We screened participants for working memory capacity using a version of the Ospan task (La Pointe & Engle, 1990) adapted for group testing (Gospan, for details see De Neys, d'Ydewalle et al. (in press). The main adaptation was that we first presented the operation from an operation-word pair on screen (e.g., 'IS $(4/2) - 1 = 5$?'). Participants read the operation silently and pressed a key to indicate whether the answer was correct or not. Responses and response latencies were recorded. After the participant had typed down the response, the corresponding word (e.g., 'BALL') from the operation-word string was presented for 800 ms. As in the standard Ospan three sets of each length (from two to six operation-word pairs) were tested and set size varied in the same randomly chosen order for each participant. The Gospan score was the sum of the recalled words for all sets recalled completely and in correct order.

Participants were tested in groups of 38 to 48 at the same time in a large computer room with an individual booth for every participant. Participants who made more than 15% math errors or whose mean operation response latencies deviated by more than 2.5

standard deviations of the sample mean were discarded. The internal reliability coefficient alpha for the Gospan was .74 and the corrected correlation between standard Ospan and Gospan score reached .70 (see De Neys, d'Ydewalle et al., in press).

Participants

Fifty-two first year psychology students from the University of Leuven, Belgium, participated in Experiment 3 in return for psychology course credit or 5 euro. These participants were identified from a larger pool of 426 first year psychology students who had participated in the Gospan task: Twenty-six participants were selected from the top quartile of the distribution ("high spans") and 26 were selected from the bottom quartile ("low spans"). Between 45 and 90 days intervened between a given individual's participation in the Gospan task and the reasoning task. None of the participants had received any training in formal logic.

Material

Eight causal conditionals from the generation study of De Neys et al. (2002) were selected for the reasoning task (see Appendix B). The conditionals were selected so that the number of available counterexamples constituted a 2 (few/many) x 2 (alternatives/disablers) design with two items per cell. The eight conditionals were embedded in the four (MP, DA, MT, and DA) inference types, producing a total of 32 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:

Rule: If Jenny turns on the air conditioner, then she feels cool

Fact: Jenny turns on the air conditioner

Conclusion: Jenny feels cool

Given this rule and this fact, give your evaluation of the conclusion:

-1-----2-----3-----4-----5-----6-----7-							
Very	Sure	Somewhat	I	Somewhat	Sure	Very	
sure		sure	I	sure		sure	
			I				
That I CANNOT draw this conclusion				That I CAN draw this conclusion			

Each of the 32 arguments was presented in this way. The premises and conclusion were presented in yellow. The remaining text appeared in white on a black background.

Procedure

All participants were tested individually for the reasoning task. Reasoning task instructions 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. Instructions stated that there were no time limits. The experimental session was preceded by one practice trial.

Participants said out loud the number reflecting their decision. The experimenter typed this number down on a keyboard connected to the computer running the reasoning task. Following the key press participants saw the text "NEXT ITEM" (gray letters on black background) for 800 ms after which the next item was presented.

We constructed four sets of eight inferences each. All eight inferences in a set were based on different conditionals. There were two sets with four MP and four AC inferences in each set and another two sets with four MT and four DA inferences in each set. The order of presentation of the inferences within a set was random. The conditionals for the four inferences of the same type in a set were taken from the four different cells within the 2 (few/many) x 2 (alternatives/disablers) design that the conditionals constituted. Half of the participants received the sets in the order MP/AC, MT/DA, MP/AC, and MT/DA. For the other half the sets were presented in the reversed order MT/DA, MP/AC, MT/DA, and MP/AC. After two sets (i.e., 16 inferences) were evaluated, item presentation was paused until participants decided to continue.

As we pointed out, the task instructions did not mention to accept the premises as true or to endorse conclusions that

follow necessarily. Instead participants were told they could evaluate the conclusions by the criteria they personally judged relevant (see Cummins, 1995).

Results

Each participant evaluated inferences based on two different conditionals within each cell of the 2 (number of alternatives) x 2 (number of disablers) x 4 (inference type) cell of the design. The mean of these two observations was calculated. These means were subjected to a 2 (span group) x 2 (number of alternatives) x 2 (number of disablers) x 4 (inference type) mixed model ANOVA with span group as between-subjects factor and number of alternatives, number of disablers, and inference type as within-subjects factor. The within subject part of the design is a replication of Cummins (1995) and De Neys et al. (2002). The primary foci of the present study are the interactions with the span group factor.

Main effects involving repeated measures with more than two levels were analyzed with multivariate ANOVA tests.

As Figure 3 shows there was an interaction between inference type and span group, $Rao R(3, 48) = 2.64, p < .06$. As predicted, high spans accepted the MP and MT inferences more than low spans, while the low spans accepted AC and DA more than high spans, $F(1, 50) = 7.7, MSE = 3.47, p < .009$.

