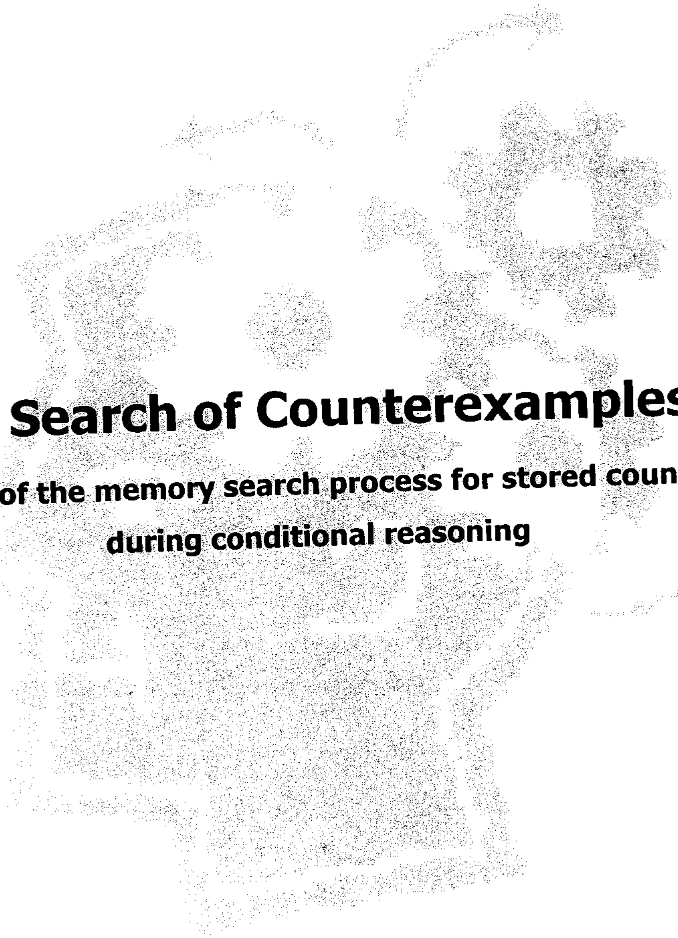
Katholieke Universiteit Leuven
Faculteit Psychologie en Pedagogische Wetenschappen
Laboratorium voor Experimentele Psychologie



In Search of Counterexamples

A specification of the memory search process for stored counterexamples during conditional reasoning

Proefschrift aangeboden
tot het verkrijgen van de graad van
Doctor in de Psychologische Wetenschappen
door Wim De Neys

Promotor: Prof. Dr. G. d'Ydewalle
Copromotor: Prof. Dr. W. Schaeken

2003

DE NEYS Wim, In search of counterexamples. A specification of the memory search process for stored counterexamples during conditional reasoning.
Dissertation submitted to obtain the degree of Doctor in Psychology, May 2003.
Supervisor: Prof. Dr. G. d'Ydewalle, Co-supervisor: Prof. Dr. W. Schaeken

When people reason with meaningful conditionals in daily life (e.g., 'If the brake is pushed, then the car slows down') they rely on stored background knowledge about the conditional. Most people will, for example, remember that cars can also slow down when the gear is shifted or that in case of black ice the car might start slipping instead of slowing down. Retrieval of this kind of background knowledge, also known as counterexamples, has a profound impact on the inferences people draw. The effects of retrieved counterexamples on the reasoning process are well established. However, it is not yet established how the counterexamples are actually retrieved. Current reasoning theories lack a specification of the background knowledge search process. This thesis addresses the shortcoming. The basic goal is a specification of the processing characteristics of the counterexample search during everyday conditional reasoning.

Chapter 1 presents the general framework. The impact of background knowledge on the reasoning process has become one of the most intensely studied topics in the conditional reasoning field. In particular, the role of the availability of two specific types of stored background knowledge, alternative causes and disabling conditions (or counterexamples in short), has attracted interest. Chapter 1 reviews the pioneering studies and presents the reasoning model of Markovits and colleagues (e.g., Markovits & Barrouillet, 2002) that served as a starting point for the thesis.

The pioneering studies established that the number of stored counterexamples and the strength of association of stored counterexamples affected successful retrieval. However, the importance of the strength of association factor was only examined for one type of counterexample, the alternative causes (alternatives). The study reported in *Chapter 2* solved this shortcoming and established that the outcome of the disabler search process is also determined by the associative strength factor.

The study described in *Chapter 3* started with an examination of the relation between different factors affecting the retrieval of stored counterexamples. We found high positive correlations between the number of counterexamples, their associative strength, and plausibility. This supported the central assumption that in a memory structure with many stored elements, successful counterexample retrieval is more likely. The study also characterized the time course of the search process and examined individual difference in search efficiency. As expected we found that the search time depends on the number of stored counterexamples and that the reasoning performance is affected by the efficiency of the disabler retrieval.

Chapter 4 specified the "stopping" characteristic of the retrieval process. We established that during everyday conditional reasoning the search does not stop after retrieval of a single counterexample. Instead, the search continues and, if possible, additional counterexamples will be activated. Every additionally retrieved counterexample will gradually decrease the inference acceptance.

The study reported in *Chapter 5* examined whether the counterexample search is purely automatic or whether the search also draws on executive working memory (WM) resources. A combination of correlational and dual-task studies indicated that the search starts with an automatic, spreading of activation but that WM-resources are recruited for a more efficient retrieval. Findings further established that participants highest in WM-capacity spontaneously allocate WM-resources to an inhibition of the automatic counterexample activation when the type of counterexample conflicted with the logical validity of a reasoning problem.

Chapter 6 tested specific trend predictions concerning the relation between WM-capacity and reasoning performance for different inferences types. Findings further supported the proposed WM-dependent retrieval mechanism.

In the concluding *Chapter 7*, we wrap up the findings and sketch a model of the counterexample retrieval process based on the established search specifications.

DE NEYS Wim, Op zoek naar tegenvoorbeelden. Een specificatie van het geheugenzoekproces naar opgeslagen tegenvoorbeelden tijdens conditioneel redeneren.
Proefschrift aangeboden tot het verkrijgen van de graad van Doctor in de Psychologische Wetenschappen, mei 2003.
Promotor: Prof. Dr. G. d'Ydewalle, Copromotor: Prof. Dr. W. Schaeken

Wanneer mensen met dagdagelijkse conditionele zinnen redeneren (bv. 'Als de rem wordt ingeduwd, dan vertraagt de wagen') maken ze gebruik van opgeslagen achtergrondkennis over de conditionele zin. De meeste mensen zullen bijvoorbeeld bedenken dat een wagen ook kan vertragen door naar een lagere versnelling te schakelen of dat in geval van ijzel de wagen kan slippen in plaats van te vertragen. Het terugvinden van dit soort achtergrondkennis, ook tegenvoorbeelden genoemd, heeft een sterke impact op de inferenties die mensen maken. De effecten van teruggevonden tegenvoorbeelden zijn goed gedocumenteerd. Er is echter nog niet verduidelijkt hoe mensen de tegenvoorbeelden eigenlijk gaan zoeken. De huidige redeneertheorieën missen een specificatie van het zoekproces naar de opgeslagen achtergrondkennis. Dit proefschrift tracht deze tekortkoming te verhelpen. Het doel is een elementaire specificatie van de karakteristieken van het zoekproces naar opgeslagen tegenvoorbeelden tijdens alledaags conditioneel redeneren.

Hoofdstuk 1 behandelt het algemene denkraam. Het onderzoek naar de impact van achtergrondkennis op het redeneren is de laatste jaren uitgegroeid tot dé centrale onderzoeksproblematiek binnen het conditioneel redeneerveld. Vooral het onderzoek naar de beschikbaarheid van twee specifieke types van opgeslagen achtergrondkennis, mogelijke alternatieve en verhinderende omstandigheden (of kortweg tegenvoorbeelden), heeft veel aandacht opgeëist. Hoofdstuk 1 geeft een overzicht van de baanbrekende studies en stelt het redeneermodel van Markovits (bv. Markovits & Barrouillet, 2002) voor dat als vertrekpunt voor het proefschrift fungeerde.

De vroegere, baanbrekende studies toonden aan dat het aantal opgeslagen tegenvoorbeelden en de associatiesterkte van de tegenvoorbeelden de kans dat het zoekproces met succes een tegenvoorbeeld terugvindt bepaalden. Het belang van de associatiesterkte werd echter enkel onderzocht voor één type tegenvoorbeelden, de alternatieve omstandigheden. De studie die gerapporteerd wordt in *Hoofdstuk 2* loste deze tekortkoming op en bevestigde dat de uitkomst van het zoekproces naar opgeslagen verhinderende omstandigheden ook bepaald wordt door de associatiesterkte.

De studie die beschreven wordt in *Hoofdstuk 3* begon met het in kaart brengen van de relatie tussen de verschillende factoren die een effect hebben op het vinden van een tegenvoorbeeld. We vonden hoge positieve correlaties tussen het aantal tegenvoorbeelden, de associatiesterkte van de tegenvoorbeelden en hun plausibiliteit. Dit bevestigde de centrale assumptie dat het waarschijnlijker is dat er een tegenvoorbeeld gevonden wordt in een geheugenstructuur met veel opgeslagen elementen. De studie specificeerde ook het tijdsverloop van het zoekproces en onderzocht interindividuele verschillen in de efficiëntie van het zoekproces. Zoals verwacht vonden we dat de zoektijd afhangt van het aantal opgeslagen tegenvoorbeelden en dat de redeneerprestatie afhankelijk is van hoe goed men verhinderende omstandigheden kan terugvinden.

Hoofdstuk 4 specificeert de "stop" karakteristiek van het zoekproces. We toonden aan dat het zoekproces tijdens dagdagelijks conditioneel redeneren niet stopt nadat één enkel tegenvoorbeeld is teruggevonden. Het zoekproces zal verdergezet worden en indien mogelijk zullen bijkomende opgeslagen tegenvoorbeelden geactiveerd worden. Elk bijkomend teruggevonden tegenvoorbeeld zal de inferentie aanvaarding gradueel doen dalen.

De studie die gerapporteerd wordt in *Hoofdstuk 5* onderzocht of het tegenvoorbeeld zoekproces puur automatisch verloopt of dat er een beroep wordt gedaan op het werkgeheugen (WG) bij het zoeken. Een combinatie van correlationeel en dubbeltaak onderzoek verduidelijkte dat het zoeken wel start met een automatische activatieverspreiding, maar dat nadien gebruikt wordt gemaakt van het werkgeheugen voor een efficiënter zoeken. De bevindingen wezen er verder op dat proefpersonen met de grootste WG-capaciteit hun WG-capaciteit spontaan aanwenden voor het inhiberen van het automatisch zoekproces in het geval de aard van een tegenvoorbeeld conflicteert met de logische validiteit van een redeneerprobleem.

In *Hoofdstuk 6* werden specifieke trend predicties getest in verband met de relatie tussen WG-capaciteit en redeneerprestatie voor verschillende inferentietypes. De bevindingen ondersteunden verder het voorgestelde WG-afhankelijke tegenvoorbeeld zoekmechanisme.

In het afsluitende *Hoofdstuk 7* worden de bevindingen samengevat en wordt een model van het zoekproces geschetst op grond van de gespecificeerde proceskarakteristieken.

Dankwoord

Géry d'Ydewalle heeft als promotor van mijn onderzoek heel wat beslommeringen op zich genomen. In de eerste plaats zorgde hij ervoor dat er steeds genoeg fondsen voorhanden waren om proefpersonen te betalen en aan congressen deel te nemen. Dankzij Géry stonden er in mijn officiële documenten ook geen tientallen dt-fouten, werd ik vertrouwd met de kleine kantjes van STATISTICA© en weet ik nu hoe je op een diplomatische manier duidelijk maakt dat een "reviewer" niet bepaald de meest intelligente mens op aarde is.

Mijn copromotor, Walter Schaeken, heeft het grootste deel van de inhoudelijke begeleiding voor zijn rekening genomen. Ik kon steeds bij Walter terecht voor advies en suggesties. Daarbij kon ik ook steeds in alle vrijheid blijven beslissen welke weg ik wou inslaan. Walters "adviseren zonder te dirigeren" aanpak was voor mij de ideale begeleidingsvorm. Samen met zijn inhoudelijke begeleiding zorgden Walters motiverende managerskwaliteiten er ook voor dat ik niet aan de typische "doctoraatsblues" heb geleden.

Walter heeft de voorbij jaren ook een excellente groep onderzoekers begeleid en bij elkaar gebracht. Op verschillende congressen heb ik zelf kunnen ervaren dat deze Leuvense redeneergroep heel wat aanzien geniet. De discussies met Kristien, Niki, Jean-Baptiste en Walter Schroyens hebben mijn denken enorm gestimuleerd en verfijnd. Ik heb ook heel wat praktische hulp en steun gekregen van de groep. Daarnaast zijn het ook gewoon fijne mensen met wie ik me goed geamuseerd heb. Dankzij deze redeneerders weet ik nu dat een luchthaven niet de meest comfortabele plaats is om te overnachten. Begrippen als "chicken sandwich" en "whisky sour" zullen dankzij sommige vrouwelijke groepsleden ook nooit meer hetzelfde klinken.

Ik heb het voordeel gehad dat mijn artikels meestal gereviewed werden door enkele van de grote experts in het veld. Alhoewel ik er in eerste instantie soms om gevloekt heb, hebben de opmerkingen van deze mensen een fundamentele bijdrage geleverd aan mijn onderzoek. Ik vond het ook enorm knap dat enkele van deze "grote" mannen en vrouwen zich niet te beroerd voelden om me op eenvoudig verzoek wat meer specifiek advies te geven. Daarom wil ik Randy Engle, Denise Cummins, Henry Markovits, Mike Oaksford en Ruth Byrne hier ook nog eens bedanken.

Zonder Ine Callebaut hadden mijn proefpersonen nooit de taakinstructies begrepen, was mijn werkgeheugendata nooit gescoord geraakt en had Géry me al lang ontslaan omdat er in al mijn officiële documenten honderden dt-fouten stonden. Zonder Ine had ik waarschijnlijk ook nog altijd niet aanvaard dat "pizza" geen frequent gebruikt Nederlands woord is.

Dirk Asselman zou ik willen danken voor de soft- en hardware hulp de voorbij jaren. Rob Stroobants mag ik ook niet vergeten te danken voor alle statistische hulp.

Over de vraag of de mannen van Pavel Milkovic en PNV Waasland veel hebben bijgedragen tot het welslagen van dit doctoraat kan gediscussieerd worden. Maar beloofd is beloofd. Bedankt mannen. De cateringsgewijze steun van Bart Verheyden zal ik ook niet vergeten.

Wie ik tenslotte nog wil bedanken zijn mijn vader en moeder, Hubert De Neys en Frieda Asselman. Zij hebben me altijd mijn ding laten doen, ook al werd hen dat door sommige mensen afgeraden. Het kan raar klinken, maar zonder die vrijheid was ik nu al lang gebotst met alles wat met gezag en regels te maken heeft. Dat ik hier vandaag een doctoraat sta te verdedigen en niet ergens centraler in Leuven zit, heb ik in de eerste plaats dan ook aan hen te danken.

Bedankt allemaal,

Wim

Table of Contents

List of Tables

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 5

CHAPTER 6

List of Figures

CHAPTER 3

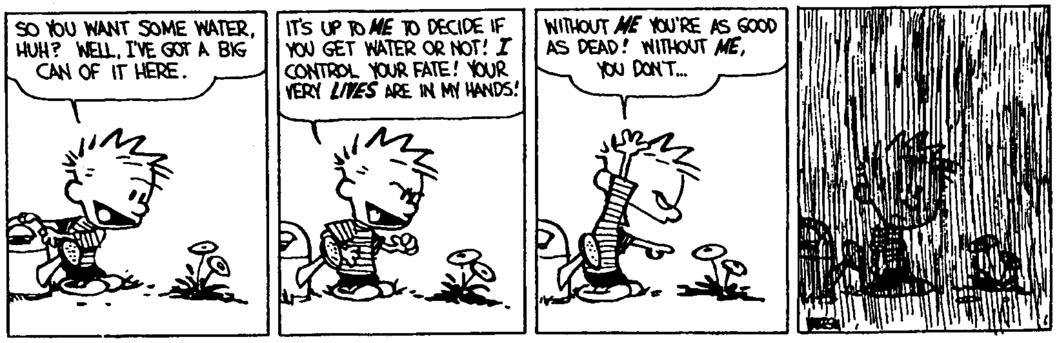
CHAPTER 4

CHAPTER 5

CHAPTER 6

CHAPTER 1

General framework and overview


SO YOU WANT SOME WATER, HUH? WELL, I'VE GOT A BIG CAN OF IT HERE.
IT'S UP TO *ME* TO DECIDE IF YOU GET WATER OR NOT! *I* CONTROL YOUR FATE! YOUR VERY *LIVES* ARE IN MY HANDS!
WITHOUT *ME* YOU'RE AS GOOD AS DEAD! WITHOUT *ME*, YOU DON'T...

Copyright © 1990 by Bill Watterson

Suppose you are told ‘If Mark turns the ignition key, then his car will start’. If you are told next that Mark‘s car started, it is likely that you will conclude that the ignition key was turned. However, suppose that you would have been reminded of the fact that the car might also be hot wired by a car thief or started with a push button. In that case you would probably have been rather reluctant to conclude that the ignition key was turned. Likewise, when you are told that Mark turned the ignition key, you might initially conclude that his car will start. However, if you would consider that the fuel tank might be empty or that the engine might be broken, you would be less prepared to accept the conclusion.

The ability to think conditional, ‘if, then’, thoughts is considered 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 is has become one of the most intensely studied topics in human reasoning research (Evans, Newstead, & Byrne, 1993).

In daily life people typically reason with meaningful, realistic conditionals (e.g., ‘If the ignition key is turned, then the car starts’). As the introductory example makes clear, our long-term memory contains relevant background knowledge about these conditionals. Accessing and retrieving this knowledge will affect the kind of inferences people draw. If you remember that cars do not start when the fuel tank is empty or that a car thief may start a car by hot-wiring it, this will prevent you from drawing certain inferences. Likewise, the cartoon on the previous page illustrates how Calvin erroneously concluded that plants would die if he did not water them because he failed to think of rain as possible cause for the plants’ survival.

The impact of background knowledge on the reasoning process has long been acknowledged (e.g., Matalon, 1962; Staudenmayer, 1975). Current reasoning theories propose a number of accounts that try to explain how retrieved background information affects the reasoning process. However, the more basic question of how the background information is actually retrieved has not yet been dealt with. This issue lies at the heart of the present dissertation. The primary goal is a specification of the characteristics of the background knowledge retrieval process during everyday conditional reasoning. The research will focus on the retrieval of two specific types of stored background knowledge known as ‘counterexamples’: Alternative causes and disabling conditions.

1. INTRODUCTION

In the last few years research on the impact of background knowledge on the reasoning process has moved to the center stage of the conditional reasoning field (Evans, 2002; Evans et al., 1993; Manktelow, 1999). In particular, the role of the availability of *alternative causes* and *disabling conditions* has attracted interest.

An alternative cause (alternative) is a possible cause that can also produce the effect mentioned in the conditional (e.g., hot wiring the car in the introductory example). A disabling condition (disabler) prevents the effect from occurring despite the presence of the cause (e.g., a broken engine in the introductory example). Consider the following conditional:

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

Possible alternative causes for this conditional are:

Taking off some clothes, the weather cools, taking a swim…

The alternatives make it clear that it is not necessary to turn on the air conditioner in order to feel cool. Other causes are also possible.

Possible disabling conditions are:

Air conditioner is broken, having fever, window open…

If such disablers are present, turning on the air conditioner will not result in feeling cool. The disablers make it clear that it is not sufficient to turn on the air conditioner in order to feel cool. Additional conditions must be fulfilled.

Investigations of conditional reasoning typically ask people to assess 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

The first part of the conditional (p) is called the antecedent, and the second part (q) is called the consequent. In standard logic MP and MT are considered valid inferences, while AC and DA inferences are considered fallacies. So, when you are told 'If the ignition key is turned, then the car will start' and you receive the information that the ignition key was turned, then logic tells you to accept the conclusion that the car will start (an MP inference). Likewise, if you receive the information that the car does not start, you should infer that the ignition key was not turned (an MT inference). On the other hand, logically speaking, from the information that the car starts, you should not infer that the ignition key was turned (an AC inference). Likewise, upon knowing that the ignition key was not turned you should reject the conclusion that therefore the car will not start (a DA inference).

In a pioneering study, Rumain, Connell, and Braine (1983) showed that when a possible alternative was explicitly presented to participants, the AC and DA inferences were less endorsed. Byrne (1989) replicated this effect and observed a similar effect on MP and MT when a possible disabler was mentioned. Consider, for example, the following conditional:

If she has an essay to write, then she will study late in the library.

Byrne (1989) found that explicitly presenting a possible alternative, for example,

If she has an essay to write, then she will study late in the library.
If she has some textbook to read, then she will study late in the library.

resulted in fewer AC ('She will study late in the library. Therefore, she has an essay to write') and DA ('She does not have an essay to write. Therefore, she will not study late in the library.') inferences. Explicitly presenting a possible disabler, for example,

If she has an essay to write, then she will study late in the library.
If the library stays open late, then she will study late in the library.

resulted in an increased rejection of the MP ('She has an essay to write. Therefore, she will study late in the library.') and MT ('She will not study late in the library. Therefore, she has not an essay to write') inferences. These findings have come to be known as the suppression effect (see Dieussaert, Schaeken, Schroyens, & d'Ydewalle, 2000, for a detailed discussion). Further on, I adopt Byrne's (1989) terminology and will refer to alternatives and disablers as counterexamples.

Further research established that the suppression effect arises even without explicit counterexample presentation (e.g., Cummins 1995; Cummins, Lubart, Alksnis, & Rist, 1991; Markovits, 1986). Cummins (1995) and Cummins et al. (1991) addressed the role of stored counterexample retrieval by examining the effect of the number of counterexamples that are available for a conditional. In a pretest participants were asked to generate as many counterexamples as possible for a set of conditionals. On the basis of the number of generated items, the conditionals were classified in groups with many or few alternatives and disablers. These conditionals were adopted for a conditional reasoning task with a new group of participants.

Cummins' (1995) results showed that people accepted the AC and DA inferences less frequently for conditionals with many possible alternatives (e.g., 'If plants are fertilized, then they grow well') than for conditionals with only few possible alternatives (e.g., 'If Dan grasps the glass with his bare hands, then his fingerprints are on it'). Furthermore, although the valid MP and MT inferences were accepted quite frequently for conditionals with only few possible disablers (e.g., 'If water is heated to 100°C, then it boils'), a large number of possible disablers (e.g., 'If Anna studies hard, then she does well on the test') also resulted in a decreased MP and MT acceptance. Alternatives and disablers were not explicitly presented, indicating that during conditional reasoning people spontaneously search their long-term memory for stored counterexamples. Thereby, the number of available counterexamples will affect the probability of successful counterexample retrieval and this will determine the conclusions people draw.

It is widely acknowledged that a theory of conditional reasoning cannot be complete without a full understanding of the counterexample retrieval process (e.g., Johnson-Laird & Byrne, 1994; Thompson, 1994). A vast amount of research in connection with the suppression effect has already resulted in a number of accounts (e.g., Byrne, Espino, & Santamaria, 1999; Johnson-Laird & Byrne, 2002; Oaksford & Chater, 1998; Politzer, in press; Thompson, 2000). However, these accounts try to explain how the retrieved information affects the reasoning process. The more fundamental question of how the information is actually retrieved has not yet been dealt with. The characteristics of the search process itself remain largely unknown (Johnson-Laird & Byrne, 1994; Oaksford & Chater, 2001). It is telling that

Johnson-Laird and Byrne use this lack-of-knowledge as an argument to counter the critique that their reasoning theory is not specifying the search process:

> "..., so far, no evidence has revealed anything about the search process itself. The theory accordingly refrains from specifying the process. To refrain from speculation seemed like prudence rather than a major flaw." (Johnson-Laird & Byrne, 1994, p. 776)

The present dissertation tries to fill this crucial 'knowledge-gap'. The goal of the research is specifying the processing characteristics of the counterexample retrieval during everyday conditional reasoning. It is obvious that the core of the research will involve an interplay between reasoning and memory research. The starting point is the work of the research group of Markovits and colleagues. Recently, these researchers proposed a first, raw sketch of the search process (e.g., Barrouillet, Markovits, & Quinn, 2001; Markovits & Barrouillet, 2002; Markovits, Fleury, Quinn, & Venet, 1998; Markovits & Quinn, 2002; Quinn & Markovits, 2002). The innovative aspect of their work is that the proposed specification was based upon general principles and assumptions from influential memory models (e.g., Anderson, 1983, 1993; Anderson & Lebiere, 1998; Cowan, 1995, 2001). The sketched search mechanism was also incorporated into a general model of conditional reasoning.

The next section gives a brief overview of the Markovits model (see Markovits & Barrouillet, 2002, for an extensive review). The reader who needs a more general review of studies that examined the impact of retrieved counterexamples on the reasoning performance is referred to Politzer (in press) and Politzer and Bourmaud (2002).

2. GENERAL FRAMEWORK

2.1 Overview of Markovits' model

2.1.1 Specification of the counterexample search mechanism

The Markovits model states that when people are confronted with a conditional, they will start constructing a basic mental representation of the elementary information the conditional contains. The elementary information concerns the antecedent and consequent of the conditional and the fact that the occurrence of the antecedent is associated with the occurrence of the consequent. This representation is maintained in working memory. Working memory is simply conceived as the activated (above a certain threshold) portion of long-term memory (e.g., Cowan, 1995; Engle & Oransky, 1999). Next, memory structures in long-term (semantic) memory storing relevant information for the evaluation of the conditional will be automatically activated. As suggested by many authors (Anderson, 1993; Cowan, 2001; Logan, 1988), it is assumed that activation will automatically start to spread from the elements in working memory or the 'focus of attention' towards related elements in long-term memory.

The crucial memory structures during conditional reasoning contain alternatives and disablers linked with the conditional. There is also an important third type of structure that contains 'complementary' elements. The complementary elements refer to cases for 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 associated with not-q). For example, the complementary structure for the conditional 'If it rains, then the streets get wet' would be composed of such cases as 'The sun shines and the street is dry' and 'It is only cloudy, and the streets are dry'.

According to many memory models (e.g., Anderson, 1983; Gillund & Shiffrin, 1984), the probability of retrieving at least one element from a semantic memory structure will increase when the number of elements stored in the structure increases. More specifically, the probability of retrieving at least one alternative from the memory structure storing alternatives will be higher for conditionals with many (vs. few) possible alternatives. Likewise, the probability of retrieving at least one disabler from the memory structure storing disablers will be higher for conditionals with many (vs. few) possible disablers.

2.1.2 Incorporation in mental models theory

Markovits incorporated this general specification of the counterexample search process in the mental models theory (MMT, see Johnson-Laird, 1983) in order to explain how a retrieved counterexample would affect the inferences one draws.

Mental models theory basically states that people reason by constructing and manipulating internal representations (mental models) of the information a reasoning problem contains. According to Markovits, the before mentioned elementary representation that people construct when confronted with a conditional amounts to a first, initial model. For example, presented the conditional 'If the brake is pushed, then the car slows down', a reasoner will construct an initial model similar to this one:

brake --- slow

The initial model links the occurrence of the antecedent and consequent of the conditional. Therefore, based on the initial model, the MP (e.g., 'The brake is pushed, therefore the car slows down') and AC (e.g., 'The car slows down, therefore the brake was pushed.') inferences will be accepted; there are no models that contradict the conclusions.

The initial model can be extended depending on the outcome of the memory search process. Successful retrieval of a complementary case will lead to the construction of an additional model that expresses the fact that the absence or non-occurrence of the antecedent is associated with the absence of the consequent. For example:

brake --- slow
not brake --- not slow

When such an additional model is constructed, the MT ('The car does not slow down, therefore the brake was not pushed.') and DA (e.g., 'The brake was not pushed, therefore the car does not slow down') inferences will be accepted. The first model tells us only something about the case where antecedent and consequent do occur and does not contradict the second, complementary model.

However, successful retrieval of an alternative will lead to the construction of an additional model representing that the consequent can occur without the antecedent. For example:

brake --- slow
not brake --- not slow
not brake --- slow

Based on this model, the AC and DA inferences will be less accepted; the AC and DA conclusions that were originally supported by the initial and complementary model are contradicted by the alternative model.

Finally, successful retrieval of a disabler will result in the construction of a model where the occurrence of the antecedent is not associated with the occurrence of the consequent. For example:

brake --- slow
not brake --- not slow
brake --- not slow

This model will lead to an increased rejection of the MP and MT inferences; the additional disabler model contradicts the MP and MT conclusions that were originally supported by the initial and complementary model.

Markovits' conditional reasoning model provides a good explanation for Cummins' (1995) findings. For conditionals with few possible alternatives the probability of successful alternative retrieval will be rather low. Thus, for most people the alternative search will be unsuccessful. Therefore, they will stick to the initial and complementary models and consequently accept the AC and DA inferences. For conditionals with many possible alternatives the probability of successful alternative retrieval will be rather high. Therefore, most people will construct an additional alternative model that will result in a decreased AC and DA acceptance. Likewise, the probability of successful disabler retrieval will increase for conditionals with many possible disablers. Therefore, it will be more likely that an additional disabler model is constructed and consequently MP and MT acceptance will decrease for the conditionals with many possible disablers.

2.2 Focus and scope of the research

2.2.1 Everyday reasoning

The present thesis does not examine the counterexample retrieval in a formal, deductive reasoning task, but rather in a situation closer to everyday reasoning. The studies adopt realistic conditionals so that participants have access to relevant background knowledge about counterexamples. Furthermore, 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). Instead, participants are told that they can use the criteria they personally judge relevant for the evaluations in the reasoning task. Although participants are still situated in a laboratory setting, these task characteristics should allow and encourage people to reason as they would in everyday life (Cummins, 1995; see also Galotti, 1989).

Markovits (2002; Quinn & Markovits, 2002) has specified that his model primarily describes the retrieval process in a formal, deductive reasoning task. That is, although Markovits' studies use realistic conditionals, participants are nevertheless explicitly instructed to reason logically. Reasoning in a formal context may alter the characteristics of the reasoning mechanisms (Evans, 2002; Johnson-Laird & Byrne, 1991; Markovits, 2002; Oaksford, Chater, & Larkin, 2000). It may be the case that some characteristics of the retrieval process differ in everyday and formal reasoning. Therefore, the specifications and revisions that will be suggested in this dissertation should not be conceived as a mere critique of Markovits' model, but rather as an attempt to extend it to reasoning in everyday life.

Accounting for peoples daily life reasoning behavior is considered the ultimate goal of cognitive reasoning research (Johnson-Laird, 1983; Oaksford & Chater, 1998). With this goal in mind I simply opted to study the counterexample retrieval directly during everyday reasoning instead of developing a model of the counterexample retrieval in a formal reasoning task first and trying to account for everyday reasoning in a later stage (e.g., Evans, 2002).

2.2.2 Causal conditionals

The research focuses on a specification of the counterexample search with realistic, causal conditionals. Causal conditionals are simply conditionals that express a causal relation: The antecedent specifies a cause and the consequent specifies an effect associated with the occurrence of the cause in question (e.g., 'If Mark turns the ignition key, then the car will start'). Thus, the studies do not include other possible conditional contents such as promises (e.g., 'If you do your homework, you get some candy'), class-based information (e.g., 'If an animal is a cow, then it has four legs'), or abstract information (e.g., 'If it is an A, then it is a 2').

The rationale is that one cannot assume from the outset that the counterexample retrieval for these different conditionals has precisely the same characteristics. For example, for an abstract conditional there is by definition no background knowledge about possible disablers available. The counterexample retrieval for class-based conditionals may differ, for example, because it will be much easier to retrieve the counterexamples. Alternatives for class-based conditionals typically belong to 'common, well established categories' (Barsalou, 1983), instances of which are readily accessible in memory (e.g., 'animals with four legs'). The stored alternatives for these conditionals are typically strongly associated with the consequent and therefore they are easily accessed and retrieved. The counterexamples for causal conditionals resemble what Barsalou (1983) called 'ad hoc' categories (e.g., 'things to take from one's home during a fire'). Ad hoc categories are less well established in memory and retrieving instances from these categories is more difficult.

Searching counterexamples for causal conditionals requires constructing a 'naïve causal structure' (Cummins, 1995). For example, for the causal conditional 'If a rock is thrown at a window, the window will break' searching alternatives implies constructing the class of things that can break windows, which in turn requires specifying such characteristics as 'hard things', that reflect theories about how things can be broken (Markovits & Barrouillet, 2002). The point is that the search process for causal conditionals will require more complex processing. This may result in discrepancies between findings with causal and class-based conditionals (Markovits & Barrouillet, 2002).

Therefore, as Cummins (1995) and most of Markovits' work, the research initially focuses on conditional reasoning with causal conditionals.

2.2.3 Reasoning theory neutrality

Markovits incorporated his specification of the counterexample search process into the mental models theory (MMT, Johnson-Laird, 1983). Such an incorporation makes good sense because the theory already presupposed such a process (i.e., searching counterexamples for a conclusion derived from an initial model, note that the problem is precisely that a clear specification of this search component is missing). There is a free 'slot' in the theory where the search specification can be fitted in. The theory already clarifies how the result of the search process will further affect the reasoning process. For rival reasoning theories, for example, mental logic (Braine & O'Brien, 1998; Rips, 1994) or the probabilistic approach (Oaksford & Chater, 1998, 2001; Oaksford et al., 2000), this is not yet the case.

Mental logic claims that the content of a reasoning problem and related background knowledge will affect the encoding of the problem into a more abstract representation. The resulting abstract representation determines which inference rule will be applied. However, the crucial problem is that, to date, mental logicians have refrained from presenting a clear description of the encoding process (Johnson-Laird & Byrne, 1993; Rips, 1994; but see Politzer & Bourmaud, 2002, for some first suggestions).

The probabilistic approach posits that people interpret conditionals probabilistically. Conditional reasoning is presumed to amount to a conditional probability calculation. Currently, the probabilistic approach focuses on a computational explanation of the reasoning process (i.e., 'what' is computed, not 'how' it gets computed, see Oaksford & Chater, 2001). It is unclear how people would derive the crucial probabilities. Thus, a major advantage of the MMT is that the theory provides an evident 'junction point' for the counterexample retrieval process.

However, it is important to stress that, in essence, a specification of the counterexample search process can be incorporated in all the different reasoning theories. The point is merely that at present such accounts have not yet been provided. It should also be noted that Markovits' model adopts only the general principles of MMT. At certain points the model strongly deviates from the standard theory as proposed by Johnson-Laird and Byrne (1991, 2002) or Johnson-Laird, Byrne, and Schaeken (1992, 1994).

The primary goal of my research is a specification of the characteristics of the counterexample retrieval independent from a further incorporation in a specific reasoning theory. In this sense the research will be reasoning theory neutral. As one will note, for completeness, the studies in this dissertation do sometimes suggest a possible reasoning theory incorporation (e.g., Chapter 4) and typically these suggestions are based on MMT. It should be stressed that the choice for MMT in these cases is purely pragmatic. This allowed us to maintain a link with Markovits' work and to rely on the expertise of the Leuven Reasoning Group with respect to MMT (and revisions of the theory, see Schaeken, De Vooght, Vandierendonck, & d'Ydewalle, 2000; Schroyens, Schaeken, & d'Ydewalle, 2001). It should be clear that this does not imply a judgment about the incorporations in other theories. Contrary to some previous dissertations from the Leuven Reasoning Group, the different reasoning theories will therefore not be played off against one another. The studies explicitly intend to provide a theory neutral characterization of the counterexample retrieval so that a wide range of researchers can adopt the results for a further development of their accounts.

3. OVERVIEW OF THE STUDIES

The next section presents a general overview of the studies reported in this thesis. The overview will stress the basic goals, manipulations, and links between the different experiments rather than presenting a detailed description of the results.

3.1 Chapter 2: Strength of association: The disabling condition case

Cummins (1995) showed that the counterexample retrieval is affected by the number of stored counterexamples. Quinn and Markovits (1998) claimed that in addition to the number of stored elements in a memory structure, the associative strength of the individual elements should also affect the retrieval outcome: The stronger the association between a stored counterexample in a memory structure and the conditional (or the elementary representation of the conditional), the higher the probability that the counterexample will be retrieved. Although Quinn and Markovits did not give a formal definition of associative strength, the individual elements in a memory structure can be thought of as nodes in a semantic network (e.g., Anderson, 1983). A higher associative strength then amounts to a lower activation threshold.

In order to test their hypothesis, Quinn and Markovits (1998) first determined the associative strength of the possible counterexamples of a conditional. Participants had to generate as many alternatives as possible for a specific effect (e.g., 'a dog scratches constantly') within 30 s. The generation frequency (% of participants that generated a specific alternative) functioned as associative strength index. Next, they identified effects for which there existed only one strongly associated cause. Two conditionals were constructed by combining the effect with the strongly (e.g., 'If a dog has fleas, then he scratches constantly') and a weakly associated cause (e.g., 'If a dog has skin disease, then he scratches constantly').

Both the 'strong' and 'weak' conditionals have exactly the same number of possible alternatives since the consequent stays the same. However, the memory structure storing alternatives will contain only weakly associated alternatives in case of the 'strongly associated' conditional. For the 'weakly associated conditional' there will still be a strongly associated alternative available in the memory structure. If associative strength affects the probability of successful retrieval, retrieving an alternative should be easier for the weakly associated conditional. Consequently, the AC and DA inferences should be less frequently accepted for the 'weak' (vs. 'strong') conditionals. Quinn and Markovits' (1998) results confirmed the predictions.

While Cummins (1995; Cummins et al., 1991) had shown that the number of available counterexamples was important both for alternatives and disablers retrieval, Quinn and Markovits' (1998) associative strength findings concerned only alternatives. In a first study (De Neys, Schaeken, & d'Ydewalle, in press-a), described in Chapter 2, we therefore examined whether the associative strength also affected the retrieval of disablers. In a generation task we presented complete conditionals. Participants were asked to generate possible disablers and we identified conditionals for which there existed only one strongly associated disabler. We constructed inference problems in which one premise was expanded with the negation of a strongly or weakly associated disabler. Thus, in both cases one possible disabler was eliminated. For example, a possible expanded MP problem was:

> If water is heated to 100°C, then the water boils.
> The water is heated to 100°C and *the water is pure* (strongly associated).
> The water is heated to 100°C and *the pressure is normal* (weakly associated).

Both expanded reasoning problems have an equal number of possible disablers (i.e., the original number minus one). When a weakly associated disabler is eliminated the resulting disabler set will still contain a strongly associated element. However, when a strongly associated disabler is eliminated, only weakly associated elements will remain in the memory structure storing possible disablers. As expected, results indicated that MP and MT were less accepted when a weakly associated disabler was eliminated. The study thus established that the outcome of the disabler retrieval is also determined by the associative strength of stored disablers.

3.2 Chapter 3: Testing the semantic memory framework

The study described in Chapter 3 (De Neys, Schaeken, & d'Ydewalle, 2002) started with an examination of the relation between different factors affecting the retrieval of stored counterexamples. We already knew that the number of stored counterexamples and the associated strength of the counterexamples were important for successful retrieval. However, the relation between these factors was not clear. For example, it is possible that conditionals with many possible counterexamples have less strongly associated counterexamples than conditionals with only few counterexamples. Therefore, one cannot a priori assume that the probability that a counterexample is successfully retrieved is indeed higher for conditionals with many (vs. few) counterexamples. It is precisely this assumption that is crucial for the explanation of Cummins' (1995) findings in Markovits' reasoning model.

We addressed this issue in a first experiment. In a generation task, participants were asked to generate as many counterexamples as possible for a set of causal conditionals. We recorded the number of generated counterexamples for every conditional and the generation frequency of every counterexample. In addition, participants were asked to rate the plausibility of the generated counterexamples. We found high positive correlations between the number of counterexamples, their associative strength, and plausibility: Conditionals with many counterexamples have also more strongly associated and more plausible counterexamples. This supported the central assumption that in a memory structure with many stored elements successful counterexample retrieval is more likely.

The impact of the number of stored disablers and alternatives on inference acceptance is well established. However, there are no studies that have examined the impact on the inference latencies. Nevertheless, whether or not one has stored many disablers or alternatives has clear processing consequences. Of primary interest here is the fact that the memory studies of Conway and Engle (1994) indicated that the time course of a memory search process is affected by the number of elements that are retrieved from a memory structure: A semantic search process will take longer when the number of elements that are retrieved from a memory structure increases. If we assume that for conditionals with many disablers or alternatives more of these elements will be retrieved, we can expect that the counterexample search process will take more time for the 'many' than for the 'few' conditionals. Such a longer search process should result in longer inference times. This prediction was explored in a second experiment by recording both acceptance ratings and reaction times for the different inferences in a replication of Cummins (1995).

Results showed that AC inferences took more time for conditionals with many (vs. few) alternatives. MP latencies increased when many (vs. few) disablers were available. These findings are consistent with the assumption that the counterexample search lasts longer when the memory structures contain many counterexamples.

In a third experiment we looked at inter-individual differences in the efficiency of the search process. If the outcome of the retrieval process is a crucial component of the reasoning process, then inter-individual differences in the retrieval efficiency should affect the reasoning performance. We presented participants a standard conditional inference task. Afterwards, participants' disabler retrieval efficiency was measured by asking them to generate as much disablers as possible for a set of conditionals with only 30 s generation time for each conditional. The total number of generated disablers served as retrieval efficiency index. Remember that successful disabler retrieval results in an increased MP and MT rejection. As expected, a higher disabler retrieval capacity was related to a higher MP and MT rejection in the inference task. This finding illustrates the importance of the characteristics of the counterexample search process for a conditional reasoning theory: Inter-individual variation in the efficiency of the search process determines the kind of conclusions people draw.

3.3 Chapter 4: Every counterexample counts?!

In the study (De Neys, Schaeken, & d'Ydewalle, in press-b) described in Chapter 4 we tried to specify the 'stopping' characteristic of the counterexample search during everyday conditional reasoning. The research focused on the question whether or not the search process ends after retrieval of a single counterexample. In principle, it is possible that the search terminates after retrieval of only one counterexample and that other stored counterexamples are not taken into account for the inference evaluation. For example, in mental models theory (MMT) and Markovits' initial, MMT-based search process specification, the number of retrieved counterexamples is not important. What matters is that at least one counterexample is retrieved. This determines whether or not an additional model is constructed and consequently whether or not the inference is accepted. In MMT, inference acceptance is an all-or-nothing phenomenon. Either an additional model is constructed and the inference is completely rejected, or one sticks to the initial (and complementary) model and accepts the inference. There are no intermediate or graded states of inference acceptance.

After one successfully retrieves a single counterexample, the extra model is constructed, and the inference completely rejected. Retrieval of additional counterexamples can have no further impact on the inference acceptance. It follows that the search can stop after retrieval of a single counterexample.

We propose an alternative specification of the search process where the search process does not stop after retrieval of a single counterexample. Instead, the search would continue and an attempt is made to activate additional counterexamples. Here, the number of stored counterexamples and associative strength factors would not merely determine the probability that at least one counterexample will be retrieved, but rather the number of counterexamples that can be retrieved. The number of retrieved counterexamples would then determine the *degree* to which an inference will be accepted. The latency findings in Chapter 3 (De Neys et al., 2002) already lent some credence to the 'additional retrieval hypothesis'. It was assumed that the longer latencies for the 'many' conditionals resulted from additional counterexample retrieval. However, the findings were also open to alternative explanations. The experiments in Chapter 4 present a direct and more conclusive test of the hypothesis.

In a first experiment we examined the effect of additional counterexample retrieval on the inference acceptance by explicitly providing possible counterexamples. As in traditional suppression studies (Byrne, 1989; Byrne et al., 1999; Rumain et al., 1983), we simulated the effect of successful counterexample retrieval by explicitly presenting the counterexamples to participants. The crucial manipulation was that we varied the number of presented counterexamples. Each participant received five different conditionals with the number of presented counterexamples ranging from zero to four.

In a second experiment we tested the effect of additional counterexample retrieval without using an explicit presentation. A set of causal conditionals that varied in the number of possible disablers and alternatives (see Cummins, 1995; De Neys et al., 2002) was adopted. In a pretest we first assessed the number of alternatives or disablers a participant could retrieve for every conditional in the set. One month after the pretest, the same participants were re-invited for a reasoning task with the conditionals from the pretest. Thus, we exactly knew how many counterexamples every participant had stored for every conditional. We then looked at a participants' acceptance ratings (on a seven point rating scale) of the different inferences for each conditional in function of the number of counterexamples the participant had been able to retrieve for that specific conditional.

The Markovits-MMT based specification of the search process predicts a stepwise trend in the acceptance ratings in function of the number of counterexamples one has stored: Up to a certain number of available counterexamples (e.g., zero or one) inferences will tend to be accepted. After successful retrieval of a counterexample (e.g., when at least two counterexamples are stored), the inferences will be rejected, the search terminates, and additionally available counterexamples will not affect inference acceptance any further. The alternative specification we propose should result in gradually decreasing acceptance ratings with every additionally available counterexample.

Results basically supported the alternative search specification: AC acceptance linearly decreased with every additional alternative and MP acceptance linearly decreased with every additional disabler.

3.4 Chapter 5: Working memory and the retrieval and inhibition of stored counterexamples

The study described in Chapter 5 (De Neys, Schaeken, & d'Ydewalle, 2003) further characterized the counterexample retrieval by examining the possible role of working memory in the retrieval process.

Working memory (WM) is often conceived as a hierarchically organized system in which specific storage and maintenance components (i.e., the 'slave' systems or short-term memory systems, see Engle, Tuholski, Laughlin, & Conway, 1999) subserve a central component responsible for the control of information processing (e.g., Baddeley & Hitch, 1974; Cowan, 1995; Engle & Oransky, 1999). The crucial controlling component or 'central executive' consists of a limited-capacity system that regulates the allocation of attentional resources.

Memory studies have established that while some forms of memory retrieval are rather automatic and effortless, other forms demand executive WM-resources for their proper functioning (Kane & Engle, 2000; Moscovitch, 1994, 1995; Rosen & Engle, 1997). Moscovitch (1995) labeled these forms associative and strategic retrieval, respectively. The crucial characteristic of the strategic search is that it draws on WM (Rosen & Engle, 1997). Rosen and Engle posited that memory retrieval always starts with an associative, automatic spreading of activation. In case of a real strategic search, WM-resources would be used next for an active generation of cues to access new instances. The active cue generation would allow a much more efficient retrieval than the passive spreading of activation.

Markovits and Barrouillet (2002) only state that activation will automatically start to spread from the elementary representation of the conditional (maintained in WM) towards related elements in long-term memory. It is not addressed whether the retrieval also involves an active, strategic component. Nevertheless, some of Markovits' findings (e.g., Markovits & Quinn, 2002) do fit with the strategic claim. Our experiments present a direct test of the hypothesis that WM is involved in the retrieval of stored counterexamples.

If WM-resources are crucial for the counterexample retrieval, then efficiency of the retrieval should depend on the availability of WM-resources (i.e., the amount of resources that can be allocated to the active cue generation). This prediction was tested in Experiment 1. We developed a Dutch and group administrable adaptation of the OSPAN (La Pointe & Engle, 1990) to measure participants' WM-capacity (see De Neys, d'Ydewalle, Schaeken, & Vos, 2002). Participants in Experiment 1 were given the WM-capacity measure and a counterexample generation task. Results established that higher WM-capacity was associated with better counterexample retrieval.

Experiment 2 adopted a dual-task methodology to examine the causal nature of this relation. In order to eliminate alternative explanations we studied the impact of a secondary task load on the generation efficiency. If WM-capacity is involved in the counterexample retrieval, then burdening WM with a secondary task should reduce the efficacy of the retrieval process. A purely automatic retrieval process should not be affected by a WM-load.

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

Results showed that the retrieval efficiency declined when WM was burdened with the attention demanding, complex tapping task, while the undemanding cascade tapping had no effect on the number of generated counterexamples. This established that counterexample retrieval involves a WM-dependent, strategic component.

In a third experiment we compared the performance of a group of low and high spans (participants in the bottom and top quartile of first-year psychology students' WM-capacity distribution, respectively) in an actual reasoning task. The first experiments established that people higher in WM-capacity are more successful at retrieving counterexamples. Our previous research already established that more successful alternative retrieval leads to lower acceptance ratings of the AC and DA inferences in a reasoning task. Therefore, if high spans are indeed more efficient at retrieving alternatives, one expects 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, note that 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. There might be a dissonance between searching disablers for the MP and MT inferences and the valid status of these inferences.

All people are assumed to have a basic "contextualisation" tendency to search stored counterexamples associated with the reasoning problem. However, individual difference studies indicate that at least people of higher cognitive capacity (e.g., high spans) also appear to have a logical, "decontextualisation" tendency: A basic ability to put background knowledge aside when it conflicts with the logical standards (e.g., Klaczynski, 2001a; Stanovich & West, 2000). If high spans would have an elementary notion of logical validity, this should conflict with the tendency to search disablers. Based on this assumption we hypothesized that high spans would use their WM-resources for an active inhibition of the spontaneous disabler search. Note that the inhibition of responses deemed inappropriate is considered as one of the key executive functions (e.g., Baddeley; 1996; Engle, Tuholski, Laughlin, & Conway, 1999; Miyake & Shah, 1999; Shallice & Burgess, 1993). Thus, when it concerns retrieving disablers, high spans would not use their WM-resources for an active search but rather for an inhibition of the automatic disabler retrieval. Despite the better intrinsic retrieval capacities for high spans, this inhibition process should result in higher MP and MT acceptance ratings for the high (vs. low) spans. Results indeed showed that AC and DA acceptance ratings were highest for the low spans, whereas MP and MT acceptance ratings were highest for the high spans.