Insert Figure 3 here

The results for the number of alternatives, number of disablers, and inference type factors completely replicated the standard findings of Cummins (1995) and De Neys et al. (2002): There was an effect of inference type, $Rao R(3, 48) = 29.86, p < .001$, and this effect interacted with number of disablers, $Rao R(3, 48) = 22.39, p < .001$, and number of alternatives, $Rao R(3, 48) = 23.98, p < .001$. The number of disablers primarily affected MP and MT acceptance ratings; for conditionals with many disablers MP, $F(1, 50) = 26.9, MSE = .4, p < .001$, and MT, $F(1, 50) = 10.43, MSE = 1.2, p < .003$, were accepted less than for conditionals with few disablers. The number of alternatives primarily affected AC and DA

acceptance ratings; AC and DA were accepted more when there were only few possible alternatives than when there were many of them, $F(1, 50) = 125.29$, $MSE = 1.1$, $p < .001$, and $F(1, 50) = 94.38$, $MSE = 1.16$, $p < .001$. As De Neys et al. we also observed an impact of alternatives on MP, $F(1, 50) = 24.65$, $MSE = .36$, $p < .001$, and MT acceptance, $F(1, 50) = 11.81$, $MSE = .9$, $p < .005$, and of disablers on AC, $F(1, 50) = 30.78$, $MSE = .53$, $p < .001$, and DA, $F(1, 50) = 18.05$, $MSE = .96$, $p < .001$, acceptance.

More important is the question whether the effects of number of alternatives and disablers interacted with span group. Figure 4 illustrates the main findings (see Appendix C for a complete overview).

High and low spans were equally affected by the number of alternatives; the Span x Number of Alternatives, $F(1, 50) = 2.35$, $p > .13$, and the Span x Number of Alternatives x Inference Type interactions were not significant, $Rao R(3, 48) < 1$. More specifically, neither on AC, $F(1, 50) < 1$, nor on DA, $F(1, 50) < 1$, span group affected the effect of number of alternatives.

Insert Figure 4 here

The effect of the number of disablers did not interact with span group, $F(1, 50) > 1$. However, we did observe a Span x Number of Disablers x Inference Type interaction, $Rao R(3, 48) = 2.48$, $p < .08$. The effect of the number of disablers on MP and MT seemed somewhat smaller for the high spans. The interaction even reached significance for the MP inference, $F(1, 50) = 3.52$, $MSE = .4$, $p < .07$. However, even on MP the number of available disablers affected both low, $F(1, 50) = 24.94$, $MSE = .4$, $p < .001$, and high spans, $F(1, 50) = 5.48$, $MSE = .4$, $p < .025$.

Discussion

Results of Experiment 3 are in line with the hypothesized role of working memory in conditional reasoning: The lower AC and DA acceptance ratings for high spans support the hypothesis that WM-capacity mediates the retrieval of counterexamples during reasoning. The higher WM-capacity is, the more efficient the search process is, and the less AC and DA

will be accepted. The lower MP and MT acceptance ratings for high spans, despite their higher retrieval capacity, indicate that high spans manage to inhibit the impact of disablers on inference acceptance.

Both high and low spans' inference acceptance was affected by the number of possible alternatives. These findings establish that even for high spans the retrieval of alternatives is a crucial factor in the inference evaluation.

The number of disablers of a conditional also affected both span groups. This indicates that high spans' disabler inhibition is not complete. Remember we proposed that the counterexample retrieval starts with an automatic spreading of activation. Except in specific cases this automatic retrieval process would not be very successful for causal conditionals. The specific cases will be counterexamples that are very strongly associated with the conditional. The strength of association between a counterexample and a conditional has been shown to affect successful retrieval (De Neys et al., in press; Quinn & Markovits, 1998). Conditionals with many counterexamples have typically also strongly associated counterexamples (De Neys et al., 2002, Experiment 1). Thus, more instances will have to be inhibited for the conditionals with many possible disablers. Therefore the inhibition will be less successful; occasionally a disabler might "slip through". Consequently, although MP and MT acceptance will overall be higher for high spans, even high spans' MP and MT acceptance will be affected by the number of available disablers.

The present results are supported by a recent experiment of Markovits, Doyon, and Simoneau (2002), conducted independently from our study. While Markovits et al. did not examine the crucial effect of the number of counterexamples, they did observe that the higher a participant's score on a WM-task, the less frequently AC and DA inferences were accepted and the more frequently MP inferences were accepted (no relation was observed for MT). Although the results may have been affected by the fact that Markovits et al. explicitly instructed participants to reason logically, the data pattern does point to the generality of the present findings.

We mentioned in the introduction that De Neys et al. (2002, Experiment 3) found that

a higher efficiency of the disabler retrieval process is associated with lower MP and MT acceptance. Given that high spans are better at retrieving disablers, these findings might seem to contradict the presently observed stronger tendency to accept MP and MT for high spans. However, De Neys et al.'s sample consisted of 40 randomly selected participants. We only claim that the inhibition occurs for people with the highest cognitive abilities. Therefore, the present study specifically selected participants from the top quartile of first year psychology students' WM-capacity distribution. Thus, when the top WM-levels are excluded (or are small in number, as was probably the case in De Neys et al.) we would indeed expect that higher WM-capacity (and thus better retrieval) results in lower MP and MT acceptance.

While the results of Experiment 3 support the hypothesized role of working memory in the retrieval and inhibition of counterexamples, the findings are not decisive. Additional evidence is required. Experiment 4 presents a more direct test of the hypothesis.

Experiment 4

If working memory resources are used for retrieval and inhibition of counterexamples during reasoning, putting a load on working memory should interfere with the proper functioning of these processes. We hypothesized that low spans primarily allocate their WM-resources to retrieval. Under load conditions retrieving counterexamples will be less efficient and thus successful retrieval of alternatives and disablers will be less likely. Since successful alternative retrieval will decrease AC and DA acceptance, while disabler retrieval decreases acceptance of MP and MT, we predict that under load conditions (due to the less efficient retrieval process) low spans' acceptance of the four different inferences will increase. Since high spans will also have difficulties in retrieving alternatives under load conditions we expect that the secondary task will also increase high spans' level of AC and DA acceptance.

The inhibition hypothesis states that high spans are using WM-resources to prevent the retrieval of disablers. Since the inhibition process will be WM-resource demanding, the inhibition should be less successful under load. Therefore, in contrast to AC and DA acceptance, we predict that a WM-load will

tend to decrease high spans' MP and MT acceptance. Thus, for high spans the load effects on AC/DA and MP/MT should interact.

Special care was taken to select an appropriate secondary task. One should note that before reasoners can start retrieving or inhibiting counterexamples they have to read and mentally represent the premises of the inference problem first. Such reading or comprehension processes will also demand WM-capacity (e.g., Just, Carpenter, & Keller, 1996). We wanted a secondary task that would interfere with retrieval and inhibition of counterexamples but that would leave the more elementary representational processes unaffected. Indeed, if the secondary task would be so demanding that participants would not be able to represent the premises, the findings would not be very informative (e.g., the load would probably only cause a general inference rejection).

While complex tapping decreased the retrieval performance in Experiment 2, participants were still able to generate some counterexamples under load. This would not be possible if complex tapping prevented participants to process the conditional and factual information of a generation task item. The information in our generation task items closely resembled the information presented in a conditional inference problem. We therefore decided to use the complex sequence tapping task of Experiment 2 as the secondary task in the present experiment.

Method

Design

We selected two new groups of high and low spans for the present experiment. Participants were presented the reasoning task of Experiment 3 while working memory was burdened with an attention demanding secondary task. The performance of the high and low spans in Experiment 3 served as a baseline for the effect of introducing the WM-load.

Participants

Thirty-three first year psychology students from the University of Leuven, Belgium, participated in the experiment in return for course credit or 5 euro. These

participants were identified from the same pool of screened students as the participants in Experiment 3. None of the participants in Experiment 4 had participated in Experiment 3. Seventeen participants were selected from the top quartile of the Gospan-distribution ("high spans") and 16 were selected from the bottom quartile ("low spans"). Between 102 and 133 days intervened between a given individual's participation in the Gospan screening task and participation in Experiment 4. None of the participants had received any training in formal logic.

Material

Participants were presented the same reasoning task with the same procedure as in Experiment 3. A program executed by a second computer collected the finger tapping data. Participants tapped on the "V", "B", "N", and "M" on the (Querty-) keyboard of the second computer

Procedure

Each participant was tested individually. All participants started with the complex (i.e., index-ring-middle-pinkie) tapping practice trials as in Experiment 2. The procedure was similar to the one used in Experiment 2 for the complex tapping condition. Thus, three 30 s practice trials with accuracy feedback preceded a 60 s and 30 s trial with both accuracy and response time feedback.

The response time feedback differed slightly from Experiment 2. Participants in Experiment 2 sometimes noted that when they wanted to get back in pace after an occasional slowing down, they needed a few taps to do this. The successive response time error tones were then perceived as somewhat distracting. Therefore, in the present experiment the response time feedback tone was only given for one of the fingers (i.e., pinkie) in the sequence.

After the final 30 s practice trial participants received the instructions for the reasoning task. The experimenter explained that the primary job in the upcoming generation task was to maintain practice tapping speeds throughout while trying to evaluate the conclusions.

The reasoning task began with a "BEGIN TAPPING" instruction screen. This "baseline tapping" signal remained onscreen for 15 s, during which participants tapped with response time and accuracy feedback.

Following the 15 s baseline-tapping a signal ("NEXT ITEM", presented for 1 s on a blue background) announced the beginning of the reasoning task. The experimenter typed down participants' oral responses on a keyboard connected to the computer running the reasoning task. The computer program kept also track of the mean time elapsed between the presentation of an item and the experimenter's key press. Both in Experiment 3 and 4 we used this time record as a raw measure of participants' mean inference latency. We only recorded the timing to have a general indication of the secondary task impact on the time participants needed for an inference. For this goal the less reliable nature of the present procedure is not problematic. For a specific study on inference latencies in a similar reasoning task we refer to De Neys et al. (2002, Experiment 2).

Following the key press participants saw the text "NEXT ITEM" (gray letters on black background) for 800 ms after which the next item was presented. After each set of eight inferences the reasoning task was paused. As in Experiment 3 the set order was reversed for half of the participants in each span group (MP/AC, MT/DA, MP/AC, and MT/DA vs. MT/DA, MP/AC, MT/DA, and MP/AC) and the order of presentation of the inferences within a set was random.