Although Experiment 3 supported the predictions, the evidence remained correlational and open to alternative explanations. Experiment 4 provided a direct test of the hypothesized role of WM in the retrieval and inhibition of counterexamples. We tested the effects of a secondary working memory load (tapping the complex finger pattern of Experiment 2) on reasoning performance. 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. Results showed that low spans' acceptance of all four inference types increased when working memory was burdened by the complex tapping task. This supports the hypothesis that WM-capacity is important for the retrieval of counterexamples in everyday reasoning.

For high spans the load effect interacted with the type of inference. The working memory load increased 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 results in 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 at retrieval. When this retrieval becomes less efficient under load, MP and MT will be more accepted.

3.6 Chapter 6: Working memory span and retrieval: A trend analysis

The final study, described in Chapter 6, presents further evidence for a WM-dependent mediation of the counterexample retrieval and inhibition. Our previous study (De Neys, Schaeken, & d'Ydewalle, 2003; Chapter 5) indicated that when it concerns searching disablers, participants highest in cognitive capacity use WM-resources to inhibit the counterexample activation. Note that we explicitly assumed that the inhibition would only occur for people highest in WM-capacity. However, we only compared a group of participants from the bottom and top quartile of the WM-capacity distribution. If the assumption that the inhibition occurs only for people highest in WM-capacity is correct, it follows that people with medium WM-capacities should show the lowest MP and MT acceptance. Indeed, on one hand medium spans (vs. high spans) should not inhibit the disabler retrieval. On the other hand, medium spans will have a more efficient counterexample retrieval than low spans because they can allocate more resources to the search. Thus, the disabler retrieval during conditional reasoning should be most successful for medium spans. Consequently, it is expected that the MP and MT acceptance ratings in function of WM-capacity follow a U-shaped curve: Due to the limited resources, people lowest in WM-capacity will not be very successful at retrieving disablers and should therefore show rather high levels of MP and MT acceptance. Because of the more efficient disabler retrieval, MP and MT acceptance should decrease for the medium spans. Because of the disabler inhibition, MP and MT acceptance ratings should increase again for reasoners higher in WM-capacity.

Since retrieving alternatives results in the rejection of AC/DA inferences and accepting AC/DA is erroneous in standard logic, there is no conflict between a basic logical notion and the retrieval of alternatives. More precisely, it is assumed that the basis of high spans' disabler inhibition is a minimal notion of the logical principle that the truth of the antecedent implies the truth of the consequent. While this notion conflicts with the possibility that the consequent does not occur when the antecedent occurs (i.e., a disabler), it does not conflict with the possibility that the consequent occurs in the absence of the antecedent (i.e., an alternative). Thus, the process where alternatives are retrieved from long-term memory should not be inhibited. Therefore, the higher WM-capacity is, the more efficient the alternative retrieval will be, and the less AC and DA should be accepted. Contrary to MP and MT, AC and DA acceptance ratings should therefore follow a negative linear trend in function of WM-capacity. These predictions were tested in Experiment 1. A large sample of participants were given a measure of WM-capacity and a conditional reasoning task with everyday, causal conditionals. By examining the specific trends in the acceptance ratings of the different inferences over the whole WM-capacity distribution the trend predictions could be tested.

Results of Experiment 1 confirmed the predicted trends. Experiment 2 and 3 further generalized the findings by showing that the quadratic MP trend could even be replicated when a disabler was explicitly presented in the reasoning task. These findings underlined the robustness of the inhibition phenomenon.

3.6 Chapter 7: Where do we stand? Where do we go from here?

The concluding chapter wraps up the findings by sketching a model of the counterexample retrieval process based on the established search specifications. Guidelines for further research are suggested and some broader implications of the findings are highlighted.

CHAPTER 2

Strength of association: The disabling condition case

Cummins (1995) showed that reasoning with conditionals involving causal content is affected by the relative number of available alternative and disabling conditions. More recent evidence (Quinn & Markovits, 1998) indicates that, beside the number of stored conditions, the relative strength of association of the alternative conditions with the consequent term is another important factor that affects conditional reasoning. In this study we examined the effect of the strength of association for the disabling conditions. We identified causal conditionals for which there exists only one highly associated disabler. With these conditionals we constructed conditional inference problems in which the minor premise was expanded with the negation of a strongly or weakly associated disabler. Results of two experiments indicate that strength of association of stored disabling conditions is affecting reasoning performance: Acceptance of Modus Ponens and Modus Tollens increased when there was no strongly associated disabler available.

Based on: De Neys, W., Schaeken, W., & d'Ydewalle, G. (in press-a). Causal conditional reasoning and strength of association: The disabling condition case. European Journal of Cognitive Psychology.

INTRODUCTION

Cognitive psychologists studying human reasoning have devoted a great deal of research to conditional reasoning. This kind of reasoning consists in making inferences on the basis of 'if p then q' sentences. In a standard conditional inference task people are asked to assess arguments of the following four kinds:

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

Under the material implication interpretation of propositional logic, MP and MT are considered valid inferences while DA and AC are regarded as fallacies. Much of the work on conditional reasoning has tried to identify the factors that influence performance on these four problems (for a review, see Evans, Newstead, & Byrne, 1993).

A growing body of evidence is showing that peoples knowledge about the relation between the p (antecedent) and q (consequent) part of the conditional has a considerable effect on the underlying reasoning process (e.g., see Byrne, 1989; Byrne, Espino, & Santamaria, 1999; Markovits, 1984; Newstead, Ellis, Evans, & Dennis, 1997; Rumain, Connell, & Braine, 1983; Thompson, 1994).

In the case of reasoning with conditionals involving causal content (e.g., 'If cause p, than effect q') seminal work has been done by Cummins and her colleagues (1995; Cummins, Lubart, Alksnis, & Rist, 1991). Following Byrne (1989), Cummins examined the effect of the alternative and disabling conditions of a causal conditional. An alternative condition is a possible cause that can produce the effect mentioned in the conditional while a disabling condition prevents the effect from occurring despite the presence of the cause. Consider the following conditional:

If the brake is depressed, then the car slows down

Possible alternative conditions for this conditional are:

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

The occurrence of these conditions will result in the car slowing 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 the 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 it is not sufficient to depress the brake in order to slow down the car. There are additional conditions that have to be fulfilled.

When people (fallaciously) accept DA and AC inferences, they fail to see that there are other causes that may lead to the occurrence of the effect beside the original stated one. Cummins (1995) and Cummins et al. (1991) found that peoples acceptance of DA and AC inferences decreased for conditionals with a high number of possible alternative conditions. This showed that a crucial factor in making the fallacious inferences is the number of alternative causes people can think of. In addition, she found that the number of disabling conditions affected the acceptance of the valid MP and MT inferences: If there were many conditions that could disable the relation between antecedent and consequent, people also tended to reject the valid inferences.

Recently, Quinn and Markovits (1998) have identified another factor that may influence reasoning with causal conditionals. They showed that not only the number of alternative conditions is important, but also what they call the 'strength of association' of the alternative conditions. Quinn and Markovits developed a framework (see also Markovits, Fleury, Quinn, & Venet, 1998) where reasoning performance is being linked to the structure of semantic memory. In this framework it is assumed that, when confronted with a causal 'if p then q' conditional, reasoners will access a causal structure in semantic memory that corresponds to 'ways of making q happen' (i.e., alternative conditions). Within the structure, there will be causes that will be more strongly associated with q than others. The more strongly associated a specific cause is, the higher the probability that it will be retrieved by the semantic search process.

Quinn and Markovits (1998) measured strength of association by frequency of generation: In a pretest, participants were asked to write down as many potential causes for a certain causal consequent (effect, e.g., 'a dog scratches constantly'). This allowed the construction of conditionals with a strongly (e.g., 'If a dog has fleas, then it will scratch constantly') and weakly (e.g., 'If a dog has skin disease, then it will scratch constantly') associated cause. Because the consequent is the same in both conditionals, the number of possible alternative conditions is kept constant.

Quinn and Markovits (1998) reasoned that with both the 'strong' and 'weak' type of conditional, people would try to activate and retrieve 'alternative ways of making q happen'. However, reasoners given the weak conditional will be able to activate the strongly associated cause, while for the strong conditional they will have to activate some other, less closely associated term. Thus, it will be more difficult to retrieve an alternative condition in case of the strong conditional, which would lead to a greater acceptance of DA and AC inferences. The results of the study were consistent with the predicted response pattern.

The identification of the strength of association effect raises the question whether this effect is also present for the disabling conditions. Indeed, although knowledge of disabling conditions is also stored in semantic memory, Quinn and Markovits (1998) restricted their case to an analysis of the alternative conditions. Cummins (1995) already showed that both the number of alternatives and disablers is affecting reasoning performance. In addition, Elio (1998) has shown that the process of disabler retrieval is not only important in conditional reasoning but also in the field of belief revision and non-monotonic reasoning: Belief in a conditional after contradiction was lower when people could find many disablers. Thus, both for reasoning psychologists and the psychological and AI community studying belief revision, examining the effect of associative strength of disablers can identify a new factor affecting the crucial disabler retrieval. The present study focused on this topic.

The framework developed by Quinn and Markovits (1998) was adopted and extended to the disabling conditions. It was assumed that when presented a causal conditional, people will not only access a causal structure with alternative conditions but also one that corresponds to 'ways that prevent q to occur' (see Markovits 2000; Vadeboncoeur & Markovits, 1999). When such disabling conditions are retrieved, p will no longer be perceived as a sufficient condition for q what renders the MP and MT conclusions uncertain.

In a generation task we identified strongly and weakly associated disablers for a number of conditionals. We constructed experimental items by expanding the original antecedents of the conditionals with the negation of the strongly or weakly associated disabler. Suppose that for a certain conditional we find that S is a strongly associated disabler, while W is a weak one. This allows the construction of the expanded conditionals: 'If P and not S, then Q' (strongly expanded conditional) and 'If P and not W, then Q' (weakly expanded conditional). These expanded conditionals have an equal number of possible disablers (i.e., the original number minus one). However, reasoners presented 'If P and not W, then Q' will still be able to activate the strongly associated disabler S, while with 'If P and not S, then Q' they will have to activate a less closely associated one. Thus, it will be harder to access and retrieve disablers for the strong conditionals. This access-to-disablers manipulation rests solely on the strength of association of the disablers and not on the number of accessible disablers.

Retrieving disablers from semantic memory will decrease the acceptance of MP and MT inferences. Therefore, we predict that acceptance ratings for MP and MT inferences will be higher for the strongly expanded conditionals than for the weakly ones.

In the present experiment we did not manipulate the access to alternative conditions. The classical findings of Cummins (1995) indicate that retrieving disablers has no effect on DA and AC. Given these findings, one might expect that the access-to-disablers manipulation will have no effect on DA and AC acceptance.

EXPERIMENT 1

Pretest

The material for the present experiment was selected from previous pilot work (see De Neys, Schaeken, & d'Ydewalle, 2000), where 20 participants wrote down as many disabling conditions as possible for a set of 20 causal conditionals (with 1.5 min generation time for each conditional). The set included the 16 conditionals from Cummins (1995, Experiment 1) and four additional ones. The generation protocols were scored by 2 independent raters in order to identify unrealistic items and items that were simple variations of a single idea (e.g., for the example above 'skid due to water on road', 'skid due to mud', 'skid due to ice on road'). Fewer than 5% of the generated disablers were disallowed by the raters (interrater reliability was .83).

For every conditional we established the relative frequency of appearance of the disablers that participants wrote down. We needed conditionals with a set of disablers in which there was one specific disabler that was very frequently generated. The expanded conditionals manipulation also forced us to take an additional criterion into account. We could not allow disablers that express a quantification of the original antecedent (e.g., 'brake not depressed hard enough'). Expanding the original with this kind of disablers would result in inconsistencies for some problems (e.g., DA, 'The brake was not depressed and the brake was depressed hard enough'). We selected 3 conditionals that met these criteria. From each set of disablers one infrequently generated disabler was selected. This weakly associated disabler had to meet the non-quantification criteria. Furthermore, if the strongly expanded conditional contained an explicit negation (e.g., 'If the apples are ripe and they are not picked'), we opted to express the selected weakly associated disabler in an explicit negated way too. The negation criterion should guarantee that the strongly and weakly expanded conditionals have comparable lexical complexity. Finally, the selected disablers had to sound as natural (according to our intuitions) as possible (e.g., 'not too little wind' was not accepted). Table 1 presents the material that was selected for the experiment.

Table 1

Relative Frequency of Generation of the Most Frequently Mentioned Disablers for the Three Selected Conditionals

If the apples are ripe, then they fall from the tree

Picked (65%)
Too little wind (25%)
Not enough weight (20%)
Not ripe enough (20%)
Apples caught in branches (10 %)

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

Hands not greasy (50%)
Grasped glass with palms only (35%)
Prints wiped off (30%)
Glass was wet (25%)

If water is heated to 100° C, then it boils

No pure water (75%)
No normal pressure (30%)
Bad temperature measure (30%)

Note. The disablers are given in order of frequency (%). Selected strongly and weakly associated disablers are highlighted.

One could remark that our pretest allowed 90 s generation time (as in Cummins, 1995), while Quinn and Markovits (1998) allowed only 30 s. It could be argued that this might confound the strength of association classification. However, De Neys et al. (2000) also used a 30 s generation task for some of the conditionals. When generation frequency for the disablers in the 1,5 min and 30 s generation task was compared, results indicated that although the shorter generation time decreased the absolute frequency of generation, the crucial relative ranking of the disablers was not affected[1].

We also note that the pilot study supplemented Quinn and Markovits (1998) by addressing some possible reservations about the use of generation frequency as strength of association measure. We looked at the relation between generation frequency and other possible associative strength measures such as plausibility and generation order. After the generation task we asked participants to judge the plausibility of the generated disablers. We also calculated the probability that a certain disabler was generated as the first one for a specific conditional. It was shown that the frequency of generation measure correlated with the other strength of association measures: More frequently generated disablers were judged more plausible and tended to be generated first[2].

Method

Participants

89 first-year university students participated in the experiment.

Material

Participants received a 4-page booklet. Page one included the instructions for the task. On the top of each of the next three pages appeared the selected conditionals. Each conditional was embedded in the four inference types (MP, DA, MT, AC). So, each of the three pages included one conditional with four inference problems. For each conditional there was a specific presentation order of the four inferences (AC, MT, DA, MP or MP, MT, DA, AC or MP, DA, MT, AC). The three pages were bound into booklets in randomized order.

[1] Generation frequency of the disablers generated for 8 different conditionals could be compared. The Spearman Rank Order Correlation reached .84 [t(82) = 13.71, p<.0001].

[2] For every disabler generated for a specific conditional we calculated the mean plausibility rating and the overall (# generated first/ # participants) and relative (# generated first/ # generated) probability that this disabler was the first generated one. Frequency of generation was associated with plausibility [Spearmann Rank Order test, r_s = .36, t(162) = 4.83, p<.0001; r_s = .50, t(35) = 3.41, p<.005, when only disablers generated by at least 50% of participants were considered] and overall [r_s = .63, t(162) = 10.22, p<.0001] and relative [r_s = .40, t(162) = 5.52, p<.0001] probability that the disabler is generated first.

Below each inference problem appeared a seven point rating scale. This resulted in the following item format:

Rule: If water is heated to 100°C, then it boils

Fact: The water is heated to 100°C and the water is pure
Conclusion: The water boils

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

This is an example of the MP problem. On the same page participants would also find the MT, DA and AC problem. The access to disablers manipulation consisted in the presentation of two different minor premises (the information under the heading 'Fact'); the above example would belong to the strongly associated group were the original information was expanded with the negation of the strongest associated disabler. Similarly, in the weakly associated group, the negation of the selected weakly associated disabler was added to the 'Fact:'-information.

Table 2

Different Content in the Groups of Experiment 1

Expanded strongly associated:

(a) Water is heated to 100°C and the water is pure
(b) The apples are ripe and they are not picked
(c) John grasps the glass with his bare hands and his hands are greasy

Expanded weakly associated:

(a) Water is heated to 100°C and the pressure is normal
(b) The apples are ripe and they are not caught in the branches
(c) John grasps the glass with his bare hands and the glass is dry

Note. Both groups only differ by the kind of information that is presented in the minor premise. This is an example of the minor premises for the MP problem.

In both expanded groups, the original conditionals appeared on top of the item pages. Thus, participants were not presented explicit expanded conditionals but rather conditional inference problems with expanded minor premises. All the items in a single booklet belonged to the same group. Table 2 gives an overview of the different material in the two groups (for an MP problem).

Procedure

The booklets were randomly given out to students who agreed to participate in the experiment. No time limits were imposed. The instructions that were presented on the first page of the booklet were read aloud by the experimenter. The instructions explained the specific item format of the task. Participants were told that the task was to decide whether or not they could accept the different conclusions. The instruction page showed an example problem (always standard MP) together with a copy of the rating scale. Care was taken to make sure that participants understood the precise nature of the rating scale. As in Cummins (1995), participants were NOT specifically instructed to accept the premises as always true and to endorse only conclusions that follow necessarily. With Cummins we assume that this encourages people to reason as they would in everyday circumstances. However, one should note that strictly speaking the task is therefore not a deductive inference task (see Evans, 2000). Thus, accepting the MP/MT and AC/DA inferences should not be considered correct or incorrect reasoning. When we refer to the standard nomenclature, a nominalist stance is adopted towards the use of the terms 'valid inferences' and 'fallacies'.

Results and discussion

Participants rated each of the four inference types three times. For every inference type the mean of these three ratings was calculated. This resulted in a 4 (inference type, within-subjects) x 2 (group, between-subjects) design. All hypotheses were tested with planned comparison tests and rejection probability of .05.

Table 3 shows the overall mean acceptance ratings for the four inference types in the expanded weakly and strongly associated group. Acceptance ratings in both groups differed significantly [F(1, 87) = 4.55, MSe = 3.85, p < .04]. As expected, we obtained higher MT [F(1, 87) = 4.99, MSe = 2.67, p < .03] ratings in the strongly associated group, where the strongly associated disabler was eliminated. A similar tendency was observed for the MP inference, but the effect did not reach significance.

In line with Cummins' findings both expanded groups did not significantly differ in terms of AC and DA acceptance. However, we neither observed a significant interaction between the different inferences types. All inferences tended to be accepted more when no strongly associated disabler was available. Thus, a possible impact of the strength of association of stored disablers on the fallacies cannot be discarded.

Table 3

Mean Acceptance Rating for the Four Inference Types in the Strongly Associated and Weakly Associated Groups

Inference type	Group	
	Expanded weakly associated (n=45)	Expanded strongly associated (n=44)
MP	5.7	5.92
DA	4.78	5.11
MT	4.37*	5.14*
AC	4.98	5.44

Note. The rating scale ranged from 1 (very sure cannot draw this conclusion) to 7 (very sure can draw this conclusion) with 4 representing cannot tell. * planned contrast, p<.05.

The MT findings do provide some preliminary support for the hypothesis that retrieving disablers from semantic memory is affected by their strength of association. As predicted, peoples acceptance of MT inferences increased when there was no strongly associated disabler available.

We suspect that the non-significance of the expected effect on MP may be due to some specific task characteristics of the experiment. In the pretest, relatively few disablers were generated (less than the overall mean) for the three conditionals that were adopted for the experiment. Cummins (1995) already obtained high MP acceptance ratings for these conditionals. The 'expansion' manipulation in the present experiment then further decreased the available number of disablers. This may have resulted in a ceiling effect. It could be the case that MP acceptance was already at the top in the weakly associated group. Mean acceptance for MP in the weakly associated group (Mean = 5.7, see Table 3) indeed tended to the 'sure that I can draw this conclusion' rating, located at the upper end of the scale. As in

Cummins (1995), acceptance ratings on the (more difficult) MT inference were lower, what allowed the associative strength effect to show up.

Since there is a strong tendency to accept MP in both expanded groups, it is also possible that the seven point rating scale was not sensitive enough to detect an effect. In addition, the strength of association factor was manipulated between groups of subjects which may have further hampered the detection of the MP effect.

In order to avoid these possible problems we decided to look at the effect of the strength of association factor in a second experiment.

EXPERIMENT 2

Experiment 2 further examined the effect of the strength of association of the disabling conditions. We tried to avoid the highlighted problems in Experiment 1 by making a number of changes to the materials and design used there. First, in the pretest, we looked for conditionals with a large number of disablers (higher than the overall mean). Since there are more disablers available, acceptance for these conditionals will be lower, what should reduce an eventual ceiling effect. Second, by using an 11-point rating scale instead of the seven point scale in Experiment 1 we tried to offer participants the possibility for a more detailed discrimination at the extreme levels of the scale. Finally, all participants received both a strongly and weakly expanded conditional (with different content) in order to allow a within subjects test of the strength of association effect.

Participants were presented MP, MT and AC problems with expanded minor premises based on the selected conditionals. Since the problems possibly resulting in the non-significance of the MP effect in Experiment 1 were avoided, we expected to find an effect of the associative strength manipulation on both MP and MT.

The AC problem was included in order to further examine the effect of associative strength of disablers on the fallacies. In Experiment 1 we found no significant effects on the AC and DA inferences. However, we neither observed a significant interaction between the different inference types. Especially for AC, acceptance ratings tended to increase when no strongly associated disabler was available. By presenting participants AC inferences in addition to the MP and MT inferences we could examine the generality of this trend.

Method

Participants

37 first-year psychology students volunteered to participate in the experiment

Material

For the strength of association manipulation it was crucial to find conditionals with a set of disablers in which there is only one which is very frequently generated. Although most conditionals with many disablers in the pretest had typically also a larger number of very frequently generated (> 50 % of participants) disablers, we could still select two conditionals (i.e., 'If Jenny Turns on the air conditioner, then she feels cool' and 'If the brake is depressed, then the car slows down') that were appropriate for the present purpose. The most frequently generated disabler was mentioned by 95% and 85% of the participants, while the second most frequently generated disabler was mentioned by only 50% and 45% of the participants, respectively. Both conditionals have also a comparable number of possible alternative conditions (see Cummins, 1995; De Neys et al., 2000). Strongly and weakly associated problem versions were constructed by expanding the minor premise with the negation of the strongest and second strongest associated disabler, respectively. The different versions are presented in Table 4 (for an MP problem).

Table 4

Material for Experiment 2

Expanded strongly associated:

(a) Jenny turns on the air conditioner and the air conditioner is not broken (95%)
(b) The brake is depressed and the brake is not broken (85%)

Expanded weakly associated:

(a) Jenny turns on the air conditioner and she has no fever (50%)
(b) The brake is depressed and the road is not slippery (45%)

Note. The generation frequency (%) of the corresponding disablers in the pretest is presented between brackets.

As in Experiment 1, the original conditional was presented on top of an item page. The expanded MP, MT and AC inferences were then presented on the same page. Below each inference problem appeared the following 11-point rating scale:

					I					
---5---	---4---	---3---	---2---	---1---	---0---	---1---	---2---	---3---	---4---	---5---
Very Sure				Some what	I I I	Some what				Very Sure
That I CANNOT draw this conclusion					I					That I CAN draw this conclusion

Participants received a 3-page booklet with instructions on the first page. On the second and third page appeared the two selected conditionals, one with strongly expanded inferences problems and the other with weakly expanded ones. In half of the booklets the strongly expanded inference problems were presented first, while the order was reversed for the other half. Each conditional was used for the weakly expanded inference problems in approximately half of the booklets and for the strongly expanded problems in the other half.

Procedure

The procedure was similar to that used in Experiment 1.

Results and discussion

An ANOVA was performed on the acceptance ratings with inference type (MP, MT, AC) and strength of association (weakly or strongly expanded) as within-subject factors. Due to the within-subjects manipulation of the strength of association factor, each participant rated both weakly and strongly expanded inferences. Although the three strongly and weakly expanded inferences were based on different conditionals, it is still possible that the order in which the strength of association manipulation was presented in the booklets (weakly expanded inference problems presented before or after the strongly associated ones), biased the ratings. Therefore, presentation order was entered as a between-subjects factor in the ANOVA.

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

Results showed that the inference type and strength of association factor interacted significantly [F(2,70) = 3.66, MSe = 2.22, p<.05]. Planned contrast tests indicated that the acceptance ratings for MT [F(1,35) = 4.54, MSe = 3.76, p<.05] were higher in the strongly associated group, while the strength of association factor had no effect on AC acceptance ratings (see Table 5). We also found a significant effect of the strength of association manipulation on the MP acceptance ratings: MP was more accepted when there was no strongly associated disabler available [F(1,35) = 12.41, MSe = .53, p<.005].