Participants continuously tapped the sequence until the reasoning task was paused. As in Experiment 3 participants could take as much time for every inference as they wanted. After the break participants started with 15 s baseline-tapping after which the warning signal announced the presentation of the next set of inferences.

Results

Reasoning Task

We tested the effects of WM-load on inference performance by comparing inference acceptance of the high and low spans in Experiment 3 (no-load) and 4 (load). We calculated participants mean inference acceptance for every inference type. This

resulted in a 2 (load) x 2 (span group) x 4 (inference type) design with load and span group as between-subject factors and inference type as within-subject factor. The Span group x Inference Type part of the design was already tested in Experiment 3. Here we focus on the effect of the Load factor.

Figure 5 shows the mean acceptance of the four different inference types for high and low spans under no load and load conditions.

Insert Figure 5 here

We tested our hypotheses with planned contrasts. For low spans there was a main effect of WM-load. As Figure 5 clearly shows, low spans tended to accept all inferences more when WM was burdened with a secondary task, $F(1, 81) = 5.32$, $MSE = 10.22$, $p < .025$. For high spans this was not the case. Indeed, the effect of the WM-load interacted with span group, $F(1, 81) = 2.96$, $MSE = 10.22$, $p < .09$. As expected, for high spans the load effects on AC and DA on the one hand and MP and MT on the other hand differed significantly, $F(1, 81) = 3.74$, $MSE = 3.17$, $p < .06$. As Figure 5 indicates, the WM-load tended to increase AC and DA acceptance while it decreased MP and MT acceptance.

The increase of AC and DA acceptance ratings under load tended to be somewhat larger for the low than for the high spans but the effect did not reach significance, both $F(1, 81) < 1$. The load impact on MP, $F(1, 81) = 5.9$, $p < .02$, and MT, $F(1, 81) = 5.28$, $p < .025$, did differ for low and high spans. While the low spans accepted MP and MT more under load, burdening WM resulted in lower MP and MT acceptance for the high spans.

We had no specific expectations about the impact of the WM-load on the effect of the number of alternatives and disablers. For completeness we entered the number of alternatives and disablers as within-subject factors in the design. This resulted in a 2 (load) x 2 (span group) x 2 (number of alternatives) x 2 (number of disablers) x 4 (inference type) mixed model factorial. An ANOVA showed that besides a significant Load x Span x Number of Alternatives interaction, $F(1, 81) = 4.22$, $MSE = 1.26$, $p < .045$, the load factor did not affect any other factor or interaction of

factors in the design (see Appendix C for the raw data). The significant interaction seemed to be caused by the fact that while high spans' inference acceptance under load increased both for conditionals with many and few alternatives, low spans' increase was more pronounced for the many alternative conditionals. This makes good sense. Under no load conditions successful retrieval for conditionals with few alternatives will already be unlikely for the low spans. Thus, an additional WM-load will not make much difference here. Consequently, the load effect for low spans will be stronger for the many alternatives conditionals.

Inference latencies

The time that elapsed between the presentation of an item and the experimenter's key press after participants had evaluated the inference was used as a measure of participants' inference latency. The mean inference latencies were subjected to a 2 (load) x 2 (span group) between-subjects ANOVA. Results showed that both high and low spans needed about 2.5 s longer to evaluate an inference when concurrently tapping the finger pattern; mean inference latency was 8.42 s ($SD = 1.79$) under no load and 11.05 s ($SD = 2.91$) under load, $F(1, 81) = 26.34$, $MSE = 5.31$, $p < .0001$, with no effect of Span, $F(1, 81) < 1$ or Span x Load interaction, $F(1, 81) < 1$.

Tapping Task

For the tapping task we analyzed the mean number of correct taps per second across two relevant tapping periods: The "baseline" period present the average tapping performance during the four 15 s periods that preceded the presentation of each inference set. The "reasoning" period refers to the average tapping performance during the periods between presentation of an inference and the participant's evaluation response. Table 2 present the results.

Insert Table 2 here

A 2 (period, within-subjects) x 2 (span group, between-subjects) ANOVA indicated that tapping performance decreased when

participants were reasoning, $F(1, 31) = 40.21$, $MSE = .09$, $p < .0001$. The decrease in tapping performance was similar for high and low spans, period $\times$ span interaction, $F(1, 33) < 1$. High spans tended to perform somewhat better than low spans but the effect did not reach significance, $F(1, 31) < 1$. The data thus establish that high and low spans were not differentially trading-off reasoning and tapping performance.

Discussion

As predicted, low spans' acceptance of all four inference types increased when working memory was burdened by the tapping task. Experiment 2 already showed that counterexample retrieval was less efficient when the tapping task demanded WM-resources. The present experiment clearly established the link between the decreased search efficiency and inference acceptance. This supports the hypothesis that WM-capacity is crucial for the retrieval of counterexamples in everyday reasoning.