Table 5

Mean Acceptance Ratings for the Strongly and Weakly Expanded Inference Problems

Inference type	Expansion type	
	Expanded weakly associated	Expanded strongly associated
MP	3.78**	4.38**
MT	2.03*	2.97*
AC	2.76	2.43

Note. The rating scale ranged from -5 (very sure cannot draw this conclusion) to 5 (very sure can draw this conclusion) with 0 representing can't tell. * planned contrast p<.05, ** p<.01.

The presentation order factor did not affect the strength of association effect for any of the three inference types, what confirms that the results were not biased by the order in which participants rated the expanded conditional inferences.

One might note that, although not significant, AC acceptance in Experiment 2 tended to be somewhat lower in the expanded strongly associated group. In Experiment 1 however, there was a trend in the opposite direction with higher AC and DA acceptance when there was no strongly associated disabler available. Detailed examination of these (reversed) AC and DA trends indicated they could be attributed to a complication due to the specific nature of the presented thematic material.

The complication lies in the fact that some disablers can also qualify as possible alternative (e.g. see Manktelow and Fairley, 2000). In Experiment 1, this was the case for one of the strongly associated disablers we eliminated (i.e., 'If the apples are ripe and they are not picked, then they fall from the tree'). In an alternatives generation task (see De Neys et al., 2000), 55% of the participants generated 'apples dropped by picker' as possible alternative. Now, if the apples are not picked, they cannot be dropped by the picker. Thus, the manipulation also eliminated a possible alternative. Therefore, the number of available alternatives was smaller in the expanded strongly associated group, what should result in slightly increased DA and AC acceptance. For the other presented disablers the complication was not present. In fact, when the conditional 'If the apples are ripe, then they fall from the three' was removed from the analysis in Experiment 1 the trends on AC and DA dissolved, while the trends on MP and MT were unaffected[3]. In Experiment 2, a similar complication was present but now in the expanded weakly associated group. One of the presented disablers (i.e., 'If Jenny turns on the airco and she has no fever, then she feels cool') indeed qualified as possible alternative. In an alternative generation task 35% of participants generated 'having cold fever' as possible alternative. Consequently, in Experiment 2, the number of available alternatives was somewhat smaller in the expanded weakly associated group. Thus, the DA and AC trends in Experiment 1 and 2 can be explained by an impact on the availability of alternatives. In line with what Cummins (1995) found for the number of disablers, this indicates that associative strength of a stored disabler has no effect on the fallacies per se.

GENERAL DISCUSSION

The experiments in this study allow us to conclude that MP and MT acceptance increases when there is no strongly associated disabler available. The results from the second experiment established that, as expected, the manipulated availability of disabling conditions affected both MP and MT acceptance. These results support the hypothesis that in addition to the number of disabling conditions (Cummins, 1995), retrieving disablers from semantic memory is affected by their strength of association.

Taken together, the outcome of the two experiments suggests there is no impact of the associative strength of stored disablers on the DA and AC inferences. The observed trends could be attributed to an impact of the disabler manipulation on the availability of the alternatives. This is consistent with Cummins (1995) findings, where the number of stored disablers had no effect on AC or DA.

Nevertheless, one should note that deviations of Cummins' findings have been reported. Some studies (e.g. Liu, Lo, & Wu, 1996; Markovits & Potvin, 2001, Experiment 3) did report an impact of disabler retrieval on AC and DA acceptance. This would not be surprising when the outcome of the disablers search process could affect the efficiency of the alternatives search process. In that case one could hypothesize that successful disabler

[3] Mean acceptance ratings in strongly vs. weakly expanded group: MP = 6.08 vs. MP = 5.77, DA = 5.60 vs. DA = 5.60, MT = 5.38 vs. MT = 4.66, AC = 5.96 vs. AC = 5.80.

retrieval will be resource demanding what might burden the search for alternatives. However, at present, such a possible interplay between the search processes has not yet been examined. Together with the 'disablers qualifying as alternative' complication this makes it clear that more research is needed to establish the precise relation between the disablers and alternatives retrieval processes. In the absence of such more specific studies, the present DA and AC conclusions should be interpreted with some caution.

In this study we adopted Quinn and Markovits' (1998) notion of a semantic search process and extended it to the disabling conditions. We should note that Quinn and Markovits (see also Markovits et al., 1998) incorporated the postulated semantic search process in the mental models theory (Johnson-Laird and Byrne, 1991). They argued that successful retrieval of an alternative would lead to the construction of an additional [not-P Q] model. This model would represent the possibility that the effect specified in the conditional occurs without the occurrence of the specified cause. Therefore, AC and DA would no longer be supported.

The above account can be easily extended to incorporate the present findings. When the semantic search process retrieves a stored disabler an additional [P not-Q] model would be constructed. This model will represent that it is possible that occurrence of the antecedent will not result in the occurrence of the consequent. With such a model MP and MT will not longer be supported. Since the probability of successful retrieval is lower when there is no strongly associated disabler available, people will be less likely to construct the additional [P not-Q] model. This would account for the higher MP and MT acceptance. In mental models theory, the AC and DA inferences are only affected by models where the consequent does occur [Q] or the antecedent does not occur [not-P]. Thus, whether or not the [P not-Q] model is constructed is not important for AC and DA. This would explain why the associative strength of disablers has no effect here.

However, it is important to notice that the present study only focused on the semantic search process during conditional reasoning. As Quinn and Markovits (1998), we examined a factor that affects successful memory retrieval. This research does not address how the retrieved disabler is actually incorporated in the reasoning process. We therefore refrained from making specific claims about the nature of the basic inferential principles (i.e., mental models or mental inference rules). The general semantic search process can be incorporated in other reasoning theories like mental logic (Braine & O'Brien, 1998; Rips, 1994) or the probabilistic approach (Oaksford & Chater, 1998; Oaksford, Chater, & Larkin, 2000).

In the probabilistic approach, for example, acceptance of MP and MT depends upon the value of an 'exceptions parameter' (i.e., the probability of not-q given p, see also

Stevenson & Over, 1995). This parameter represents the probability that exceptions (disablers) will occur. The higher the exceptions value the less MP and MT will be accepted. Now, it is reasonable to assume that a reasoner determines this probability by searching his/her memory for known exceptions. Thus, one can also predict that successful disabler retrieval, by increasing the exceptions parameter, will decrease MP and MT acceptance. Comparing these different implementations is not within the scope of the present experiment or the Quinn and Markovits study.

We mentioned the relevance of the present study for the work of Elio (1997, 1998) and other researchers in the domain of belief revision and non-monotonic reasoning. Elio established that the number of stored disabling conditions affected peoples belief revisions and stated that conditional reasoning and belief revision are guided by the same memory search process. Our results show that successful retrieval is not only affected by the number of stored disabling conditions but also by their strength of association.

Finally, the present study can be related to the work of Chan and Chua (1994). They examined the effect of 'relative salience' of disabling conditions. This factor can be interpreted as strength of association. Chan and Chua presented participants inference problems with two conditionals (e.g., 'If p then q, If r then q, p, thus q?'). The second conditional mentioned a possible disabling condition while the categorical premise was not expanded (see Byrne, 1989). Acceptance of MP and MT decreased with the strength of association of the mentioned disabler. However, a crucial difference with our study is that the present manipulation specifically affected the retrieval of disablers from semantic memory. In Chan and Chua's experiment, reasoning was affected by the strength of association of the mentioned disabler per se. The expansion of the categorical premise in the present experiment eliminated a strongly or weakly associated disabler and thereby affected the strength of association in the residual disabler set.

In sum, our study indicated that the conditional inferences people draw are influenced by the strength of association of the disabling conditions. This complements Quinn and Markovits' (1998) contention that the strength of association of elements in semantic memory is an important factor in predicting conditional reasoning performance.

CHAPTER 3

Testing the semantic memory framework

This study tests and refines a framework that proposes a mechanism for retrieving alternative causes and disabling conditions during reasoning. Experiment 1 examined the relation between different factors affecting retrieval. The test revealed high correlations between the number of possible alternative causes or disabling conditions and their strength of association and plausibility. Experiment 2 explored the hypothesis that due to a more extended search process, conditional inferences would last longer when many alternative causes or disabling conditions were available. Affirmation of the Consequent (AC) and Modus Ponens (MP) latencies showed the hypothesized pattern. Denial of the Antecedent (DA) and Modus Tollens (MT) inferences did not show latency effects. The experiment also identified an effect of the number of disabling conditions on AC and DA acceptance. Experiment 3 measured efficiency of disabler retrieval by a limited time, disabler generation task. As predicted, better disabler retrieval was related to lower acceptance of the MP and MT inferences.

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

INTRODUCTION

Conditional reasoning involves making inferences on the basis of an 'if-then' relation and is considered one of the cornerstones of human reasoning. Research on conditional reasoning has been trying to identify the factors and processes that affect the performance on these 'if-then' inference problems. In a standard conditional inference task people are asked to assess arguments of the following four kinds:

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 propositional logic, MP and MT are considered valid inferences while DA and AC are regarded as fallacies.

A growing body of evidence is showing that people's knowledge about the relation between the p (antecedent) and q (consequent) part of the conditional has a considerable effect on the reasoning process. In particular, the role of knowledge of alternative causes and disabling conditions has attracted interest (see Politzer, in press, for a review).

An alternative cause (alternative) is a possible cause that can produce the effect mentioned in the conditional, whereas a disabling condition (disabler) prevents the effect from occurring despite the presence of the cause. Consider the following conditional:

If the air conditioner is turned on, then you feel cool

Possible alternative causes for this conditional are:

Taking off some clothes, the weather cools, swimming...

The alternatives make it clear that it is not necessary to turn on the air conditioner in order to feel cool. Other causes are also possible.

Possible disabling conditions are:

Air conditioner is broken, having fever, window open…

If such disablers are present, turning on the air conditioner will not result in feeling cool. The disablers make it clear that it is not sufficient to turn on the air conditioner in order to feel cool. Additional conditions must be fulfilled.

Rumain, Connell, and Braine (1983) showed that when a possible alternative was explicitly presented to participants, the AC and DA inferences were less endorsed. Byrne (1989) found a similar effect on MP and MT when a possible disabling condition was mentioned. In addition, using familiar relations (e.g., If an animal has feathers, it is a bird) for which people have ready access to alternatives, Markovits (1986) showed that even without explicit presentation, awareness of a possible alternative decreased the number of AC and DA inferences.

Cummins and colleagues (1995; Cummins, Lubart, Alksnis, & Rist, 1991) have conducted further seminal work on the impact of alternatives and disablers in conditional reasoning. Cummins (1995) and Cummins et al. (1991) directly addressed the role of stored knowledge of alternatives and disablers by examining the effect of the number of available alternatives and disablers. In a pretest they identified causal conditionals (i.e., conditionals that express a causal relation) for which participants generated many or few alternatives and disablers. These conditionals were then adopted for a conditional reasoning task. Results showed that people's acceptance of DA and AC inferences decreased for conditionals with many possible alternatives. In addition, the number of disabling conditions affected the acceptance of the MP and MT inferences: If there were many conditions that could disable the relation between antecedent and consequent, people tended also to reject these valid inferences. Alternatives and disablers were not explicitly presented, showing that a crucial factor in causal conditional reasoning is the number of alternative causes and disabling conditions people can think of.

Cummins (1995) argued that finding possible alternative causes and disabling conditions affects peoples interpretation of the necessity and sufficiency of a cause for bringing about the effect in question. As such, Thompson (1994) showed that the number of disablers and alternatives effect generalized to non-causal conditional relations (e.g., permissions, obligations, and definitions).

The results of Cummins' experiments imply that during a conditional reasoning task, people search their memory for stored knowledge of alternatives and disablers. Since the outcome of this retrieval process determines which conclusions people are willing to draw, it is of crucial importance for the reasoning community to clarify how the search process is affecting reasoning (Johnson-Laird & Byrne, 1994; Oaksford & Chater, 1998)

In a number of studies, Markovits and collaborators have started to specify this search mechanism, which constitutes the core of their general model of conditional reasoning (see Janveau-Brennan & Markovits, 1999; Markovits, 2000; Markovits, Fleury, Quinn, & Venet, 1998; Markovits & Potvin, 2001; Quinn & Markovits, 1998).

The model states that while making conditional inferences, reasoners will automatically access structures with relevant information in semantic memory[1]. Such a structure contains semantically or propositionally related elements. In causal conditional reasoning, the structures would consist of possible alternative causes and disabling conditions. Alternatives and disablers would be stored in different structures. According to many influential models of long-term memory (e.g., Anderson, 1983; Gillund & Shiffrin, 1984), the probability of retrieving at least one element from such a semantic memory structure will depend on the number of elements within the structure. Thus, the probability of retrieving at least one element from the structure storing alternative causes will be higher for conditionals with many possible alternative causes. In the same way, the probability of retrieving a disabling condition will be higher for conditionals with many possible disablers.

The model thus accounts for the effect of number of alternatives and disablers on the underlying reasoning process: The number of possible alternatives and disablers affects successful retrieval of such an element from the corresponding semantic memory structures. When an alternative cause is retrieved, the original antecedent will no longer be perceived as necessary for bringing about the consequent. As a consequence, the DA and AC inferences will be less accepted. Retrieval of a disabling condition will decrease the perceived sufficiency of the original antecedent for bringing about the consequent. This will result in increased rejection of the MP and MT inferences.

We group these ideas under the heading of 'semantic memory framework'. Although the framework starts specifying a crucial component of the reasoning process, a number of properties are not addressed and remain untested. The present study focuses on this issue. We present three experiments where the characteristics of the semantic search process during conditional reasoning are further tested and explored.

[1] As the introductory examples make clear, alternatives and disablers are typical instances of a person's general knowledge about the world, usually stored in semantic memory (Tulving, 1983). Of course, this does not exclude that some alternatives or disablers, tied to a specific personal experience, might be retrieved from episodic memory.

First, we address a neglected issue concerning the relation between different retrieval factors. Remember that one of the framework's central claims is that the probability of successful retrieval is higher for conditionals with many alternatives or disablers. Although it is well established that the probability of retrieval is affected by the number of stored elements, other factors are known to affect the retrieval too. The relation between these factors has to be taken into account in order to validate the claim. A prior concern here is the recently identified impact of 'strength of association' on conditional reasoning (see De Neys, Schaeken, & d'Ydewalle, in press-a; Quinn & Markovits, 1998).

Quinn and Markovits identified the associative strength factor within the semantic memory framework: In the memory structure with possible alternative causes of a conditional, some alternatives will be more strongly associated with the consequent in question than will others. For example, the cause 'the dog has fleas' will be more strongly associated with the consequent 'a dog scratches constantly' than the cause 'the dog has skin disease'. Quinn and Markovits showed that in addition to the number of alternative elements in a structure, the relative strength of association facilitated retrieval. In a related study, De Neys et al. (in press-a) found that the strength of association effect was also present for the disabling conditions.

Participants in Experiment 1 were asked to generate disablers and alternatives for a set of conditionals. As in Cummins' (1995) pretest, we recorded the number of generated alternatives and disablers in order to classify the conditionals in groups with many and few disablers and alternatives. We also recorded how frequently each individual alternative and disabler was generated to measure strength of association (see Quinn and Markovits, 1998). This allowed us to establish the relation between both retrieval factors.

In addition, plausibility ratings for the individual disablers and alternatives were collected. Since more readily available stored elements will tend to be judged more likely (Kahneman, Slovic, & Tversky, 1982) this allowed to take an additional retrieval factor into account.

The impact of the number of stored disablers and alternatives on inference acceptance is well established. In the second experiment we looked for the first time at the impact on the inference latencies. The additional latency data can contribute to a further characterization of the semantic memory framework: Whether or not one has stored many disablers or alternatives will have clear processing consequences. Of primary interest here is the fact that some studies (e.g., Conway & Engle, 1994) have shown that the time course of a memory search process is affected by the number of elements that are retrieved from a memory structure. It is likely that for conditionals with many disablers or alternatives more of these elements will be retrieved. Therefore, within the semantic memory framework, one might expect that the time needed for an inference will also be affected by the available number of disablers and alternatives. This was explored in Experiment 2 by recording both acceptance ratings and reaction times for the different inferences in a replication of Cummins (1995). The replication also allowed us to examine the generality of a previously reported (e.g. Liu, Lo, & Wu, 1996; Markovits & Potvin, 2001) additional effect of the number of disablers on AC and DA acceptance.

Finally, in a third experiment we try to obtain further evidence for the role of the memory search process in conditional reasoning. If the outcome of the search process determines the inferences people draw, we should expect that individual differences in the efficiency of the retrieval process will affect reasoning performance. Janveau-Brennan and Markovits (1999) already showed that elementary school children's capacity to generate possible alternative conditions was related to their performance on the AC and DA inferences. Here, we tested whether individual differences in adult reasoners capacity to retrieve disabling conditions affected performance on MP and MT inferences. The semantic memory framework states that the increased rejection of MP and MT results from the successful retrieval of a disabling condition. Since the probability of successful retrieval will be higher for people with a more efficient search process, we predict that the better one is at retrieving disablers from semantic memory, the less MP and MT should be accepted.

EXPERIMENT 1 – RELATION BETWEEN RETRIEVAL FACTORS

Whether or not the semantic search process retrieves an alternative or disabler depends on more than the number of stored elements. The probability of successful retrieval will also increase with the associative strength or plausibility of the stored elements. In Experiment 1 we examined the relation between different factors affecting the retrieval of stored alternatives and disablers. Participant generated disablers and alternatives for a set of conditionals. On the basis of the number of generated items, conditionals were classified in groups with many and few alternatives or disablers. Associative strength was measured by recording the frequency of generation of the individual alternatives and disablers. After, the generation task participants rated the plausibility of the generated alternatives or disablers.

Within the semantic memory framework one would expect to see positive relations between the factors of number, associative strength, and plausibility. That is, if conditionals with more disablers have also more strongly associated and plausible disablers, the claim that the probability of successful retrieval is higher for conditionals with many alternatives or disablers would be validated.

Method

Participants

Forty first-year students in psychology were enrolled. Half the participants were required to generate possible alternative causes and the other half to generate possible disabling conditions. All participants were native Dutch speakers.

Material

The 16 causal conditionals from Cummins (1995, Experiment 1A) were used for the generation task. The conditionals were selected by Cummins because they constituted a 2 (few vs. many alternatives) x 2 (few vs. many disablers) manipulation of the factors of number of alternative causes and number of disabling conditions. Another four causal conditionals that seemed to vary in terms of possible alternatives and disablers were adopted from the literature. Item format and instructions were similar to the generation task of Cummins (1995) and Cummins et al. (1991). Thus, the following format was used:

> 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.

This is an example of the alternative causes generation task. The format of the disabling conditions generation task was similar except that under the heading fact would appear 'The air conditioner was turned on, but you don't feel cool'. Formats like these were constructed for each of the 20 conditionals; they were typed one to a page in a booklet. The order in which the conditionals appeared in the different booklets was randomized. Task instruction stressed the importance of producing items that were reasonably realistic and different from each other. Participants were instructed that simple variations of the same, simple 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.

Procedure

Participants were run in groups of two to six. The top sheet of the generation task booklet included the written instructions, which were read aloud to the participants. Participants had 1.5 min to write down their answers for each conditional.

After the generation task, participants received written instructions for the plausibility rating task. They were asked to rate the plausibility of the disablers or alternatives they had generated on an 11-point scale, (0 = *very implausible*; 10 = *very plausible*). Ratings had to be written down in the booklets alongside each generated disabler or alternative. Participants received a practice conditional ('If Bart drinks coffee in the evening, then he doesn't sleep well') with possible disablers or alternatives. Instructions for the disabler rating task made clear we wanted participants to rate how plausible they judged the generated disablers as 'explanations' for the non-occurrence of the specified effects (e.g., ' Bart drank coffee in the evening, but he slept well. How plausible is it that this was due to the fact that the coffee was decaffeinated?'). Participants who had generated alternatives were instructed to rate the plausibility of the generated factors as alternative causes for the specified effects (e.g., 'Bart did not sleep well. How plausible is it that this was due to the fact that it was too noisy?).

As in Quinn and Markovits (1998) and De Neys et al. (in press-a), associative strength was measured by recording how frequently an individual alternative or disabler was generated across participants. For example, if 12 out of the 20 participants would generate "taking off sweater" as alternative for "If the air conditioner is turned on, then you feel cool", "taking off sweater" would receive an associative strength of 60%.

The generation protocols were scored by two independent raters in order to identify unrealistic items and items that were variations of a single idea.

Results and discussion

Overall, 5.6% of the generated items were disallowed by the raters. Interrater reliability for the alternatives generation task was .95 and for the disablers generation task .83. These figures are in line with the data of the Cummins (1995) generation task.

In general, the mean number of generated alternatives and disablers for the 16 'Cummins' conditionals was also very similar to what was originally reported. Nevertheless, two original conditionals ('If Alvin reads without his glasses, then he gets a headache' and 'If the doorbell is pushed, then it will ring') needed to be replaced because the number of generated disablers differed substantially from Cummins (1995). Table A1 in the appendix presents the 16 selected conditionals. The table contains the respective means together with the most frequently generated alternatives and disablers for every conditional.

Table 1

Mean Number and Mean Plausibility of Generated Alternatives/Disablers for Conditionals Classified as Having Many or Few Alternatives/Disablers

		Alternatives		Disablers	
		Many	Few	Many	Few
Mean number		4.03	1.91	4.03	2.31
Mean plausibility		6.34	5.08	6.29	5.21
Mean number AS	50%	3	.88	2.88	1.75
	75%	1.13	0.25	1.5	0.75
	90%	0.63	0.13	0.5	0.5

Note. There are eight conditionals in each group. The last three rows present the mean number of alternatives and disablers generated by at least 50, 75, and 90 percent of the participants (AS= associative strength). The plausibility rating scale ranged from 0 (very implausible) to 10 (very plausible).

We were interested in the differences in generation frequency and plausibility ratings of alternatives and disablers for conditionals classified as having many or few of these conditions. Results are summarized in Table 1. Overall, when a conditional had many alternatives or disablers, the disablers (34% vs. 32%) and alternatives also (30% vs. 22%) tended to have a higher strength of association. Spearman Rank Order Correlations confirmed this positive relation between a conditional's number of possible disablers and the disablers' mean strength of association [$r_s = .43$, $n = 16$, $t(14) = 1.77$, $p<.1$]. The alternatives showed a similar relation [$r_s = .65$, $n = 16$, $t(14) = 3.23$, $p<.01$]. To provide a more specific picture, we also looked at the number of strongly associated alternatives/disablers in both groups. Three frequency levels (disabler or alternative generated by at least 50%, 75% and 90% of participants) were used as criterion.

As Table 1 indicates, comparing the number of strongly associated elements supported the global analysis: The group of conditionals with a high number of alternatives has also a higher mean number of strongly associated alternatives (both at the 50%, 75% and 90% level). The disablers show the same trend at the 50% and 75% levels. The high correlations between the number and number of strongly associated (>50%) alternatives [r_s = .82, n = 16, t(14) = 5.82, p<.001] and disablers [r_s = .73, n = 16, t(14) = 4.01, p<.002] supported these findings[2].

Thus, conditionals with a higher number of possible disablers or alternatives will also have more strongly associated disablers and alternatives. Since both the number and strength of association increase the probability of retrieval, these results validate the claim that successful retrieval is more probable for conditionals with many alternatives or disablers.

Table 1 further shows that the generated disablers and alternatives for conditionals classified as having many disablers/alternatives are also rated more plausible. For the 16 selected conditionals, Spearman Rank Order Correlations showed a high positive correlation between the number of generated disablers and the mean plausibility of these disablers [r_s = .70, n = 16, t(14) = 4.01, p<.002]. The same trend was observed for the alternatives [r_s = .62, n = 16, t(14) = 2.97, p<.02]. Given the often observed 'availability heuristic' (Kahneman et al., 1982), one can assume that the higher plausibility ratings reflect easier retrieval. In this sense, the findings present converging evidence for the strength of association conclusion.

The plausibility rating gives us also an indication of the 'quality' of the retrieved disablers and alternatives. Chan and Chua (1994) showed that this 'quality' of a presented disabler (i.e., the salience or perceived importance of the disabler in relation to the occurrence of the consequent) affected MP and MT acceptance. This could imply that in addition to the higher retrieval probability, a disabler (or alternative) that is retrieved from a memory structure with many stored elements will have a stronger impact on the inferences acceptance. It is worthwhile to note that this might be an additional factor that contributes to the effect of the number of alternatives and disablers on conditional reasoning.

[2] One should note that Quinn and Markovits (1998) measured strength of association in a generation task that only allowed 30s generation time, while the present study allowed 1.5 min (as in Cummins, 1995). It could be argued that this longer generation time confounded the strength of association measure. However, for the retrieval efficiency measure in Experiment 3 we used a disabler generation task with 30s generation time. This allowed to compare the frequency data for six conditionals with 30s and 1.5 min generation time. The crucial relative ranking of the disablers was hardly affected: Spearman Rank Order Correlation reached .84 [t(82) = 13.71, p<.0001].