For high spans the load effect interacted with the type of inference. The working memory load tended to increase AC and DA acceptance, as with low spans, but in contrast to the low spans, MP and MT acceptance tended to decrease under load. This pattern corroborates the hypothesis that high spans are using their working memory to inhibit retrieved disablers. The inhibition process explains the higher MP and MT acceptance for high spans in the absence of a WM-load. However, since the inhibition is resource demanding, it will be less efficient under load. Therefore, automatically activated disablers that are otherwise inhibited will decrease MP and MT acceptance. Low spans on the other hand allocate their working memory resources primordially at retrieval. When this retrieval becomes less likely under load, MP and MT will be more accepted.

Interestingly, even low spans showed a higher DA and MT acceptance under load. In contrast to AC and MP, the DA and MT inferences involve negations. Therefore, DA and MT are typically labeled more complex inferences than AC and MP (Braine & O'Brien, 1998; Johnson-Laird & Byrne, 1991; Oaksford, Chater, & Larkin, 2000; Schroyens, Schaeken, & d'Ydewalle, 2001). Markovits and Barrouillet (2002) for example have

argued that accepting MT and DA requires that people retrieve an instance of a 'complementary class' from memory. Such a complementary class refers to cases in which both the relationship and the events concerned are complementary to those specified in the original conditional (i.e., cases where events different from p are related to not- q). For example, the complementary class for the conditional 'If it rains then the streets get wet' would be composed of related events such as 'If the sun shines, the streets are dry' or 'If it is only cloudy, the streets are dry'.

The fact that low spans' MT and DA acceptance increased under load indicates that they could still retrieve a complementary class example. One could suggest this finding shows that retrieval of a complementary class example is rather automatic so that it is not affected by the WM-load. However, this would conflict with the general contention that processing the negation in the categorical premise of the DA and MT problem requires additional processing capacity (e.g., De Neys et al., 2002; Johnson-Laird & Byrne, 1991; Oaksford & Chater, 2001; Schroyens et al., 2001; Schroyens, Schaeken, Fias, & d'Ydewalle, 2000). Markovits and Barrouillet (2002) however suggest an explanation in terms of priority of retrieval instead of demands of retrieval. Markovits and Barrouillet state that when reasoners are confronted with an MT or DA problem they will first search for instances of the complementary class before alternatives or disablers are retrieved (see, also Schroyens et al., 2001, for a parameterized model that captures this processing assumption). Consistent with this claim the increased acceptance might point to the priority of the complementary class search. Because of the priority, the few WM-resources that are still available under load would be primordially allocated to retrieval of the complementary example. Therefore, retrieval of a complementary instance could still be successful, but additional disabler or alternative retrieval would become very unlikely. Consequently, DA and MT acceptance should increase under load.

To our knowledge the present experiment is the first one that used a dual task methodology to study reasoning with meaningful, realistic conditionals. The findings present an interesting extension of previous

dual task studies with abstract material (e.g., Toms et al., 1993; Meiser et al., 2001). These studies typically found that burdening working memory gave rise to reasoning errors (e.g., a higher rejection of the valid MT under load in Toms et al.). This supported the general contention of reasoning theories like mental logic and mental models that capacity limitations in working memory are a major source of fallacious reasoning. Traditional reasoning theories have focused on the role of WM in the manipulation and storage of the basic mental representations (be it mental rules or mental models) of a reasoning problem. We deliberately selected a secondary task that would not interfere with these basic representational processes. This allowed a more subtle examination of the WM contribution to everyday conditional reasoning. We observed for example that low spans' MP and MT acceptance increased under load. Remember that MP and MT are both logically valid. Thus, working memory limitations resulted in a better logical performance here. Despite their larger pool of WM-resources, high spans did not show the same effect. To explain these findings reasoning theories will need to take count of the role of working memory in the retrieval and inhibition of background knowledge.

General Discussion

The present study examined the role of working memory (WM) capacity in everyday life conditional reasoning. We reported four experiments which are the first ones in the field to introduce a dual task methodology. Over these studies a stable pattern emerged that establishes the central role of WM-capacity. We focused on two crucial functions: Retrieval and inhibition of counterexamples stored in long-term memory. While some forms of memory retrieval are rather automatic, other forms demand WM-resources for their proper functioning. Our experiments clearly indicate that counterexample retrieval during everyday conditional reasoning is of the latter form. Experiment 1 and 2 showed that WM-capacity contributes to the efficiency of the counterexample retrieval: People with higher WM-capacity performed better in a counterexample generation task and performance decreased when WM was

burdened by an attention demanding tapping task.

Experiment 3 compared the performance of a group of people classified as high and low spans in an everyday conditional reasoning task. Successful retrieval of an alternative during conditional reasoning decreases acceptance of the AC and DA inferences, while disabler retrieval decreases MP and MT acceptance. Consistent with the predictions, low spans less efficient alternative retrieval resulted in higher acceptance ratings for the AC and DA inferences. Experiment 4 showed that making counterexample retrieval less likely by burdening WM led to a higher acceptance of every inference type for the low spans. Thereby, the availability of WM-resources for counterexample retrieval determines the kind of inferences people draw.