EXPERIMENT 2 – INFERENCE STUDY

First, Experiment 2 explored the effect of the number of stored alternatives and disablers on the conditional inference latencies. Previous studies have focused solely on the impact on inference acceptance. However, latency data may allow us to further characterize the crucial semantic search process during conditional reasoning.

Experiment 1 showed that for conditionals with many disablers or alternatives the memory structure storing these elements will typically contain a number of strongly associated elements. Available evidence from memory studies (Conway & Engle, 1994) suggests that (up to four elements) a semantic search process will take longer when the number of elements that are retrieved from a memory structure increases. When making conditional inferences, such a longer search process should result in longer inference times. Thus, availability of many alternatives and disablers might result not only in lower AC/DA and MP/MT inference acceptance but, due to an extended search process, also in longer inference latencies. This expectation was explored by recording both acceptance ratings of the conclusions and the time needed to evaluate them in a replication of Cummins (1995, Experiment 1A).

Second, Experiment 2 allowed us to examine a possible impact of the number of possible disablers on the AC and DA inferences. Although not detected by Cummins (1995), Liu et al. (1996) and Markovits and Potvin (2001, Experiment 3) observed that AC and DA were more accepted when many disablers were available. At present, the semantic memory framework does not incorporate such an effect. Since procedural variations in the cited studies might restrict comparability with Cummins' work, the present large scale (100 participants) replication will allow to assess the generality of the trend and possible implications.

It is important to note that in the present study we are interested in the general inference latency pattern. Therefore, contrary to more traditional reasoning studies, latencies are analyzed in function of the number (few vs. many) of stored alternatives and disablers independent of the specific acceptance ratings. Furthermore, as in Cummins (1995), a graded rating scale was used 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 indicates lower acceptance ratings).

In sum, except for the possible impact of disabling conditions on AC and DA we expect to replicate Cummins' (1995) findings on the acceptance ratings: MP and MT should be more rejected for conditionals with many disablers than for conditionals with few disablers. AC and DA should be more rejected for conditionals with many alternatives than for conditionals with few alternatives. For the latency effects, given the extended search hypothesis one expects longer MP and MT latencies for conditionals with many disablers than for conditionals with few disablers. Likewise, we expect longer AC and DA latencies for conditionals with many alternatives than for conditionals with only few alternatives.

Method

Participants

One hundred and one undergraduate students from the University of Leuven participated as paid volunteers or for partial fulfillment of a course requirement. All were native Dutch speakers and none had had training in formal logic.

Materials

The 16 conditionals selected in the generation task (see Table A1) were used. The conditionals yielded a 2 (few/many) x 2 (alternatives/disablers) design with four items per cell. The 16 conditionals were embedded in the four (MP, DA, MT, and DA) inference types, producing a total of 64 inferences for each participant to evaluate.

The experiment was run on computer. The item format was based on Cummins (1995). 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:

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

Type down the number that best reflects your decision about the conclusion:_

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

Procedure

Participants were run in groups of two to eight. 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. Participants used the keypad to type down the number reflecting their decision and pressed the Enter-key when finished. The next item was presented 750 ms after the Enter-key was pressed. The instructions made clear that there were no time limits, but it was stressed that once participants had made their final decision, they had to press the Enter-key immediately. The time between the presentation of the item and pressing the Enter-key was recorded together with the acceptance rating. After half of the inferences were evaluated, item presentation was paused until participants decided to continue. Instructions stressed that no other breaks were allowed and participants were expected to work through the items one immediately after another. The 64 items were presented in random order. The experimental session was preceded by one practice trial.

It should be pointed out that, as in Cummins' (1995) study, participants were not explicitly instructed to accept the premises as true and to endorse only conclusions that follow necessarily. Instead, participants could evaluate the conclusions by the criteria they personally judged relevant. With Cummins we assume that this encourages people to reason as they would in everyday circumstances. However, we should note that strictly speaking the task is therefore not a deductive inference task (see Evans, 2000). Endorsing the logically valid (MP and MT) and invalid inferences (AC and DA) should therefore not be considered correct or incorrect reasoning. When we refer to the standard nomenclature, we adopt a nominalist stance towards the use of the terms 'valid inferences' and 'fallacies'.

Results and discussion

Each participant evaluated inferences based on four different conditionals within each 2 (number of alternatives) x 2 (number of disablers) x 4 (inference type) cell of the design.

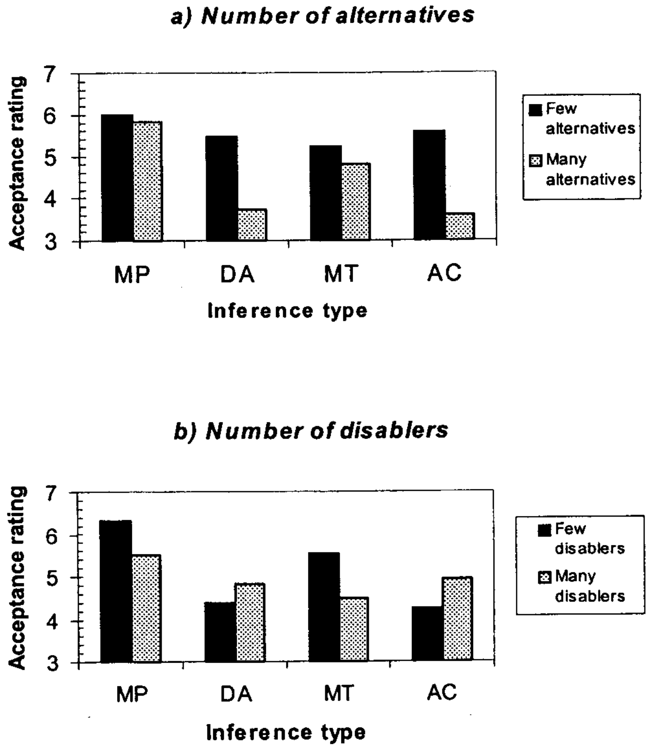
The mean of these four observations was calculated. These means were subjected to a 2 x 2 x 4 within-subjects ANOVA for acceptance ratings and reaction times.

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

The data from one participant were discarded. Mean reaction times differed from the mean reaction times of the sample by more than 3 *SD* on MP, DA, MT and 2.5 *SD* on AC.

Acceptance ratings

The main effects of inference type, number of disablers and number of alternatives were all significant [Rao R (3,97) = 82.19, p<.0001; F(1,99) = 17.94, MSe = .89, p<.0001; F(1,99) = 399.39, MSe = 1.17, p<.0001]. As in Cummins (1995), the interactions between inference type and number of disablers [Rao R (3,97) = 106.41, p<.0001] and inference type and number of alternatives [Rao R (3,97) = 96.34, p<.0001] were also significant. These interactions are depicted in Figure 1.


a) Number of alternatives

b) Number of disablers

Figure 1. The effect of (a) the number of alternatives and (b) disablers on mean acceptance ratings for the four inference types. The rating scale ranged from 1 (very sure cannot draw this conclusion) to 7 (very sure can draw this conclusion), with 4 representing can't tell.

Planned contrast tests showed that acceptance of DA [F(1, 99) = 353.59, MSe = .85, p<.0001] and AC [F(1,99) = 383.59, MSe = 1.04, p<.0001] was significantly lower for conditionals with many alternatives. Further in line with Cummins (1995), the number of disablers affected acceptance of MP [F(1,99) = 137.85, MSe = .48, p<.0001] and MT [F(1,99) = 121.41, MSe = .94, p<.0001], with lower ratings for conditionals with many possible disablers.

These findings replicate Cummins (1995). The present study also identified some additional effects[3]. First, the number of alternatives not only affected AC and DA but also MP [F(1,99) = 7.88, MSe = .37, p<.01] and MT acceptance [F(1,99) = 29.8, MSe = .6, p<.0001]: For conditionals with many alternatives, slightly lower MP and MT ratings were obtained. Nevertheless, there was still an interaction between inference type and number of alternatives. As is clear from Figure 1a, the MP and MT effects were smaller than the impact of alternative retrieval on AC and DA.

The MP and MT findings can be explained if one takes into account that in our set of conditionals there is a positive correlation (r_s = .37, n.s.) between a conditional's number of possible alternatives and disablers (see Thompson, 2000, Appendix E, for a similar observation). Therefore, conditionals with many alternatives will also tend to have a slightly higher number of disablers than conditionals with few alternatives. Consequently, since retrieving disablers will be somewhat more probable for the many alternative conditionals, the lower MP and MT acceptance is not surprising.

Second, the findings of Liu et al. (1996) and Markovits and Potvin (2001, Experiment 3) were replicated: In addition to the impact on MP and MT, the number of disablers also affected DA and AC acceptance. Both DA [F(1,99) = 38.16, MSe = .47, p<.0001] and AC [F(1,99) = 74,13, MSe = .58, p<.0001] were more accepted when many disablers were available. This effect cannot be explained by the correlation between the number of alternatives and disablers since this would result in a trend in the opposite direction.

A possible explanation points towards an interplay between the disablers and alternatives search processes. Markovits and Potvin (2001) already observed that disabler retrieval can occur during a search for alternative causes. This suggest that retrieval of alternatives and disablers is not occurring in complete isolation. Note that at present the specific relation between both search processes is not specified in the semantic memory

[3] We refer to additional effects because these were not significant in Cummins (1995). However, there were trends in the same directions (see Figure 3, p 653).

framework. We can assume that retrieving stored memory elements is resource demanding. A plausible hypothesis would be that retrieving (many) disablers puts a load on the cognitive system. This load would then burden the search for alternative causes. Thus, retrieval of alternatives would be less efficient for conditionals with many disablers which would result in a lower retrieval probability and thus higher AC and DA acceptance. Consistent with this hypothesis, both on AC [F(1,99) = 6.56, MSe = .34, p<.015] and DA [F(1,99) = 6.04, MSe = .5, p<.02] the effect of the number of disablers was mediated by the available number of alternatives: Disabler retrieval affecting AC and DA was less pronounced for the conditionals that had only few alternatives. Indeed, when only few alternatives are available successful retrieval is not very likely anyway. Thus, an eventual burden of the search process should not be expected to have a major impact here.

While interesting, the explanation is not unproblematic. For example, given that alternative retrieval did not increase MP and MT acceptance, one needs to explain why retrieval of disablers would burden retrieval of alternatives, while the reverse does not occur. Such a pattern suggests that retrieving disablers has somehow priority over retrieving alternatives. It is clear that this issue demands further research. Meanwhile, in the light of recent demonstrations of the interplay between disabler and alternative retrieval processes (Markovits & Potvin, 2001) the overload mechanism should not be discarded.

Reaction times

In order to eliminate biased measures, a trimming procedure was applied to the reaction times before they were subjected to analysis. For every inference type, any latency that was more than 3 standard deviations above a persons mean reaction time for that inference type was discarded. This procedure affected less than 1% of all observations.

The (M)ANOVA revealed a significant main effect of inference type [Rao R (3,97) = 43.03, p<.0001]. As for the acceptance ratings, we also observed significant interactions between inference type and number of alternatives [Rao R (3,97) = 4.82, p<.004] and inference type and number of disablers [Rao R (3,97) = 2.7, p<.05]. These interactions are depicted in Figure 2.

Planned contrasts tests indicated that the AC [F(1,99) = 5.47, MSe = 7.52, p<.025] inference took more time (641 ms) for conditionals with many alternatives (see Figure 2a). Likewise, the MP [F (1,99) = 11.52, MSe = 5.27, p<.001] inferences required more time (776 s) for the conditionals with many disablers (see Figure 2b).

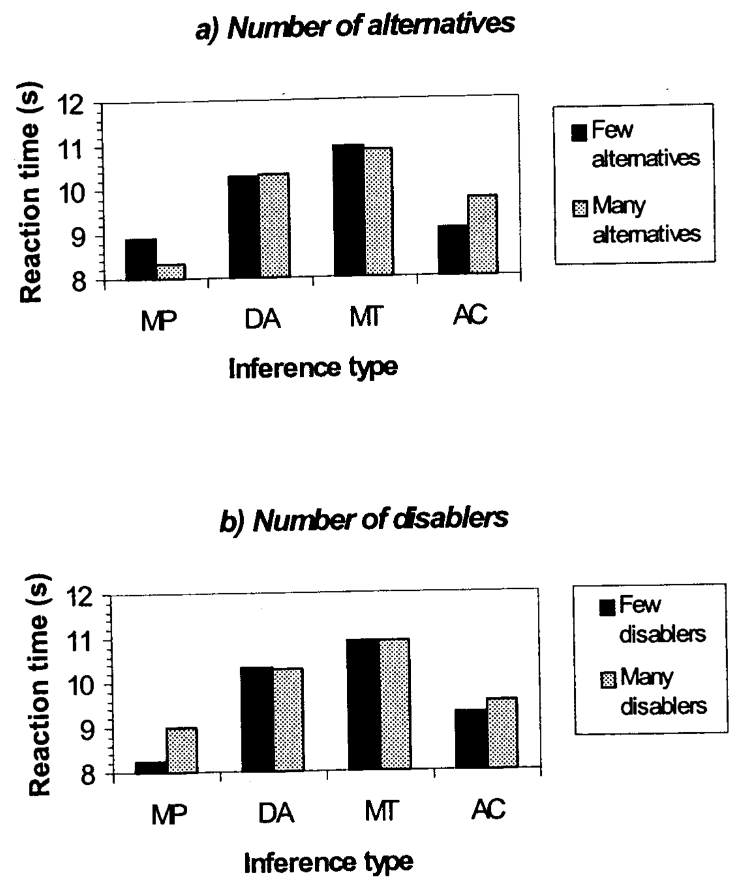


Figure 2. The effect of (a) the number of alternatives and (b) disablers on mean reaction times for the four inference types.

These findings show that the higher number of stored alternatives and disablers results not only in lower AC and MP acceptance, but also in increased inference latencies. This corroborates the hypothesized extended semantic search process for conditionals with many alternatives or disablers: When many (strongly associated) alternatives and disablers are available, retrieval of additional alternatives and disablers will result in longer inference latencies.

Although the number of available disablers and alternatives affected MT and DA acceptance, no effects were observed on the latencies. Thus, DA and MT latency data suggest there is no extended search process for these inferences. The findings can be reconciled if one takes into account that the semantic search process during conditional reasoning is possibly affected by inference complexity. DA and MT inferences are more complex than AC and MP inferences. DA and MT, the so called 'denial inferences', involve negations and reasoning theories typically state these demand more cognitive (working memory) resources (Johnson-Laird & Byrne, 1993; Oaksford, Chater, & Larkin, 2000; Braine & O'Brien, 1998). Rosen and Engle (1997) have shown that semantic memory retrieval is affected by working memory load. Therefore, the increased complexity may affect the semantic search process. Due to the higher load the search would be less extensive. The additional need to process negations would thus override the extended search for DA and MT inferences.

In general, this implies that for DA and MT based on conditionals with many alternatives/disablers, less of these elements will be retrieved than for MP and AC. Therefore, one might expect a less pronounced effect of the number of alternatives and disablers on the DA and MT acceptance ratings. However, one has to consider that when the reasoning system is already burdened, retrieval of even a single alternative might cause a cognitive overload. Such an overload can result in increased inference rejection (e.g. Toms, Morris, & Ward, 1993). Therefore, even with a less extensive search process, availability of many alternatives/disablers can still result in considerable decreased MT and DA acceptance. Consistent with this claim planned contrast test showed that the number of disablers had a similar impact on MP and MT acceptance ratings. Likewise, AC and DA acceptance were not differentially affected by the number of available alternatives.

As Figure 2a indicates, MP inferences also lasted longer (595 ms) when only few alternatives were available [F(1,99) = 11.9, MSe = 2.97, p<.001]. This effect of alternatives on MP reaction times is puzzling. The direction of the effect, with many alternative conditionals having shorter reaction times, was the opposite of the other latency effects. At present, we have no clear cut explanation for this finding.

EXPERIMENT 3 – INDIVIDUAL DIFFERENCES IN DISABLER RETRIEVAL CAPACITY

In Experiment 3 we tried to obtain further evidence for the role of the semantic search process during conditional reasoning. We tested whether individual differences in adult reasoners capacity to retrieve disabling conditions affected performance on MP and MT inferences. The semantic memory framework states that the increased rejection of MP and MT results from the successful retrieval of a disabling condition. Since the probability of successful retrieval will be higher for people with a more efficient search process, one should predict that the better one is at retrieving disablers from semantic memory, the less MP and MT should be accepted.

We do not predict an effect on the inference latencies. Reasoners with a more efficient retrieval process can be expected to be faster at retrieving disablers. Although more efficient retrievers should retrieve more disablers, the faster running search process will probably compensate the extra time needed for the additional retrieval.

Retrieval efficiency was measured by looking at the number of generated disablers in a generation task that was presented to forty participants after the inference task of Experiment 2. The generation task used four new conditionals and four conditionals already presented in the inference task. The new conditionals were included because generation for the old conditionals may be biased by familiarity. On the other hand, retrieval of disablers for adult participants may be highly specialized and conditional specific. In this latter case, generation of disablers for the new conditionals would not be informative for the retrieval efficiency for conditionals in the inference task. Comparing the results for the old and new conditionals allows to sidestep this possible complication.

Method

Participants

Forty participants that also participated in Experiment 2 took part in the present experiment.

Materials

The disabler retrieval capacity was measured by requesting participants to generate possible disablers for eight conditionals. Half of the conditionals were already presented to participants in the inference study (= old conditionals), while the other half were different (= new conditionals). Half of the old and new conditionals were classified in previous work as possessing many disabling conditions, while the other half had only few possible disabling conditions. Item format and task instructions were similar to the generation task presented in Experiment 1. The conditionals appeared in the same order in all booklets. The old conditionals were presented after the new ones.

Procedure

The retrieval measure was presented after all participants of a group had finished the inference study. Generation time was limited to 30 s for the new conditionals as in Janveau-

Brennan and Markovits (1999). To account for faster reading times (and thus more retrieval time) due to previous presentation, generation time was limited to 28 s for the old conditionals. Participants had to write down their answers. Instructions stressed the importance of writing down only the general core of the retrieved disablers, in order not to lose time because of the writing itself. The generated disablers were scored according to the list of accepted disablers as provided by the raters in our previous studies.

Results and discussion

We disallowed 5.66% of the generated disablers in the retrieval measure, mainly because they expressed variations of the same idea according to the previous classifications. The Spearman Rank-Order correlation between the generated number of disablers for the new and old conditionals was rather high ($r_s = .65$, $n = 40$, $t(38) = 5.30$, $p<.001$).

Acceptance ratings

We first analyzed the relation between participants' mean MP, DA, MT and AC acceptance rating in Experiment 1 and the total number of disablers they generated for the eight conditionals in the retrieval task. Spearman Rank-Order correlations indicated that both MP [$r_s = -.35$, $n = 40$, $t(38) = -2.33$, $p<.03$] and MT [$r_s = -.40$, $n = 40$, $t(38) = -2.67$, $p<.02$] showed the expected significant negative relation: The more disablers people could retrieve in a limited time period, the less they accepted MP and MT. Acceptance of DA ($r_s = .09$) and AC ($r_s = .13$) was not significantly related to the disabler retrieval capacity.

The same analysis was run with the number of generated disablers for the new and old conditionals as separate indexes of retrieval capacity. Both indexes pointed to the same conclusions: There were significant or marginally significant negative correlations with MP [old; $r_s = -.31$, $n = 40$, $t(38) = -1.97$, $p<.06$, new; $r_s = -.31$, $n = 40$, $t(38) =-1.98$, $p<.06$] and MT acceptance ratings [old; $r_s = -.41$, $n = 40$, $t(38) = -2.75$, $p<.01$, new; $r_s = -.31$, $n = 40$, $t(38) = -.98$, $p<.06$], but there was no effect on DA and AC. Together with the high correlation between the generated number of disablers for the old and new conditionals, this indicates that both indexes are measuring the same capacity. Therefore, in the remaining analyses we used only the total number of generated disabler as the index of retrieval efficiency.

To provide a more global picture of the observed effects, we classified participants in three groups according to the total number of generated disablers in the retrieval task.

Participants giving 17 or fewer disablers were classified as Low (n=10, Mean= 14.2), participants giving 23 or more disablers were classified as High (n=12, Mean= 24.8), while the remaining participants were classified as Intermediate (n=18, Mean=20). For each of the four inference types, we performed an ANOVA on the acceptance ratings with retrieval capacity as a between-subjects factor and number of alternatives and disablers as a within-subjects factor. This resulted in a 3 (retrieval) x 2 (many/few alternatives) x 2 (many/few disablers) design. Because the pattern of results for the number of alternatives and disablers variables repeated what had been found in the inference-study, we only report results relating to the retrieval capacity factor.

Acceptance of both MP [F(1,37) = 4.19, MSe = 2.64, p<.03] and MT [F(1,37) = 4.40, MSe = 4.94, p<.02] was affected by the retrieval factor. These effects are depicted in Figure 3. Newman-Keuls tests showed that participants with low (6.29) and intermediate (6.23) capacity to retrieve disablers from semantic memory, accepted MP significantly more than the high capacity group (5.44). A similar effect was observed for MT, with the low (5.51) and intermediate groups (5.44) giving higher ratings than the participants in the high group (4.33). The differences between the low and intermediate groups were not significant. Both for MP and MT, interactions of the retrieval and number of disabler and alternatives factors were not significant, which indicates that the retrieval factor has a similar effect on conditionals with few and many disablers.

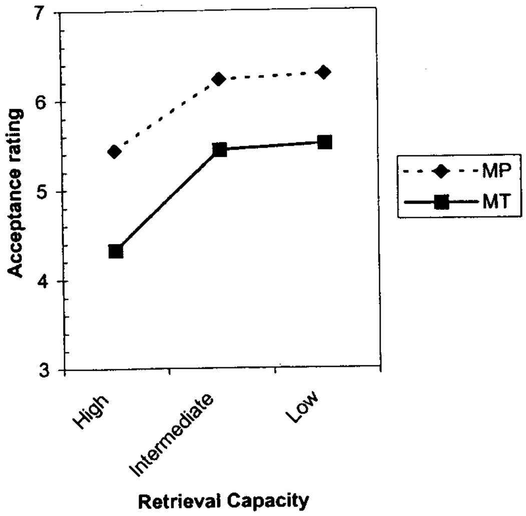


Figure 3. The effect of disabler retrieval capacity on MP and MT acceptance ratings. Retrieval capacity is measured by the number of generated disablers (Low, High, or Intermediate) in the generation task.

We mentioned that Janveau-Brennan and Markovits (1999) already found that children's capacity to retrieve alternatives was related to acceptance of DA and AC. The more alternatives children could retrieve, the less they accepted both inferences. This corroborates the present findings. However, in addition Janveau-Brennan and Markovits also observed that an efficient retrieval of alternatives resulted in lower MP acceptance. Although there was no effect on MT, they interpreted this as evidence for a general retrieval capacity. In the present study DA and AC acceptance ratings were not affected by participants' capacity to retrieve disablers from semantic memory. This suggests that for adult reasoners, retrieval capacity for alternatives and disablers is not related.

Reaction times

Spearman Rank Order correlations between participants' mean MP, DA, MT and AC latencies in the inference task and the total number of generated disablers for the eight conditionals in the retrieval task were calculated. Although the DA ($r_s = -.25$), AC ($r_s = -.20$) and MT ($r_s = -.16$) correlations showed a negative trend, none of the correlations were significant (not even at the .1 level). When the analysis was run with number of generated disablers for the new and old conditionals as separate indexes of retrieval capacity, the same, non-significant, pattern was observed.

As expected, these findings indicate that good disabler retrievers do 'more in the same time'. Although they will retrieve more disablers, which is time consuming, the retrieval will also be faster than for people with less good retrieval capacities. Hence, the inference time is not affected.

GENERAL DISCUSSION

The results of the present study supported and refined the semantic memory framework of conditional reasoning. This framework explains the effect of the number of possible disablers and alternatives of a conditional and integrates it with the underlying reasoning process. It states that the possible alternatives and disablers of a conditional are stored in semantic memory structures and that reasoners access these structures when presented a conditional inference problem. The framework further assumes that the number and strength of association of the elements in the memory structures affect successful retrieval. Retrieval of a disabler will decrease the perceived sufficiency of the original antecedent for bringing about the consequent. This results in an increased rejection of the MP and MT inferences. When an alternative cause is retrieved, the original antecedent will no longer be perceived as necessary for bringing about the consequent. As a consequence, the DA and AC inferences are less accepted.

We started the study with an examination of the relation between different factors affecting retrieval of alternatives and disablers. We observed high correlations between the different factors: Conditionals with many elements in the semantic memory structures had a higher number of strongly associated elements and the elements were also rated as more plausible. This supported the framework's central claim that successful retrieval is more likely when many alternatives or disablers are available.