In contrast with AC and DA, MP and MT are logically valid inferences. Based on Stanovich and West's (2000) work on individual differences in reasoning we hypothesized that people of high cognitive ability (for example people in the upper regions of the WM-capacity distribution) would have a basic notion of logical validity. Since disablers lead to rejection of MP and MT, this logic notion should conflict with the disabler retrieval. We hypothesized that high spans use WM-resources to inhibit activated disablers. Consistent with this claim high spans in Experiment 3 showed higher MP and MT acceptance ratings than low spans. If high spans are indeed using WM-resources to inhibit disabler retrieval in a conditional reasoning task, the efficiency of the inhibition should be affected by an attention demanding secondary task. Results of Experiment 4 corroborated the inhibition hypothesis.

The present study contributes to recent research that aims to characterize and model the background knowledge search process during (everyday) conditional reasoning (e.g., De Neys et al., in press, 2002; De Neys, Schaeken, & d'Ydewalle, 2003; Markovits & Barrouillet, 2002). Rosen and Engle's (1997) memory retrieval model could be extended to sketch the elementary components of the counterexample retrieval process. We summarize the components once more below. The findings have also implications for broader issues in the reasoning domain. We will point to the relevance of logic in everyday life, and

elaborate on issues related to dual processing theories of reasoning.

Components of counterexample retrieval

It is assumed that the retrieval process starts with an automatic spreading of activation from a retrieval cue. In a conditional reasoning task this retrieval cue will be the conditional and the categorical premise (e.g., Markovits & Barrouillet, 2002). More precisely, the cue would be the mental representation of these premises stored in working memory. The spreading of activation requires little in the way of executive attention and this component is important for both high and low spans. Both span groups then use their working memory resources to monitor the automatic retrieval to prevent errors and re-access of previously retrieved counterexamples. Finally, available WM-resources will be used for an active generation of cues to access new instances.

The active cue generation will allow for a much more efficient retrieval than the passive spreading of activation. Educated, adult reasoners (e.g., undergraduate students) will typically use WM-resources for active cue generation. The more resources that are available, the more successful the retrieval will be.

This mechanism can be used for the retrieval of both alternatives and disablers. When it concerns retrieving disablers however, people from the top of the WM-capacity distribution will not use their WM-resources for active cue generation but for an inhibition of automatically activated disablers.

Note that the present model characterizes the inhibition process as targeted at preventing the retrieval of disablers. However, this does not imply that the disabler inhibition cannot occur at a later stage in the reasoning process. The inhibition might also contribute to the discarding of a disabler after successful retrieval rather than to the prevention of the retrieval per se. We make no strong claims about the exact locus of the inhibition phenomenon. The crucial point is that the inhibition will prevent that the disablers have their full impact on the reasoning process.

Currently, reasoning theories mainly focus on a specification of how retrieved counterexamples affect the reasoning process (e.g., Byrne, Espino, & Santamaria, 1999;

Johnson-Laird & Byrne, 2002; Oaksford & Chater, 1998; Politzer, in press; Thompson, 2000). Explaining how a reasoning problem is represented and how additional information alters these representations is of course a fundamental component of a reasoning theory. However, the equally important 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 WM-based specification of the retrieval process will contribute to filling this 'knowledge' gap.

Logic in everyday life

The evidence for the high spans' disabler inhibition indicates that people with high cognitive abilities have some basic notion of logical validity or its subsidiary assumptions. High spans apparently adhere to the basic principle in first-order conditional logic that the truth of the antecedent implies the truth of the consequent. This basic logical notion would be the direct cause of the disabler inhibition. This does not mean that high spans intuitively master the propositional calculus or that they would reason by applying formal, logical rules. The results clearly showed that even high spans' inference acceptance ratings were affected by the number of available alternatives and disablers. If high spans would be using an abstract logical 'database' to reason, such content mediation would not be expected.

What the data point to is a minimal notion of the logical fact that a conditional rule excludes the possibility that the consequent does not occur when the antecedent occurs. Previous demonstrations of peoples ability to adhere to the normative standard of first-order logic have always explicitly instructed people to adhere to this norm (Stanovich and West, 1998; Markovits et al., 2002; see also Evans & Over, 1996, Chapter 6). Although this shows that some people are able to reason in accordance with the normative standard when they are properly instructed, it does not show that this ability has something to do with reasoning in everyday life (when the norm is not provided explicitly). This critique is not applicable to the present study. Participants were not instructed to reason logically and none of them had received any training in

formal logic whatsoever. The present findings thus question the claim of a number of authors that standard logic has no bearing on everyday human reasoning (e.g., Harman, 1986; Oaksford & Chater, 1998). The disabler inhibition phenomenon indicates that for some people, be it in a minimal form, it has.

Conscious and unconscious reasoning?

Some theorists have distinguished two types of reasoning systems (e.g., Evans & Over, 1996; Sloman, 1996; Stanovich & West, 2000; see Osman, 2002 for a review). In general the first system (System-1) is characterized by unconscious and automatic processing, while the second systems is characterized by conscious and controllable processing. System-1 processes would tend toward an automatic contextualisation of a problem with prior knowledge. System-2 processes on the other hand decontextualize a problem and allow reasoning according to normative standards. Thereby, System-2 would serve as an override system for the output provided by System-1. Stanovich and West (2000) suggested that System-2 thinking would be characteristic for those higher in cognitive ability.

While the evidence for the disabler inhibition by the high spans corroborates the general claim of an override function, the present findings also point to a problematic feature of the dual process framework. Klaczynski (2001) already argued that theoretical accounts of the two systems are underspecified (also see, Schroyens, Schaeken, & Handley, in press). Osman (2002) accordingly argued that the characterization of the two systems (e.g. automatic, unconscious vs. controlled, conscious processing) is inaccurate. Consistent with this critique, our dual task experiments showed that retrieval of counterexamples during reasoning (thus a System-1 process) is certainly not automatic in nature but requires WM-resources or controlled processing. This questions the characterization of System-1 processes as automatic and effortless.

Conclusion

The present study indicates that WM-capacity plays a crucial role when people reason with conditionals for which they have

access to relevant background knowledge. By the mediation of counterexample retrieval and inhibition, WM-capacity has a profound impact on the inferences people draw in daily life reasoning.

Acknowledgements

Preparation of the manuscript was supported by grants from the Fund for Scientific Research-Flanders (FWO-Vlaanderen) and the Belgian Program on Interuniversity Poles of Attraction, Convention number P4/19. We would like to thank Walter Schroyens for his comments on an earlier draft of the manuscript.

References

- Anderson, M. C., & Bell, T. (2001). Forgetting our facts: The role of inhibitory processes in the loss of propositional knowledge. *Journal of Experimental Psychology: General*, 130, 544-570.
- Baddeley, A. D. (1996). Exploring the central executive. *Quarterly Journal of Experimental Psychology*, 49A, 5-28.
- Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. Bower (Ed.), *Recent advances in learning and motivation* (Vol. 8, pp. 47-90). New York: Academic Press.
- Barrouillet, P. (1996). Transitive inferences from set-inclusion relations and working memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 22, 1408-1422.
- Barrouillet, P., & Lecas, J. F. (1999). Mental models in conditional reasoning and working memory. *Thinking and Reasoning*, 5, 289-302.
- Barsalou, L. W. (1983). Ad hoc categories. *Memory and Cognition*, 11, 211-227.
- Braine, M. D. S., & O'Brien, D. P. (Eds.). (1998). *Mental logic*. Mahwah, NJ: Lawrence 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 and Language*, 40, 347-373.
- 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.

- Cummins, D. D. (1995). Naive theories and causal deduction. *Memory and Cognition*, 23, 646-658.
- Cummins, D. D., Lubart, T., Alksnis, O., & Rist, R. (1991). Conditional reasoning and causation. *Memory and Cognition*, 19, 274-282.
- De Neys, W., d'Ydewalle, G., Schaeken, W., & Vos, G. (in press). A Dutch, computerized, and group administrable adaptation of the operation span test. *Psychologica Belgica*.
- De Neys, W., Schaeken, W., & d'Ydewalle, G. (in press). Causal conditional reasoning and strength of association: The disabling condition case. *European Journal of Cognitive Psychology*.
- 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(6), xxx.
- De Neys, W., Schaeken, W., & d'Ydewalle, G. (2003). Inference suppression and semantic memory retrieval: Every counterexample counts. Submitted manuscript.
- 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.
- 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., 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.
- 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* (pp. 515-555). Cambridge, MA: MIT Press.
- Evans, J. St. B. T., & Over, D. E. (1996). *Rationality and reasoning*. Hove, UK: Psychology Press.
- Evans, J. St. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human reasoning: The psychology of deduction*. Hove & London: Erlbaum.
- Galotti, K. M. (1989). Approaches to studying formal and everyday reasoning. *Psychological Bulletin*, 105, 331-351.
- Gilhooly, K. J., Logie, R. H., & Wynn, V. (1999). Syllogistic reasoning tasks, working memory, and skill. *European Journal of Cognitive Psychology*, 11, 473-498.
- Goel, V., Buchel, C., Frith, C., & Dolan, R. J. (2000). Dissociation of mechanisms underlying syllogistic reasoning. *Neuroimage*, 12, 504-514.
- Harman, G. (1986). *Change in view*. Cambridge: MIT Press.
- 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*. Cambridge, MA: Cambridge University Press.
- Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction*. Hillsdale, NJ: Laurence Erlbaum.
- 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.
- Just, M. A., Carpenter, P. A., & Keller, T. A. (1996). The capacity theory of comprehension: New frontiers of evidence and arguments. *Psychological Review*, 103, 773-780.
- 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.
- Klaczynski, P. A. (2001). Analytic and heuristic processing influences on adolescent reasoning and decision making. *Child Development*, 72, 844-861.
- 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.
- Levy, B. J., & Anderson, M. C. (2002). Inhibitory processes and the control of memory retrieval. *Trend in Cognitive Sciences*, 6, 299-305.
- 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.

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.

Miyake, A., & Shah, P. (Eds.). (1999). *Models of working memory: Mechanisms of active maintenance and executive control*. Cambridge, MA: Cambridge University Press.

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.

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, & Cognition*, 26, 883-899.