The second experiment explored the impact of the number of available alternatives and disablers on conditional inference latencies. It was hypothesized that due to a more extended search process, inferences would last longer when many alternatives or disablers could be retrieved. Consistent with this claim, AC took more time for conditionals with many alternatives while MP latencies increased when many disabling conditions were available. For MT and DA inferences there was no evidence for an extended search process. We hypothesized that the additional need to process negations for MT and DA overrides the retrieval of additional alternatives or disablers. In addition to Cummins (1995) findings, Experiment 2 also identified an impact of the number of available disablers on AC and DA acceptance: When many disablers were available AC and DA were less accepted. We argued that this finding suggests an interplay between the alternative and disabler search processes.

In Experiment 3 the role of the semantic search process during conditional reasoning was further established. As expected, individual differences in the efficiency of the disabler retrieval process affected inference acceptance. The better people were at retrieving disablers, the less MP and MT were accepted.

This study supports the semantic memory framework, but it is important to remark that possible alternative accounts might be suggested. Especially with respect to the latency data, such a possible alternative explanation is provided by the mental models theory (Johnson-Laird, 1983; Johnson-Laird, Byrne, & Schaeken, 1992). One should note that unlike Cummins (1995), Markovits and collaborators (e.g., see Markovits, 2000; Markovits et al., 1998) have incorporated the semantic memory framework in this general reasoning theory.

Mental models theory could explain the increased MP and AC latencies as the result of an additional mental representation process. The theory will state that retrieval of an alternative or disabler alters the mental representation of the conditional that is used as the basis for inferences: After successful retrieval the initial representation of the conditional will be extended with an additional model (Byrne, Espino, & Santamaria, 1999). The additional model would represent the possibility that occurrence of the antecedent is not linked with occurrence of the consequent (for a disabler) or the fact that the consequent can be linked with different alternatives (for an alternative). In Byrne et al. (1999) these extra models constitute what the authors refer to as the 'recursive' and 'conditional' interpretation of a conditional. Now, the additional model construction is known to be time consuming (see Barrouillet, Grosset, & Lecas, 2000). Since the additional representation will only be constructed when retrieval of an alternative or disabler is successful, this can partially explain the longer inference latencies. Thus, the possible contribution of such a representational process to the MP and AC latency findings should at this point not be discounted.

The present study also points towards some important issues that are currently not addressed in the semantic memory framework. A prior concern is that the framework needs to establish the precise relation between the different processes taking part in conditional reasoning. At present the framework assumes that, for example, the processes of representing the conditional, processing negations, and searching for stored alternatives and disablers are occurring in complete isolation. We believe the findings of Experiment 2 suggest this is not the case. Rather, the different processes would 'compete for limited resources'. The crucial role of working memory in both conditional reasoning (e.g., Barrouillet & Lecas, 1999; Toms et al., 1993; Meiser, Klauer, & Naumer, 2001) and semantic memory retrieval (e.g., Rosen & Engle, 1997) has been established. Thus, it is likely that the processes that are crucial in the semantic memory framework are all burdening the limited working memory resources. Therefore, it would not be surprising that processing of negations affects the extent of the search process or that disabler retrieval affects the search for alternatives. Furthermore, the fact that some processes can have priority over others (e.g., disabler retrieval over alternative retrieval), might help to explain some of the puzzling findings in the present study.

However, we acknowledge that our findings should be interpreted with some caution. Experiment 2 is just the first experiment where the effect of alternative and disabler retrieval on inference latencies was examined. Establishing the specific relation between the retrieval processes and eliminating possible alternative suggestions will demand more detailed research. Nevertheless, the obtained results legitimate the further development of the framework. The present study thus established the semantic memory framework as a viable starting point for unraveling the complicated relationship between inference processes and memory retrieval processes.

Appendix

Table A1

Mean Number and Mean Plausibility (P) of Generated Alternatives (Alt) and Disablers (Dis) for the 16 Conditionals Adopted for the Inference Study

Conditional	Mean Alt	Mean P	50% alternatives	Mean Dis	Mean P	50% disablers
Many alternatives, many disablers						
1. If fertilizer is put on plants, then they grow quickly	3.7	6.84	Well watered (75%) Lots of sunlight (70%) Fertile soil (55%) Naturally fast growers (50%)	4.5	6.52	No water (75%) Plants dying (65%) No sunlight (55%) Too much/little applied (55%) Wrong type (55%)
2. If the brake is depressed, then the car slows down	4.1	6.02	Uphill (70%) Foot off accelerator (60%) Out of gas (55%) Collision (55%)	3.5	6.21	Brake broken (85%)
3. If john studies hard, then he does well on the test	4.4	5.47	Cribbing (90%) Easy test (70%) Lucky (60%)	4.9	6.26	Test too hard (75%) Not concentrated (60%) Low IQ (50%)
4. If Jenny turns on the air conditioner, then she feels cool	4.6	6.16	Took off clothes (75%) Open window (60%)	4.2	5.70	Airco broken (95%) Fever (50%)
Many alternatives, few disablers						
5. If Bart's food goes down the wrong way, then he has to cough	3.6	6.92	Catch a cold (100%) Attract attention (80%)	2.1	5.07	Choked not hard enough (80%)
6. If Marry jumps into the swimming pool, then she gets wet	4.4	6.46	Rains (100%) Shower (60%)	2.5	4.93	Pool empty (100%) Wearing dry-suit (95%)
7. If the apples are ripe, then they fall from the tree	3.4	6.16	Storm (95%) Tree shaken by agent (55%) Dropped by picker (55%)	2.1	5.23	Picked (65%)
8. If water is poured on the campfire, then the fire goes out	4	6.57	Died out (95%) Smothered with blanket/sand (80%) Rain (60%) Wind blew out (55%)	2.5	5.96	Too little water (90%) Very large fire (65%)

Table A1(continued)

Conditional	Mean Alt	Mean P	50% alternatives	Mean Dis	Mean P	50% disablers
Few alternatives, many disablers						
9. If the trigger is pulled, then the gun fires	1.7	4.24	Faulty design (55%)	3	6.40	No bullets (100%) Gun broken (75%)
10. If the correct switch is flipped, then the porch light goes on	1.9	5.68	Faulty wiring (50%)	3.9	7.21	No power (100%) Missing or broken bulb (95%) Switch broken (75%)
11. If the ignition key is turned, then the car starts	1.8	5.28	Hot wired (75%)	4.2	6.19	Engine/car broken (75%) Wrong key (50%) No fuel (50%)
12. If the match is struck, then it lights	1.8	5.85	Lit with other fire (100%)	4	5.83	Match wet (80%) Not struck hard enough (75%) Worn matchbox pad (60%) Used match (50%)
Few alternatives, few disablers						
13. If Joe cuts his finger, then it bleeds	3.1	5.52	Scraped/scratched (60%) Removed scab (50%)	2.7	5.76	Cut not deep enough (100%) Knife blunt (65%) Cut in nail/callous (65%)
14. If Larry grasps the glass with his bare hands, then his fingerprints are on it	1.6	3.97	[Still on from earlier grasp (35%)]*	1.9	4.67	Hands not greasy (50%)
15. If the gong is struck, then it sounds	2.3	4.51	Gong fell/bumped (55%)	2.7	4.53	Struck too lightly (70%) Padded/gripped (70%) Struck with light material (55%)
16. If water is heated to 100°C, then it boils	1.1	5.5	[Still warm from earlier heating (20%)]*	2	5.49	No pure water (75%)

Note. The relative frequency of generation for conditions that were mentioned by at least 50% of participants is presented in order of frequency. The plausibility rating scale ranged from 0 (very implausible) to 10 (very plausible).
* Most frequently generated alternative.

Table A2

Characteristics of the Additional Conditionals Not Selected in Experiment 1

Conditional	Mean Alt	Mean P	50% alternatives	Mean Dis	Mean P	50% disablers
1. If Alvin reads without his glasses, then he gets a headache	4.5	6.29	Bonked head (75%) Fever (65%) Hangover (55%)	3.5	6.36	Took aspirin (85%) Wore contacts (70%) Didn't read long (70%) Large print book (50%)
2. If the doorbell is pushed, then it will ring	1.8	4.37	Malfunction (65%)	3.15	6.14	Bell broken (90%) Not pushed hard enough (75%) No power (60%)
3. If Jan consumes alcohol, then he gets drunk	2	4.67	[Spinning around (40%)]*	3.15	7	Consumes only a little (85%) Ate a lot (75%) Low percent alcohol (50%)
4. If Steve goes in for sports, then he loses weight	4	6.46	Ate less (75%) Sick (75%) Stress (70%) Ate healthy food (65%)	3.2	6.58	Low intensity/frequency (100%) Bad diet (90%) Playing chess/pool (65%)

Note. * Most frequently generated alternative.

Notes on the manuscript

NOTE

In Experiment 3 we found no correlation between disabler retrieval capacity and AC/DA acceptance. We therefore suggested, contrary to Janveau-Brennan and Markovits (1999), that retrieval capacity for alternatives and disablers is not related. However, this suggestion was discarded in a subsequent study. We gave forty participants both a disabler and alternative generation task (presentation order was counterbalanced). Results clearly showed that the number of generated disablers was associated with the number of generated alternatives, $r_s = .55$, $t(38) = 4.06$, $p < .001$. The test thus directly established that the retrieval capacity for alternatives and disablers is related: The better one is at retrieving disablers, the better one will be at retrieving alternatives. This supports Janveau-Brennan and Markovits' claim for a general retrieval capacity. Presumably we did not detect an indirect correlation between AC/DA acceptance and disabler retrieval capacity because the disabler retrieval assessment was not sensitive enough.

CHAPTER 4

Every counterexample counts?!

In this study we further examined the characteristics of the counterexample search process during everyday conditional reasoning. Experiment 1 manipulated the number (zero to four) of explicitly presented counterexamples (alternative causes or disabling conditions) for causal conditionals. In Experiment 2, a generation pretest measured the number of counterexamples participants could retrieve for a set of causal conditionals. One month after the pretest, participants were presented a reasoning task with the same conditionals. The experiments indicated that acceptance of Modus Ponens linearly decreased with every additionally retrieved disabler, while Affirmation of the Consequent acceptance linearly decreased in function of the number of retrieved alternatives. Results for Denial of the Antecedent and Modus Tollens were less clear. The findings show that the search process does not necessarily stop after retrieval of a single counterexample and that every additional counterexample has an impact on the inference acceptance.

Based on: De Neys, W., Schaeken, W., & d'Ydewalle, G. (in press-b). Inference suppression and semantic memory retrieval: Every counterexample counts. *Memory & Cognition.*

INTRODUCTION

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

Cognitive scientists have spent a great deal of research to establish how people reason with these 'if, then' sentences. The research has typically focused on peoples performance on four kinds of conditional arguments: The above illustrated Modus Ponens (MP, e.g., 'If p then q, p therefore q) and Affirmation of the Consequent (AC, e.g., 'If p then q, q therefore p') inferences, Modus Tollens (MT, e.g., 'If p then q, not q, therefore not p'), and Denial of the Antecedent (DA, e.g., 'If p then q, not p, therefore not q). The first (p) part of the conditional is called the antecedent and the second (q) part is called the consequent.

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

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

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

Further studies established that the suppression effect arises even without explicit presentation of counterexamples (e.g., Cummins, 1995; Cummins, Lubart, Alksnis, & Rist, 1991; Markovits, 1986; Thompson, 1994, 1995). Cummins and colleagues examined the role of counterexample retrieval by looking at the effect of the number of possible alternatives and disablers of a conditional. In a pretest they identified conditionals for which participants generated many or few possible alternatives and disablers. These conditionals were then adopted for a reasoning task with a second group of participants.

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

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

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

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

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

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

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

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

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

The alternative specification gains some credence from related studies. In the field of 'uncertain' or probabilistic reasoning the work of Liu, Lo, and Wu (1996) is especially relevant. Participants received three different conditionals that were previously rated as having a high (e.g., If John lives in Canada, then he lives in the Northern Hemisphere), medium (e.g., If Mary moves, then she adds some furniture) or low (e.g., 'If Stan wears glasses, then he is intelligent') sufficient antecedent. Liu et al. observed that MP and MT acceptance gradually decreased with decreasing sufficiency. With this realistic, thematic material, conditionals with lower sufficiency levels will presumably have a higher number of possible disablers. While inconclusive, one might thus suggest that the lower acceptance is the result of additional disabler retrieval: The more disablers are retrieved, the less MP and MT are accepted.

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

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

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

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

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

EXPERIMENT 1

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

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

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

Method

Participants

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

Material

The material was selected from previous pilot work (see De Neys et al., 2002) where 40 participants wrote down as many alternatives or disablers for a set of 20 conditionals (with 1.5 min generation time for each conditional). Two independent raters scored the generation protocols in order to eliminate unrealistic items and items that were variations of a single idea.

The conditionals, item format, instruction and scoring procedure for the pilot were based on Cummins (1995). For every conditional the mean number of generated counterexamples and the relative frequency of generation of every counterexample was recorded. For the present experiment, we selected five conditionals with many (above the group mean) possible disablers, and five conditionals with many (above the group mean) possible alternatives.

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

Each participant received a 6-page booklet. Page one included the task instructions. On top of each of the next five pages appeared the selected conditionals in bold. One of them was presented without possible counterexample while for the others the versions with one, two, three and four counterexamples were presented. Thus, each participant received five different conditionals with the number of presented counterexamples ranging from zero to four. In every booklet, we varied which conditional was used in which counterexample version. We made sure that each of the five counterexample versions of the different conditionals was used equally often (i.e., in approximately 1/5 of the booklets). The conditional without counterexample was always presented first, while the remaining conditionals appeared in random order.

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

Rule: **If fertilizer is put on the plants, then they grow quickly**

but:
if the plants get enough water, they will also grow quickly
if the plants get a lot of sunlight, they will also grow quickly
if the plants are planted in fertile soil, they will also grow quickly

---

Fact: **The plants grow quickly**
Conclusion: **Fertilizer was put on the plants**

					I					
---5---	---4---	---3---	---2---	---1---	---0---	---1---	---2---	---3---	---4---	---5---
Very Sure				Some what	I I I I	Some what				Very Sure

That I CANNOT draw this conclusion

That I CAN draw this conclusion

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

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

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

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

Procedure

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

Participants were instructed to place a mark on one of the numbers of the scale that best reflected their decision. Care was taken to make sure the participants understood the precise nature of the rating scale: Placing a mark on the left side of the scale indicated they believed the conclusion could not be drawn, placing a mark on the right side of the scale indicated they believed the conclusion could be drawn. Marking the zero indicated they could not tell one way or the other.

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

Results

The data from four participants were discarded because they did not solve all the inferences. Of the remaining 174 participants, 88 had received booklets from the alternatives presentation group, while 86 participants had received booklets from the other group where the number of presented disablers was manipulated.

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

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

For every inference type in both counterexample presentation groups we performed separate MANOVA's on the acceptance ratings with the number of presented counterexamples as within-subjects factor. In the analyses three consecutive issues are addressed. First, we test whether there is an overall effect of the number of counterexamples factor. Then we examine the crucial issue whether presenting more then one counterexample has an additional effect on the acceptance ratings. Third, the precise trend of an eventual additional retrieval effect is analyzed.

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

Effect of number of alternatives

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

As expected, presentation of alternatives had a significant effect on AC [Rao R(4, 84) = 10.66, p<.0001] and DA acceptance [Rao R(4,84) = 9.93, p<.0001]. Newman-Keuls tests showed that for every number of presented alternatives, AC and DA acceptance was lower than when no alternative was presented. These findings were confirmed by the combined materials analysis on AC and DA [Rao R(4,6) = 60.96, p<.0001]. Both subjects and materials analyses indicated that the alternatives had no significant impact on MP. Although, there were only five conditionals for the materials analysis on MP, it was clear that there were no meaningful trends in the data. These results replicate previous suppression findings.

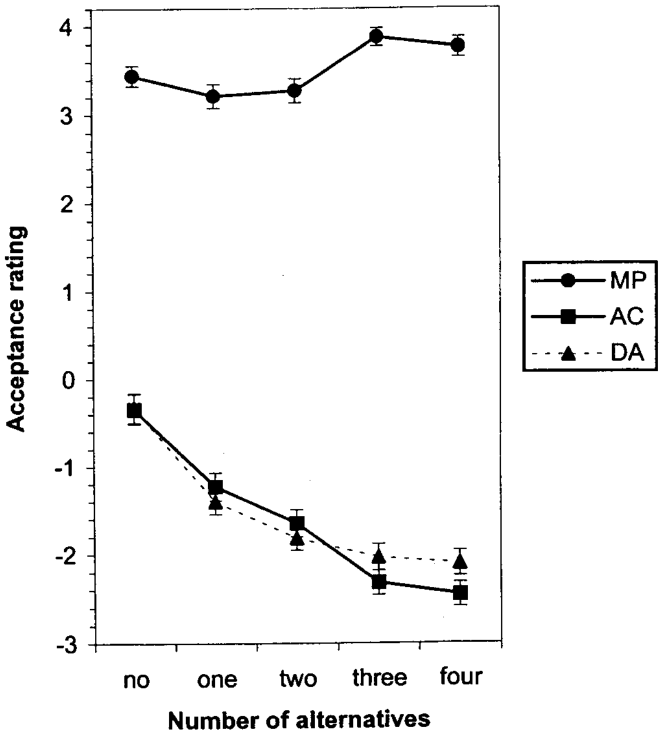


Figure 1. Inference acceptance in function of the number of presented alternatives. The rating scale ranged from –5 (very sure I cannot draw this conclusion) to +5 (very sure I can draw this conclusion). Vertical lines depict standard errors of the means.

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

The materials analysis established that there was a marginal effect of additional alternatives on the combined DA and AC acceptance [Rao R(3,7)=3.37, p<.09] and that this effect had a linear nature [F(1,9) = 8.26, MSe = .57, p<.02]. These effects are depicted in Figure 1.

Effect of number of disablers

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

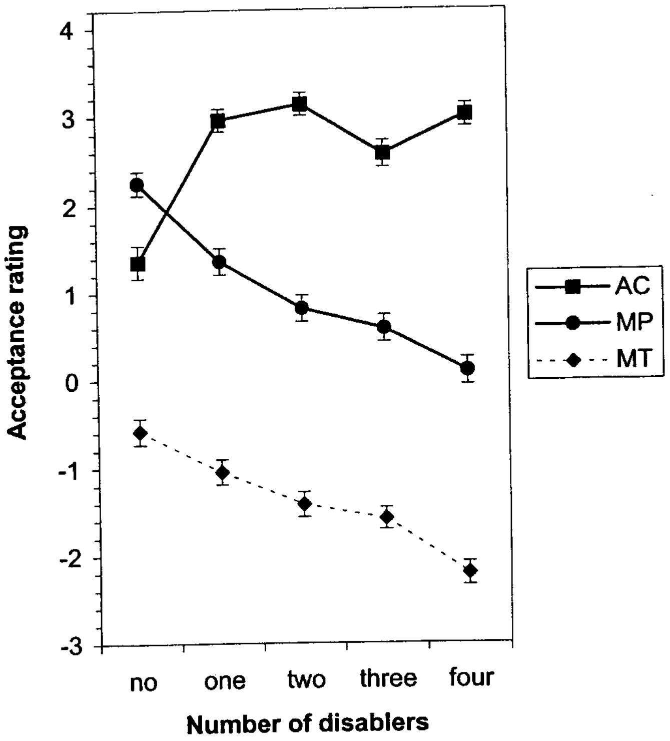


Figure 2. Inference acceptance in function of the number of presented disablers. The rating scale ranged from –5 (very sure I cannot draw this conclusion) to +5 (very sure I can draw this conclusion). Vertical lines depict standard errors of the means.

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

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

The crucial manipulation of presenting more than one counterexample had a significant effect on MP [Rao R(3,83) = 6.77, p<.0005] and MT [Rao R(3,83) = 5.56, p<.002] acceptance. Trend analysis showed that both for MP [F(1,85) = 20.46, MSe = 3.42, p<.0001] and MT [F(1,85) = 16.02, MSe = 3.58, p<.0002] there was a significant negative linear trend in the acceptance data, while higher order trends were not significant. Thus, every disabler that is retrieved in addition to the first one will result in a further decrease of MP and MT acceptance ratings. These findings were confirmed by the combined materials analysis [Rao R(4,6) = 13.44, p<.004, significant linear trend: F(1,9) = 10.18, MSe = .68, p<.015, no higher order trends].

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

Discussion

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

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

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

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

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

EXPERIMENT 2

In Experiment 1 the inference suppression linearly increased with every presented counterexample. Because of the explicit presentation procedure, this does not yet show that the actual search process retrieves additional counterexamples. However, it does imply that if people would indeed search and retrieve additional counterexamples themselves, we should see a similar linear decreasing acceptance pattern. In Experiment 2 we will look at participants' inference acceptance in function of the number of counterexamples they can retrieve for a conditional. By checking whether the same graded trends are observed, the findings of Experiment 1 can be validated.

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

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

It is crucial to stress the within-subjects nature of the analyses in the present study. The number of stored counterexamples is of course directly associated with the probability of retrieving a single counterexample. Thus, if we would compare different groups of participants (e.g., groups that retrieved one, two, three, or more counterexamples for a specific conditional) a graded effect could not be attributed to additional disabler retrieval. Indeed, it could simply be claimed that there will be a larger number of participants that retrieve a single counterexample in the successive groups. Therefore, we always compare the inference acceptance of the same participants for conditionals for which they retrieved a different number of disablers or alternatives.

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

Pretest

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

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

[2] Precise material, procedure and results of the pretest were previously reported in De Neys et al. (2002, Experiment 1). The material for Experiment 1 was also taken from the same study.

Method

Participants

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

Material

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

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

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:

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

Type down the number that best reflects your decision about the conclusion:_

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

Procedure

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

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

Results

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

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

For the main analysis we grouped all conditionals for which a participant could retrieve no or one, two, three, and four or more counterexamples. Since the vast majority of participants generated at least one counterexample for every conditional we combined the no and one group. Likewise, since rarely more than four counterexamples were generated these conditionals were combined with the four group. On average, participants generated no or one, two, three, and four or more alternatives for 3.22 (SD = 1.9), 3.11 (SD = 1.75), 3.67 (SD = 1.78), and 6 (SD = 2.81) conditionals, respectively. The average number of conditionals in the successive number of disablers groups was 1.53 (SD = 1.07), 3.53 (SD = 1.35), 5.37 (SD = 1.71), and 5.58 (SD = 1.8), respectively. For every participant we calculated the mean inference acceptance for the different conditionals in every number of counterexamples group. For every inference type, these means were subjected to a MANOVA with the number of retrieved alternatives or disablers as a within-subject factor.

[3] Since the number of possible alternatives and disablers of the conditionals in the reasoning task varied systematically, the data could be analyzed as a 2 (few/many) x 2 (alternatives/disablers) x 4 (inference type) within subjects design (see Cummins, 1995). The kind of generated counterexample (disablers or alternatives) was entered as a between subject-factor in this design. An ANOVA showed that the kind of generated counterexample factor, nor any of its interactions with the other factors reached significance.

Missing observations (e.g., a participant had no conditionals for which two alternatives were retrieved) were set to the overall mean. Both in the alternative and disabler retrieval groups this affected less than 4% of the observations.

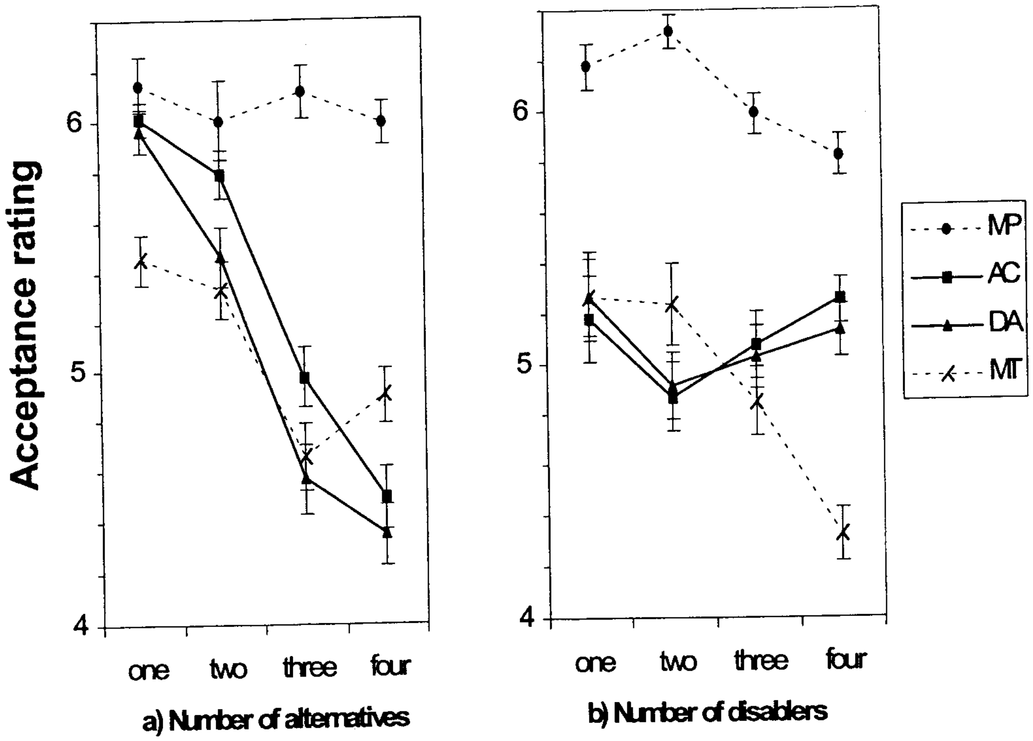


Figure 3. Inference acceptance in function of the number of alternatives (3a) or disablers (3b) participants could retrieve for a conditional. The rating scale ranged from 1 (very sure I cannot draw this conclusion) to 7 (very sure I can draw this conclusion). Vertical lines depict standard errors of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

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

As Figure 3a makes clear, the AC and DA ratings did not show a stepwise trend in function of the number of retrieved alternatives. Trend analyses established that both for AC [F(1,17) = 59.62, MSe = .43, p<.001] and DA [F(1,17) = 47.36, MSe = .62, p<.001] there was a significant negative linear trend while higher order trends were not significant. Likewise, acceptance of MP [F(1,18) = 7, MSe = .27, p<.02] and MT [F(1,18) = 6.26, MSe=, p<.03] also linearly decreased with every retrieved disabler (see Figure 3b). Higher order trends were not significant. These observations are in line with the findings of Experiment 1.

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

Table 1

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

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

The individual acceptance patterns were classified in three groups. If a participant gave three or four successive decreasing ratings, her/his acceptance pattern was classified as 'graded'. If there was a clear single step in the pattern, it was classified as 'stepwise'. A rather liberal criterion was adopted: The step had to be larger in size than the differences between the hypothesized equal ratings. For example, an acceptance rating pattern as (4, 5, 2, 3) for conditionals with respectively, one, two, three, and four counterexamples would be classified as a stepwise pattern with the retrieval threshold at 3 counterexamples. Likewise, a pattern like (6, 2, 3, 2) would be classified as a stepwise pattern with the retrieval threshold at 2 counterexamples. Patterns that could not be classified in these two categories were labeled 'other' (e.g., a pattern like 5, 3, 4, 2). Table 1 shows the classification results.

Inspection of Table 1 makes it clear that the graded trends in Figure 3 cannot be attributed to the aggregation of individual stepwise trends. For every inference type the acceptance rating of the vast majority of participants showed a graded acceptance trend. For the participants that did show a stepwise trend, the step or 'threshold' was always located at 2 (MT, DA) or 3 (AC, DA) stored counterexamples. It is interesting to note that both for the disablers [0% MP vs. 11% MT; n = 19, p<.08] and alternatives [17% AC vs. 39% DA; n = 18, p<.08] the stepwise trends seemed to pop up especially for the 'denial' inferences DA and MT.

For completeness, we report that the number of alternatives also affected MT [Rao R(3,15) = 9.14, p<.002] acceptance. As for DA and AC, the MT trend had a linear nature [F(1,17) = 22.89, MSe = .21, p<.001]. MP also tended to decrease with the number of alternatives, but the trend was not significant. Likewise, AC and DA acceptance showed an opposite trend with increasing acceptance when two and more disablers were available, but the effect did not reach significance.

Discussion

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

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

One should note here that in Experiment 1 we neither observed an effect of additional alternatives on DA. Interestingly, the evidence for additional counterexample retrieval in the latency findings of De Neys et al. (2002) was also only clear for the MP and AC inferences.

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

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

A critic could utter that the retrieval pretest in Experiment 2 showed us only the number of counterexamples a participant had stored for the different conditionals. Obviously, there is no direct evidence that these stored counterexamples were actually retrieved during the reasoning task. Here, it is crucial to stress the relation with the findings of Experiment 1. The explicit presentation illustrated the kind of effect the different number of counterexamples should have on the inference acceptance. The fact that (at least for MP and

AC) the same linear trends were observed in both experiments supports the additional counterexample retrieval hypothesis.

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

With respect to the possible procedural complications it is interesting to note that there is also converging evidence for the present findings. In a recent thinking-aloud study (Verschueren, Schaeken, De Neys, & d'Ydewalle, 2003) without generation pretest or explicit counterexample presentation we observed for example that participants spontaneously produced two, three, or more counterexamples in the evaluation of a single MP or AC argument. Such a result would be hard to explain if people would stop the search after retrieval of a single counterexample. The consistent results in these experiments indicate that the additional retrieval findings should not be attributed to a procedural artefact.

GENERAL DISCUSSION

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

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

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

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

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

Markovits' reasoning model

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

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

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

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

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

The recent work of Schaeken, Vander Henst, and Schroyens (in press) on isomeric mental models might however propose a solution to the 'computational complexity' caveat. The authors argued that people can construct more economical mental models than traditionally assumed. They showed that when two models are redundant in that they share the same information, people can combine them into a single 'isomeric' model. Results indicated that with indeterminate relational inferences (e.g., Bart is to the left of Mark. Mark is to the right of Jan. Is Bart to the right of Jan?) instead of constructing two possible specific models (e.g., Jan → Bart → Mark and Bart → Jan → Mark) people rather constructed a single isomeric model (e.g., Bart ⇔ Jan → Mark) that represented the same crucial indeterminacy. The basic idea is that people will avoid building a model of a piece of information that is already represented. This idea can be extended to the present situation.

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

Daily life or formal deductive reasoning?

Our study did not examine the counterexample retrieval process in a formal, deductive reasoning task, but rather in a situation closer to everyday life reasoning. Participants were not specifically instructed to reason logically and were allowed to give a graded acceptance rating. Therefore, if a sharp distinction is maintained between formal and daily life reasoning, the present findings should not be immediately generalized to reasoning in a formal, deductive reasoning task. Note once more that Markovits developed his model in the context of formal, deductive reasoning. Hence, the present findings do not necessarily refute Markovits' original search process specification. That is, it might be the case that in a pure deductive reasoning task, people do stop the search and take only one counterexample into account. The present work can be best characterized as an adaptation and extension of Markovits' search process specification towards everyday life reasoning. This extension is nevertheless crucial for the final evaluation of Markovits' model. Accounting for peoples daily life reasoning behavior is considered the ultimate goal of any reasoning model

(Johnson-Laird, 1983; Oaksford & Chater, 1998). The present findings do indicate that the model will need to be fine-tuned to encompass daily life reasoning.

Probabilistic reasoning models

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

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

The nature of inference suppression

We finally note the relevance of the present findings for the debate on the nature of the suppression effect (see Byrne et al., 1999 for an overview). Byrne has always stated that the suppression effect arises because whenever a counterexample is available and explicitly represented, certain inferences are no longer supported. The inference is thus suppressed but the status of the conditional itself remains unaffected. However, from the findings on reasoning with uncertain conditionals (e.g. George, 1997; Liu et al., 1996; Stevenson & Over, 1995), it has been argued that suppression arises because the counterexample may lead people to doubt the conditional. Conditionals would be interpreted probabilistically and a counterexample would directly lower the certainty status of the conditional itself.

Traditionally, the graded suppression effects of manipulating P(q/p) on MP and MT acceptance in these studies have been interpreted as support for the 'conditional doubt' position.

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

CONCLUSION

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

Appendix

Table A1

The Conditionals and Counterexamples Adopted for Experiment 1 (translated from Dutch)

ALTERNATIVES

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

But,
If An takes off some clothes, she will also feel cool
If An opens a window, she will also feel cool
If An takes a shower, she will also feel cool
If An turns on the fan, she will also feel cool

2. If fertilizer is put on plants, then they grow quickly

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

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

But,
If Mark is cribbing, he will also do well on the test
If the test is easy, he will also do well on the test
If Mark is lucky, he will also do well on the test
If Mark is very smart, he will also do well on the test

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

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

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

But,
If the fire dies out, the fire will also go out
If the fire is smothered with sand, the fire will also go out
If it rains, the fire will also go out
If there's a lot of wind, the fire will also go out

Table A1 (continued)

DISABLERS

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

But,
If the test is very hard, he will not do well on the test
If John is not concentrated, he will not do well on the test
If John is not smart enough, he will not do well on the test
If John studied the wrong subject, he will not do well on the test

2. If the match is struck, then it lights

But,
If the match is wet, the match will not light
If the match is not struck hard enough, the match will not light
If the matchbox pad is worn, the match will not light
If the match was already used, the match will not light

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

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

4. If fertilizer is put on plants, then they grow quickly

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

5. If the ignition key is turned, then the car starts

But,
If the engine is broken, the car will not start
If the wrong key is used, the car will not start
If the fuel tank is empty, the car will not start
If they key is not turned far enough, the car will not start

CHAPTER 5

Working memory and the retrieval and inhibition of stored counterexamples

Four experiments examined the contribution of Working Memory (WM) capacity to the retrieval and inhibition of stored counterexamples 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 lager 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.

Based on: De Neys, W., Schaeken, W., & d'Ydewalle, G. (2003). *Working memory and everyday conditional reasoning: Retrieval and inhibition of stored counterexamples.* (Tech. Rep. No. 295). Leuven, Belgium: University of Leuven, Laboratory of Experimental Psychology.

INTRODUCTION

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 is 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 tasks (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; Rumain, 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 (see also Markovits & Quinn, 2002). 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 involved in 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. There is a dissonance between searching disablers for the MP and MT inferences 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; Klaczynski, 2001a; Stanovich & West, 2000). It is assumed that all people have a basic "contextualisation" tendency to search stored background knowledge (e.g., counterexamples) associated with the reasoning problem. However, individual difference studies indicate that at least people of high cognitive capacity also appear to have a logical, "decontextualisation" tendency: A basic ability to put background knowledge aside when it conflicts with the logical standards (e.g., Klaczynski, 2001a, 2001b; Stanovich & West, 2000). In these studies cognitive ability is typically operationalized in terms of scores on general intelligence tests 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 tendency to retrieve 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 unbelieve conclusion like 'Some of the communist 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 of 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, 2002). 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 x 1) – 4 = 2 ? BALL
IS (3 x 4) – 5 = 8 ? MAN

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

The Ospan score 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 .05. For completeness, we always report the individual estimated p-values.

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 (highs spans) and bottom quartile (low spans) of the WM-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 x 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 involved in 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 a rather 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 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 effortlessly and without errors. Therefore, following Kane and Engle, we did not give accuracy feedback in the cascade condition. In order to force people to keep tapping at a consistent rate we did give response time feedback starting from the 60 s practice trial in the cascade condition. As 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.

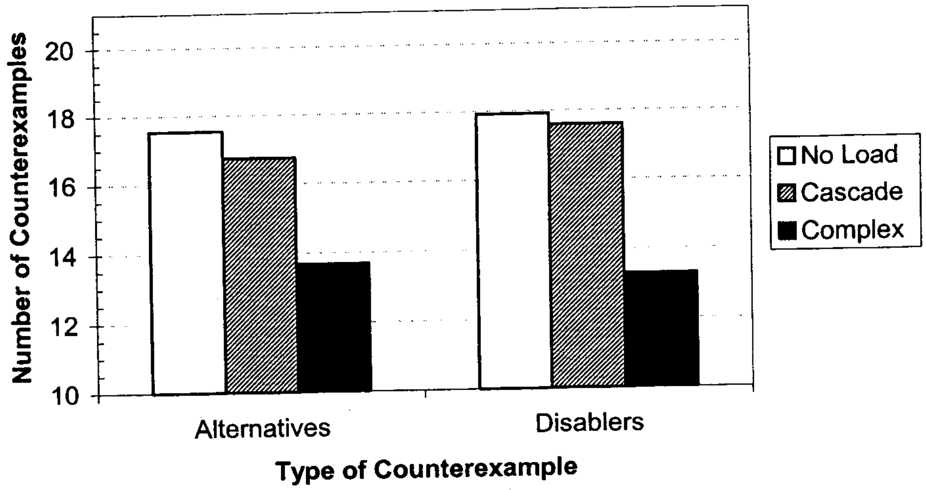


Figure 1. Mean number of generated alternatives and disablers for the eight conditionals in the generation tasks when participants concurrently tapped a cascade or complex finger pattern. The "No Load" condition presents the mean generation performance of participants in the Cascade and Complex condition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

The next analysis checked whether the complex tapping effects on counterexample generation differed for high and low spans. Following the quartile criteria of Experiment 1, 19 out of the 64 participants 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 marginal main effect

[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.

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

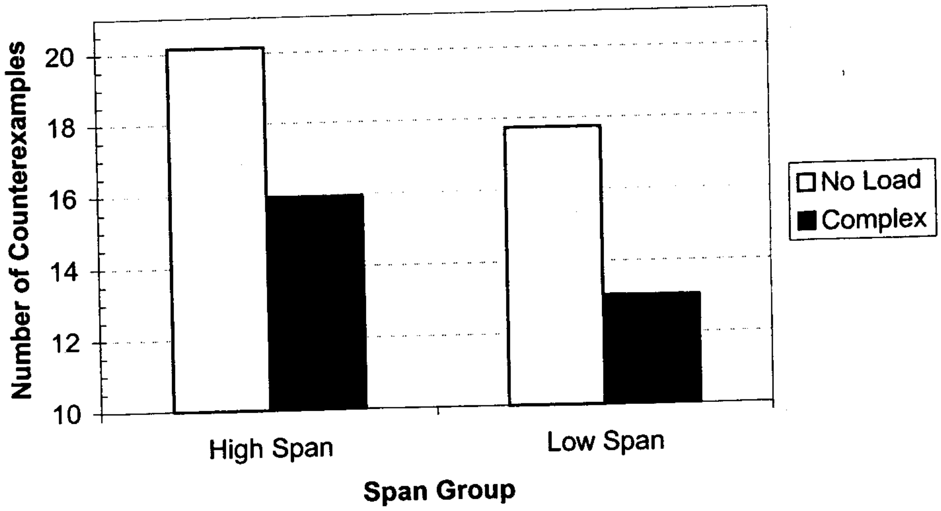


Figure 2. Mean counterexample generation performance of participants classified as High or Low Span when there was no secondary task imposed ("No Load") and when concurrently tapping the complex finger pattern.

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.

Table 1

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

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

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 tended to tap slightly 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 important 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, the present findings suggest that retrieving disablers and alternatives is equally demanding for WM.

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

Rosen and Engle (1997) proposed an interesting component model of memory retrieval. 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 then 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.

It should be stressed that the present counterexample search results do fit within the general Rosen and Engle (1997) retrieval model. The finding that WM-load also affects low spans' counterexample retrieval is not surprising if on takes the different demands of the monitor component in both retrieval tasks into account. 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-resources to actively generate cues. Because of the higher level of available resources, high spans will nevertheless be more successful in the cue generation.

The importance of the active cue generation for successful counterexample retrieval is also apparent if on looks at the nature of the counterexamples for the causal conditionals adopted in this study. As Markovits and Barrouillet (2002; see also Oaksford & Stenning, 1992) noted, counterexamples for causal conditionals resemble what Barsalou (1983) called "ad hoc" categories. In contrast to common categories (e.g., 'animal names') ad hoc categories are less well established in memory. While instances of a common category can be accessed relatively direct, 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.

In sum, the findings point to a clear-cut involvement of WM in the retrieval of stored counterexamples for causal conditionals. It should be clear that, following the Rosen & Engle (1997) model, this does not imply that there is no role for an automatic retrieval process. Indeed, it is explicitly assumed that the retrieval starts with an automatic, passive spreading of activation. The point is that except in specific instances (e.g., counterexamples that have a very low activation threshold, see for example Quinn & Markovits, 1998) the automatic process will not be very successful for causal conditionals. The recruitment of WM-resources will allow a more active and efficient retrieval.

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-a). The inhibition hypothesis states that that high spans will use their WM-resources to inhibit spontaneous disabler access. Remember that we stated that the counterexample retrieval process starts with an automatic spreading of activation after which WM-resources will be recruited for a more active search. It is assumed that when it concerns retrieving disablers, high spans will not use WM-resources for an active search but rather for an inhibition of the automatic disabler activation. If the hypothesis 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.

Following 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. Having an intuitive notion of some basic logical principle does not guarantee that one will always draw correct, logical inferences (Jacobs & Klaczynski, 2002; Klaczynski, 2001a). It is explicitly claimed that inhibiting the retrieval process is difficult and resource demanding. Although high spans might have some logical competence, the actual performance will depend on the demands of the inhibition process. We proposed that, except in specific cases, the 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 (i.e., stored counterexamples with a low activation threshold). The strength of association between a counterexample and a conditional has been shown to affect successful retrieval (De Neys et al., in press-a; Quinn & Markovits, 1998). Conditionals with many counterexamples have typically also more 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. Under the assumption that high spans' disabler inhibition is an attention demanding process we can expect that an increase in the inhibition demands will result in a less successful inhibition. 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-a, 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. (2002)). 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., 2002).

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:

			I			
------1------	------2------	------3------	------4------	------5------	------6------	------7------
Very Sure	Sure	Somewhat Sure	I I	Somewhat Sure	Sure	Very Sure
			I			
That I CANNOT draw this conclusion			I			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 tended to be 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.

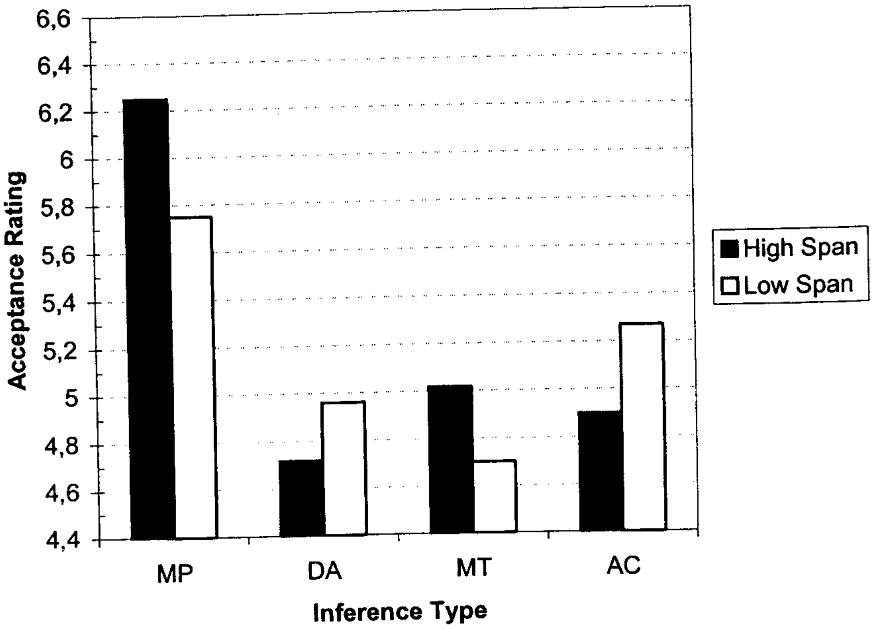


Figure 3. High and Low Spans' mean acceptance rating of the four inference types. The rating scale ranged from 1 (very sure cannot draw this conclusion) to 7 (very sure can draw this conclusion).

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.

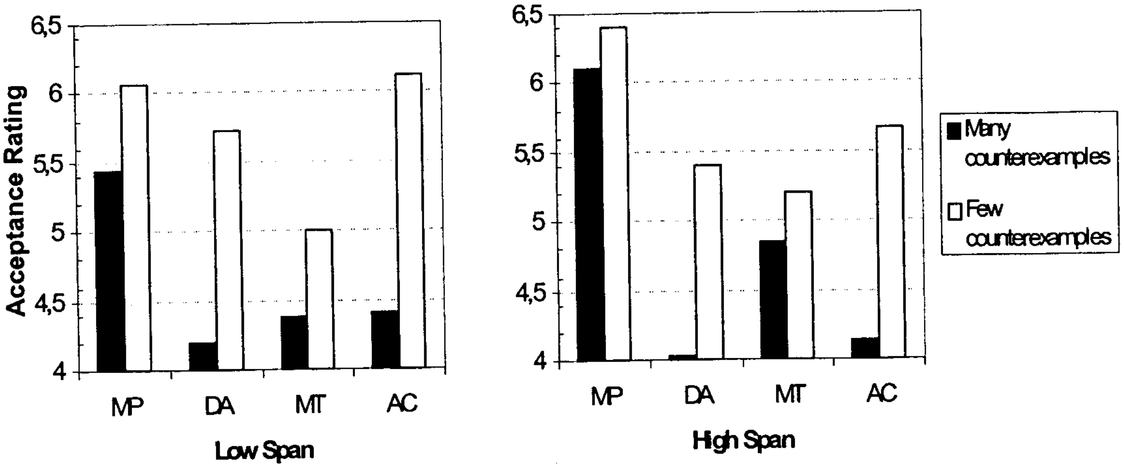


Figure 4. High and Low spans' mean inference acceptance ratings in function of the number of available disablers (MP and MT) and alternatives (AC and DA). The rating scale ranged from 1 (very sure cannot draw this conclusion) to 7 (very sure can draw this conclusion).

The effect of the number of disablers did not interact with span group, F(1, 50) > 1. We did observe a marginal 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 tended to reach 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, $p < .025$.

It could be argued that the effect of the number of disablers on high spans' MP and MT acceptance in the present analysis resulted from an aggregation confound. That is, the overall number-of-disabler effect might be caused by a subset of high spans, while the vast majority would not be affected. However, an examination of the individual acceptance patterns indicated there were only 2 out of 26 high spans (1 out of 26 low spans, $p_1 = .08$, $n_1 = 26$, vs. $p_2 = .04$, $n_2 = 26$, $p > .25$) whose mean MP and MT ratings were not affected by the number of disablers factor. Thus, it is rather unlikely that the reported findings should be attributed to an aggregation confound.

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 higher 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-a; Quinn & Markovits, 1998). Conditionals with many counterexamples have typically also more 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. Therfore, the inhibition will be somewhat 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 hypotheses.

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, and disabler retrieval will decrease 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.

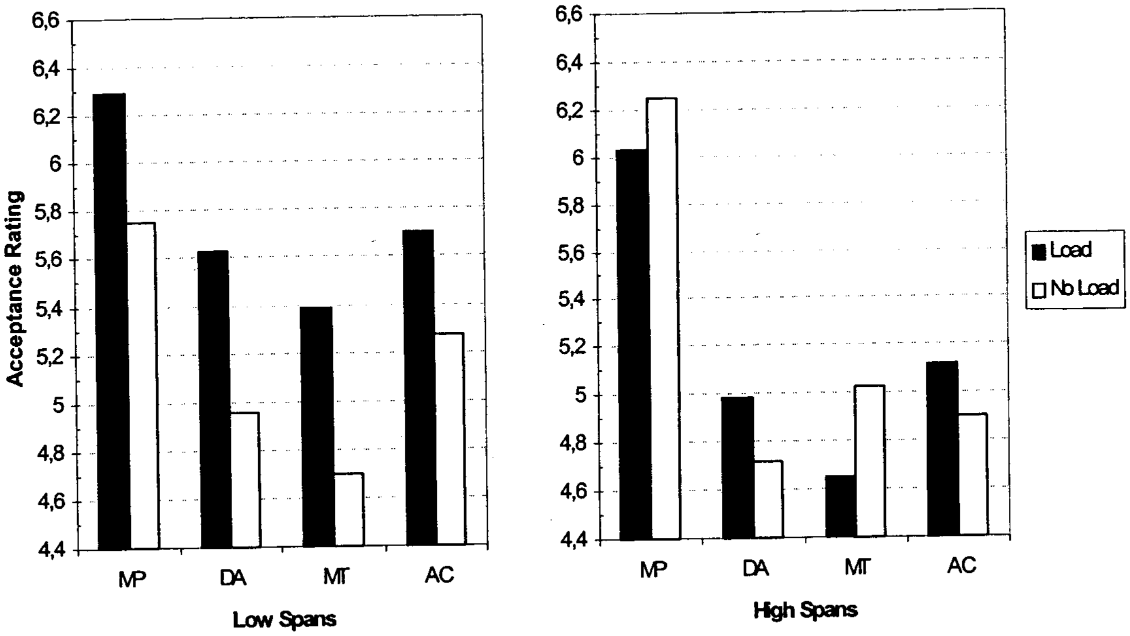


Figure 5. Low and High spans' mean acceptance rating of the four inference types while concurrently tapping the complex finger pattern ("Load") and when there was no secondary task imposed ("No Load"). The rating scale ranged from 1 (very sure cannot draw this conclusion) to 7 (very sure can draw this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

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 tended to interact 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. As Figure 5 indicates, the WM-load tended to increase AC and DA acceptance while it decreased MP and MT acceptance, F(1, 81) = 3.74, MSE = 3.17, p < .06.