Osman, M. (2002). Is there evidence for unconscious reasoning processes?. *Proceedings of the Annual Conference of the Cognitive Science Society*, 24, 732-737.

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

Radvansky, G. A. (1999). Memory retrieval and suppression: The inhibition of situation models. *Journal of Experimental Psychology: General*, 128, 563-579.

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.

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.

Schroyens, W., Schaeken, W., & d'Ydewalle, G. (2001). The processing of negations in Conditional reasoning: A meta-analytic study in mental model and/or mental logic theory. *Thinking and Reasoning*, 7, 121-172.

Schroyens, W., Schaeken, W., Fias, W., & d'Ydewalle, G. (2000). Heuristic and analytic processes in propositional reasoning with negatives. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26, 1713-1734.

Schroyens, W., Schaeken, W., & Handley, S. (in press). In search of counter examples: Deductive rationality in human reasoning. *Quarterly Journal of Experimental Psychology*.

Shallice, T., & Burgess, P. W. (1993). Supervisory control of thought and action. In A. D. Baddeley and L. Weiskrantz (Eds.), *Attention: Selection, awareness, and control: A tribute to Donald Broadbent* (pp. 171-187). Oxford: Oxford University Press.

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

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

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, N. J.: Erlbaum.

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

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

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

Wickelgren, I. (1997). Getting a grasp on working memory. *Science*, 275, 1580-1582.

Appendix A

The conditionals for the counterexample generation tasks of Experiment 1 and 2:

Alternative generation tasks:

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 stuck, 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.

Disabler generation tasks:

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 stuck, 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.

Appendix B

The conditionals for the reasoning task of Experiment 3 and 4:

If John studies hard, then he does well on the test.

If Bart's food goes down the wrong way, then he has to cough.

If the trigger is pulled, then the gun fires.

If the intensity of light increases, then the pupils of the eyes grow smaller.

If Jenny turns on the air conditioner, then she feels cool.

If water is poured on the campfire, then the fire goes out.

If the ignition key is turned then the car starts.

If Tom grasps the glass with his bare hands, then his fingerprints are on it.

Appendix C

High and Low Spans Mean Acceptance Rating of the Four Inference Types in Function of the Number of Alternatives and Disablers When No Secondary Task Was Imposed ("No Load") and When Concurrently Tapping the Complex Finger Pattern ("Load").

Counterexample Group	Low Span								High Span							
	MP		DA		MT		AC		MP		DA		MT		AC	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
No Load (Experiment 3)																
Few Alternatives	6.03	.71	5.72	.78	5.01	1.13	6.13	.58	6.38	.64	5.40	1.46	5.16	1.21	5.67	1.25
Many Alternatives	5.47	.97	4.19	1.11	4.38	.96	4.41	1.07	6.12	.78	4.03	1.58	4.88	1.17	4.13	1.64
Few Disablers	6.06	.78	4.66	1.00	5.01	1.01	4.90	.90	6.39	.65	4.43	1.46	5.20	1.21	4.71	1.29
Many Disablers	5.44	.91	5.25	.93	4.38	1.23	5.63	.75	6.11	.78	5.00	1.50	4.85	1.17	5.10	1.45
Load (Experiment 4)																
Few Alternatives	6.45	.55	6.13	.83	5.59	1.00	6.16	.86	6.26	.74	5.60	1.39	4.90	1.36	5.88	1.38
Many Alternatives	6.13	.67	5.14	1.35	5.19	1.07	5.25	1.14	5.79	.95	4.37	1.26	4.41	1.35	4.35	1.52
Few Disablers	6.42	.49	5.41	1.36	5.63	1.04	5.52	1.03	6.21	.87	4.66	1.33	4.91	1.23	5.04	1.31
Many Disablers	6.16	.63	5.86	.81	5.16	1.03	5.89	1.00	5.85	.73	5.31	1.30	4.40	1.33	5.19	1.40

Table 1

Mean Number and Standard Deviations of Correct Taps per Second During Baseline-Tapping and During Concurrent Counterexample Retrieval

Condition	Period			
	Baseline		Counterexample	
	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

Table 2

Mean Number and Standard Deviations of Correct Taps per Second During Baseline-Tapping and During Concurrent Conditional Reasoning

Span Group	Period			
	Baseline		Reasoning	
	M	SD	M	SD
High spans (n = 17)	2.60	0.73	2.08	0.54
Low spans (n = 16)	2.44	0.56	2.05	0.67

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.

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.

Figure 3. High and Low Spans' mean acceptance rating of the four inference types. The rating scale ranged from 1 (very sure cannot draw this conclusion) to 7 (very sure can draw this conclusion).

Figure 4. High and Low spans' mean inference acceptance ratings in function of the number of available disablers (MP and MT) and alternatives (AC and DA). The rating scale ranged from 1 (very sure cannot draw this conclusion) to 7 (very sure can draw this conclusion).

Figure 5. Low and High spans' mean acceptance rating of the four inference types while concurrently tapping the complex finger pattern ("Load") and when there was no secondary task imposed ("No Load"). The rating scale ranged from 1 (very sure cannot draw this conclusion) to 7 (very sure can draw this conclusion).

PBIB

009591510

+99.10 DENE 2003