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. As expected, the load impact on MP, F(1, 81) = 5.9, p < .02, and MT, F(1, 81) = 5.28, p < .025, differed 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 presents 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 shows the results.

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 x 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.

Table 2

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

	Period			
	Baseline		Reasoning	
Span Group	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

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 established the link between the decreased search efficiency and inference acceptance. This supports the hypothesis that WM-capacity is important 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 at retrieval. When this retrieval becomes less likely under load, MP and MT will be more accepted.

Although it should be noted that some individual effects in the study reached only marginal significance, a consistent and stable pattern emerged over the different experiments that lent credence to the findings. Taken together the results clearly corroborate the hypothesized role of WM.

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 gets 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 primarily 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 rather unlikely. Consequently, DA and MT acceptance should increase under load.

To our knowledge the present experiment is one of the first ones that used a dual task methodology to study reasoning with meaningful, realistic conditionals[2]. 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 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 purely 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

[2] In a different context, Oaksford, Morris, Grainer, and Williams (1996) already used dual task methodology to assess the impact of mood states on performance in a hypothesis testing task with a realistic conditional rule.

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 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-a, in press-b, 2002; 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. We also point to the broader implications of the findings for the role of logic in everyday life.

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. As suggested by many authors, it is assumed that activation will automatically start to spread from the information stored in WM (or "the focus of attention" see Cowan, 1995) to related long-term memory elements (Anderson, 1993; Cowan, 1995; see also Markovits & Barrouillet, 2002). 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 a much more efficient retrieval than the passive spreading of activation. Educated, adult reasoners (e.g., undergraduate students) will typically use WM-resources for an 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 an 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 the disablers having 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 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. Typically, previous demonstrations of peoples ability to adhere to the normative standard of first-order logic have explicitly instructed people to adhere to this norm (Evans & Over, 1996; Markovits et al., 2002; Stanovich and West, 1998; but see Klaczynski, 2001b). 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.

As in most reasoning studies, we referred to first-order, 'textbook' logic as the logical norm (Evans, 2002). Note however that despite its widespread use in psychological reasoning studies the status of standard logic as the correct normative system for conditional reasoning is debated (e.g., Edgington, 1995; Evans, 2002; Oaksford & Chater, 1998). Logicians have constructed alternative logical systems with different validity characteristics. When we claim that participants higher in WM-capacity manage to inhibit the disabler retrieval, no claims are made about the quality of the reasoning process. It is not claimed that high spans are 'better' reasoners. One could argue that low spans adhere to a different normative system where there is simply no need for a disabler inhibition. However, the fact that high spans do tend to adhere to a standard logical norm must at least give pause for thought before discarding the notion of a standard logic-based normative rationality.

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.

Appendix A

Table A1

The Conditionals for the Counterexample Generation Tasks of Experiment 1 and 2

ALTERNATIVES GENERATION TASK

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

Experiment 2:
If fertilizer is put on plants, then they grow quickly.
If the gong is 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.

DISABLERS GENERATION TASK

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

Table A2

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

Table A3

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")

	Low Span								High Span							
	MP		DA		MT		AC		MP		DA		MT		AC	
Counterexample Group	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

Notes on the manuscript

NOTE 1

One might wonder whether high and low spans' different reasoning performance can be simply attributed to a different task interpretation. High spans would interpret the reasoning task as a logical task, while low spans would perceive it as a task where as much background information as possible needs to be taken into account. We concur that at a meta-level high spans' disabler inhibition may be described as the result of a different, 'logical' task interpretation. However, one should be aware that proposing a difference in task interpretation as the actual explanation for the present findings will beg the question. Indeed, one would still need to specify why the task interpretation of both groups differs. At the end one will need to assume here that high spans base their interpretation on some elementary logic notion that low spans lack. Furthermore, one would need to explain how a different interpretation results in a different performance. Here, one will probably also need to take some kind of inhibition phenomenon into account. Thus, we object against any mere reference to a difference in task interpretation as explanation for the present findings because it does not explain anything at the psychological processing level.

One could also argue that since task instructions did not instruct participants to reason logically, high spans are actually erring in that they mistakenly treat an everyday reasoning task as a 'logical' one. We believe that such an argument is fallacious and detracts attention from the crucial message. As argued above, it is not clear at present what the correct norm for everyday reasoning should be. It might even be the case that different reasoners adhere to different norms. Therefore, we stressed that the fact that high spans seem to adhere to some basic, standard logical principle does not allow one to draw any conclusions about the 'quality' or 'correctness' of the reasoning process. Along the same lines it makes no sense to label high spans' performance as erroneous by changing the norm. The crucial finding in the present study is that in a situation where people have to evaluate everyday conditional inferences, high pans spontaneously take count of a standard logical principle. The findings indicate that this notion affects performance by means of a WM-dependent inhibition process. Whatever the normative status of standard logic may be, the message is that, to some extent, it is mediating high spans everyday reasoning. One may wonder why high spans do adhere to this particular norm. Although such considerations fall outside the scope of the present study, they are intriguing and deserve further research. Our point is that one should refrain from wondering why high spans adhere to a *faulty* norm. Coughing the discussion in these evaluative terms is not warranted and may prevent further progress.

NOTE 2

We assumed that the specific instances where the automatic counterexample search might be successful will be counterexamples that are strongly associated with the conditional. Markovits and Quinn (2002) measured the time participants needed to retrieve an alternative as indicator of retrieval efficiency. Interestingly, results showed that the retrieval times of strongly associated alternatives were less good predictors of the retrieval efficiency (as measured by the tendency to reject AC/DA inferences) than the retrieval times of alternatives with a lower associative strength. Retrieval times for the strongly associated alternatives showed only small inter-individual variation. It is generally assumed that contrary to strategic, active cognitive processes there is only little inter-individual variation in the efficiency of automatically operating cognitive processes (e.g., Evans & Over, 1996; Klaczynski, 2001a; Sloman, 1996; Stanovich & West, 2000). Therefore, Markovits and Quinn's findings support the present claim that the retrieval of a strongly associated counterexample depends on an automatic retrieval process, while retrieval of less strongly associated counterexamples requires an active, WM-resource dependent search.

NOTE 3

Our study treated the executive WM-resources as a domain-free attentional capacity and thereby adopted the WM-model of Engle and colleagues (2001; Engle et al., 1999; Engle & Oransky, 1999). While Engle's model attracted the interest of numerous North-American researchers, most European WM-research is based on the traditional Baddeley and Hitch (1974) model. One of the differences between these models is that Engle conceives WM-resources as being modality a-specific and does not distinguish a separate verbal and visual WM-resource pool (for a review of the different WM-models, see Engle & Oransky, 1999, and Miyake & Shah, 1999).

Some researchers have claimed that Engle's WM-task (i.e., the OSPAN) taps only the verbal WM-resource pool (e.g., Miyake et al., 2000). Although the claim is heavily debated one might nevertheless wonder about the possible consequences for our findings. Here it can be noted that the WM-study of Markovits et al. (2002) was conducted within the Baddeley and Hitch (1974) paradigm and used two different tests to measure a specific verbal and visual WM-capacity. Results showed that at least with meaningful, causal conditionals both tasks predicted reasoning performance equally well. As Engle would expect, these results indicate that it is unlikely that the nature of our WM-task is biasing the findings.

NOTE 4

The present study has some important implications for dual processing theories of reasoning. The recent dual processing theories distinguish 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 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 (2001a) 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 'contextualisation' or System-1 process) is not purely automatic but depends on WM-resources or controlled processing. This questions the characterization of System-1 processes as automatic and effortless.

CHAPTER 6

Working memory and retrieval: A trend analysis

This study presents further evidence for the role of working memory (WM) capacity in the retrieval and inhibition of counterexamples (alternatives and disablers) during everyday conditional reasoning. In Experiment 1, participants were given a measure of WM-capacity and a reasoning task with everyday, causal conditionals. Results showed that the acceptance ratings of the Modus Ponens (MP) and Modus Tollens (MT) inferences follow a quadratic, U-shaped trend in function of WM-capacity, while acceptance ratings of the Affirmation of the Consequent (AC) and Denial of the Antecedent (DA) inferences follow a negative linear trend. In Experiment 2 and 3, we tried to replicate the MP findings with extended MP problems that explicitly mentioned a possible disabler. Explicit disabler presentation was assumed to stimulate the disabler retrieval. Results established that acceptance ratings of the extended MP problems still followed a U-shaped, quadratic trend in function of WM-capacity. Contrasting performance with extended and standard MP problems indicated that all span groups were affected by the explicit presentation manipulation. Findings support the claim that it are specifically high spans that manage to inhibit the spontaneous disabler search and underline the generality and robustness of this inhibition phenomenon.

Based on: De Neys, W., Schaeken, W., Dieussaert, K., & d'Ydewalle, G. (2003). *Working memory span and disabler retrieval inhibition: A trend analysis*. Manuscript submitted for publication.

INTRODUCTION

Suppose that you are given the following conditional information: "*If you pull the trigger, then the gun fires*". Next, you are also told that "*The gun fired*". Would you draw the conclusion "*The trigger was pulled*" on the basis of this information? Likewise, suppose that in addition to the conditional "*If the match is struck, then it lights*" you are told that "*The match is struck*". Would you then draw the conclusion "*The match will light*"? In the present study we examine how the extent to which people do draw such conditional conclusions is related to their working memory capacity.

Research on conditional reasoning typically focuses on peoples performance on four kinds of arguments: Modus Ponens (MP, e.g., 'If p then q, p therefore q'), Affirmation of the Consequent (AC, e.g., 'If p then q, q therefore p'), Modus Tollens (MT, e.g., 'If p then q, not q, therefore not p'), and Denial of the Antecedent (DA, e.g., 'If p then q, not p, therefore not q'). The first (p) part of the conditional is called the antecedent and the second (q) part is called the consequent. In standard logic, MP and MT are considered valid inferences, while AC and DA inferences are considered fallacies. Thus, standard logic would tell you to reject the AC conclusion "*The trigger was pulled*" in the first introductory example and to accept the MP conclusion "*The match will light*" in the second example.

When people reason with realistic, content-rich conditionals, the inferences they draw are affected by prior knowledge about the conditional. For example, when one thinks about the fact that the match might be wet, one might be reluctant to infer that the match will light when it is struck. Since people typically reason with content-rich conditionals in daily-life it is crucial for cognitive reasoning theories to address the impact of background knowledge on the inference acceptance (Johnson-Laird & Byrne, 1994, 2002; Oaksford & Chater, 1998). In the last few years this issue has become one of the main foci of interest in the literature (e.g., Byrne, 1989; Byrne, Espino, & Santamaria, 1999; Cummins, 1995; De Neys, Schaeken, & d'Ydewalle, 2002; Markovits & Barrouillet, 2002; Thompson, 1994, 2000; Rumain, Connell, & Braine, 1983).

At least two important kinds of information stored in long-term memory have been shown to affect the inference acceptance: Alternative causes and disabling conditions. An alternative cause (alternative) is a condition, besides the original antecedent, that can bring about the consequent (e.g., lighting the match with another fire in the introductory example). A disabling condition (disabler) is a condition that prevents the antecedent from bringing about the consequent (e.g., the match being wet in the introductory example).

It is well established that when reasoners retrieve an alternative during conditional reasoning they will tend to reject the fallacious AC and DA inferences (e.g., Rumain et al., 1983; Cummins, 1995; Janveau-Brennan & Markovits, 1999; Quinn & Markovits, 1998). Retrieval of a disabler results in rejection of the valid MP and MT inferences (e.g., Byrne, 1989; Cummins, 1995; Thompson, 1994; De Neys, Schaeken, & d'Ydewalle, 2002, in press-a, in press-b). The impact of disablers and alternatives on the inference acceptance is known as the suppression effect (Byrne, 1989). Further on, we adopt Byrne's terminology and refer to alternatives and disablers as counterexamples.

Recently, peoples conditional reasoning performance with realistic, causal, conditionals has been related to working memory (WM) capacity (De Neys, Schaeken, & d'Ydewalle, 2003; Markovits, Doyon, & Simoneau, 2002). De Neys et al. showed that WM affects everyday conditional reasoning by mediation of the counterexample retrieval process. In a first experiment participants were asked to generate as much counterexamples as possible in limited-time for a set of conditionals. Results indicated that participants higher in WM-capacity were better at retrieving alternatives and disablers. Findings implied that in addition to an automatic counterexample search process based on a passive spreading of activation, people also allocate WM-resources to a more active and efficient search process: The larger the WM-resource pool is, the more resources can be allocated to the search, and the more efficient the search will be.

In a further experiment (De Neys et al., 2003, Experiment 3) a group of low and high spans (participants in the bottom and top quartile of first-year psychology students' WM-capacity distribution, respectively) were tested in an everyday conditional reasoning task. Because of the more efficient alternative retrieval, De Neys et al. reasoned that since retrieval of an alternative is known to decrease acceptance of the AC and DA inferences, high spans (vs. low spans) should be more likely to reject the fallacious AC and DA inferences. On the other hand, based on findings of Stanovich and West (2000), it was assumed that high spans could have a minimal notion of the norms of standard logic. Remember that in standard logic MP and MT are valid inferences. Since disabler retrieval will result in the rejection of MP and MT, a basic validity notion should conflict with the disabler retrieval process. De Neys et al. reasoned that high spans would therefore use their WM-resources for an active inhibition of the spontaneous disabler search. Thus, when it concerns retrieving disablers, high spans would not use their WM-resources for an active search but rather for an inhibition of the automatic disabler retrieval. Despite the better intrinsic retrieval capacities for high spans, this inhibition process should result in higher MP and MT acceptance ratings for the high (vs. low) spans. Results of the study supported the predictions.

In a final dual-task study the hypotheses were further confirmed. The basic assumption stated that lower spans allocate WM-capacity to the disabler retrieval, while high spans allocate WM-capacity to the retrieval inhibition. Consistent with the hypothesis, the dual-task study showed that a less efficient disabler retrieval under WM-load resulted in higher MP acceptance ratings under load (vs. no load) for low spans, while the less efficient inhibition resulted in lower MP ratings under load (vs. no load) for high spans.

Thus, there is evidence for the claim that high spans are inhibiting the disabler retrieval process during conditional reasoning. Inhibition of cognitive processes deemed inappropriate is indeed one of the key executive working memory functions (e.g., Baddeley, 1996; Levy & Anderson, 2002; Miyake & Shah, 1999). The basis of the inhibition during reasoning would be high spans' minimal notion of the standards of first-order logic. High spans would adhere to the logical principle that the truth of the antecedent implies the truth of the consequent. This principle excludes the possibility that the consequent does not occur when the antecedent occurs (i.e., a disabler).

Note that De Neys et al. (2003) explicitly assumed that the inhibition would only occur for people highest in WM-capacity. If this assumption is correct it follows that people with medium WM-capacities should show the lowest MP and MT acceptance. Indeed, on one hand medium spans (vs. high spans) should not inhibit the disabler retrieval. On the other hand, medium spans will have a more efficient counterexample retrieval than low spans because they can allocate more resources to the search. Thus, the disabler retrieval during conditional reasoning should be most successful for medium spans. Consequently, it is expected that the MP and MT acceptance ratings in function of WM-capacity follow a U-shaped curve: Due to the limited resources, people lowest in WM-capacity will not be very successful at retrieving disablers and should therefore show rather high levels of MP and MT acceptance. Because of the more efficient disabler retrieval, MP and MT acceptance should decrease for the medium spans. Because of the disabler inhibition, MP and MT acceptance ratings should increase again for reasoners higher in WM-capacity.

Since retrieving alternatives results in the rejection of AC/DA inferences and accepting AC/DA is erroneous in standard logic, there is no conflict between a basic logical notion and the retrieval of alternatives. More precisely, it is assumed that the basis of high spans' disabler inhibition is a minimal notion of the logical principle that the truth of the antecedent implies the truth of the consequent. While this notion conflicts with the possibility that the consequent does not occur when the antecedent occurs (i.e., a disabler), it does not conflict with the possibility that the consequent occurs in the absence of the antecedent (i.e., an alternative). Thus, the process where alternatives are retrieved from long-term memory should not be inhibited. Moreover, De Neys et al. (2003) already showed that the extent to which high spans accept AC and DA is mediated by the alternative retrieval. Therefore, the higher WM-capacity is, the more efficient the alternative retrieval will be, and the less AC and DA should be accepted. Contrary to MP and MT, AC and DA acceptance ratings should therefore follow a negative linear trend in function of WM-capacity. These predictions were tested in Experiment 1.

Experiment 2 and 3 present additional evidence for the disabler inhibition claim. In these experiments we constructed extended MP problems where a possible disabler was explicitly mentioned (e.g., see Byrne, 1989). We reasoned that the explicit disabler presentation would stimulate the search process. An explicitly presented disabler can serve as additional retrieval cue: If a possible disabler is presented, it will be incorporated in the elementary mental representation of the inference problem. It is assumed that this representation is held in working memory. As suggested by many authors, activation will automatically start to spread from the information stored in WM (or "the focus of attention" see Cowan, 1995) to related long-term memory elements (Anderson, 1993; Cowan, 1995; see also Markovits & Barrouillet, 2002 for an integrated account). Thereby, the given disabler will directly prompt activation towards other disablers. Thus, the spreading of activation will be more focused and stored disablers will receive relatively more activation. Consequently, it will be more likely that additional disablers will be retrieved (i.e., the activation reaches the critical retrieval threshold). Since the explicit disabler presentation will result in an increased disabler activation, the inhibition demands for the high spans will also increase.

If high spans would still manage to block the stimulated retrieval to some extent, we expect to replicate the higher MP acceptance for the high spans compared to medium spans. This prediction was tested in Experiment 2. In Experiment 3, we compared the acceptance ratings of standard and extended MP problems for the different WM-span groups. Because the inhibition is assumed to be resource demanding, higher inhibition demands should affect the efficiency of the inhibition process. Therefore, although high spans are expected to show overall a higher MP acceptance, even high spans' ratings should be affected (i.e., decrease) by the explicit disabler presentation. These experiments will allow us to establish the generality and robustness of the inhibition phenomenon.

EXPERIMENT 1

In Experiment 1, a large sample of participants were given a measure of WM-capacity and a conditional reasoning task with everyday, causal conditionals. By examining the specific trends in the acceptance ratings of the MP, MT, AC, and DA inferences over the whole WM-capacity distribution the trend predictions could be tested. The experiment completes the previous work of De Neys et al. (2003) and Markovits et al. (2002). De Neys et al. (2003) only compared a group of participants from the bottom and top quartile of the WM-capacity distribution. Thus , the crucial 'medium' span group was missing. Markovits et al. (2002) only looked at the linear correlation between WM-capacity and inference acceptance. Given the a priori expectation of a quadratic, U-shaped function for MP and MT, a mere linear correlation analysis is not informative here.

Method

Participants

A total of 292 first-year psychology students from the University of Leuven (Belgium) participated in the experiment in return for course credit. None of the students had had any training in formal logic.

Material

Working memory task. Participants' working memory capacity was measured using a version of the Operation span task (Ospan, La Pointe & Engle, 1990) adapted for group testing (Gospan, for details see De Neys, d'Ydewalle, Schaeken, & Vos, 2002). In the Ospan-task participants solve series of simple mathematical operations while attempting to remember a list of unrelated words The main adaptation in the Gospan is that the operation from an operation-word pair is first presented separately on screen (e.g., 'IS (4/2) – 1 = 5 ?'). Participants read the operation silently and press a key to indicate whether the answer is correct or not. Responses and response latencies are recorded. After the participant has typed down the response, the corresponding word (e.g., 'BALL') from the operation-word string is presented for 800 ms. As in the standard Ospan three sets of each length (from two to six operation-word pairs) are tested and set size varies in the same randomly chosen order for each participant. The Gospan-score is the sum of the recalled words for all sets recalled completely and in correct order.

Participants were tested in groups of 21 to 48 at the same time. 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 (participants already in the bottom quartile of the Gospan-score distribution were not discarded based on the latency criterion). De Neys, d'Ydewalle et al. (2002) reported an internal reliability coefficient alpha of .74 for the Gospan. The corrected correlation between standard Ospan and Gospan-score reached .70.

Reasoning task. Sixteen causal conditionals from the generation study of De Neys et al. (2002) and Verschueren, De Neys, Schaeken, and d'Ydewalle (2002) were selected for the reasoning task (see Appendix). The number of possible counterexamples of the selected conditionals varied systematically. The number of counterexamples constituted a 2 (few/many) x 2 (alternatives/disablers) design with four conditionals per cell. Two conditionals in each cell were embedded both in the MP and DA inferences, while the other two were embedded both in the AC and MT inferences. This produced 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:

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

Type down the number that best reflects your decision about the conclusion:_

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

Procedure

All participants started with the Gospan task. After a 5 min break the reasoning task was presented. Reasoning task instructions were presented 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. Participants used the keypad to type down the number reflecting their decision. The experimental session was preceded by one practice trial.

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. Participants received the sets in the order MP/AC, MT/DA, MT/DA, MP/AC. Eight conditionals were used for the first two sets and the remaining eight conditionals for the inferences in the last two sets. Thus, the inferences in the first and last two sets were always based on different conditionals. After two sets (i.e., 16 inferences) were evaluated, item presentation was paused until participants decided to continue.

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.

Results

Ten participants were discarded (about 3.5 % of the sample) because they did not meet the operation correctness or latency requirements of the WM-task. The remaining 282 participants were split in five span groups based on the quintile boundaries of the Gospan-score distribution. Mean Gospan-score for the successive span groups was 18.21 (SD = 4.1), 26.97 (SD = 1.69), 32.26 (SD = .99), 37.56 (SD = 1.35), and 47.32 (SD = 5.07) for the fifth and top quintile group.

Each participant evaluated eight inferences of the same inference type. The mean of these eight observations was calculated. These means were subjected to a 5 (span group) x 4 (inference type) mixed model ANOVA with span group as between-subjects factor and inference type as within subjects factor. The effects of inference type were analyzed with multivariate ANOVA tests.

There was a main effect of span group, F(1, 277) = 3.31, MSE = 2.35, p < .015, and inference type, Rao R(3, 275) = 180.16, p < .0001, and the two factors also interacted, Rao R(12, 727) = 1.83, p < .045. More specifically, the impact of span group on AC and DA differed from the impact of span group on MP and MT, F(1, 277) = 2.72, MSE = .33, p < .03. Figure 1 shows the results.

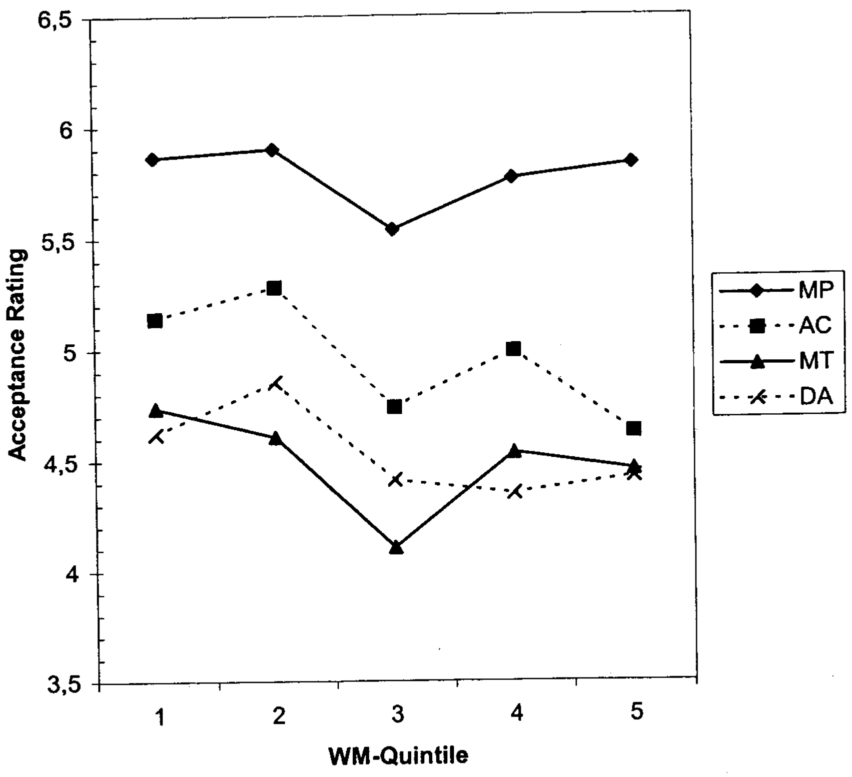


Figure 1. Mean acceptance rating of the four inference types for participants in the successive span groups. WM-Quintile 1 stands for the bottom quintile. The rating scale ranged from 1 (very sure cannot draw this conclusion) to 7 (very sure can draw this conclusion).

As figure 1 indicates, the MP and MT acceptance ratings indeed followed a U-shaped curve with people in the middle quintile ('medium' spans) showing the lowest acceptance levels. The trend analysis established the quadratic nature of this trend for MP, F(1, 277) = 3.13, MSE = .55, p < .08, and MT, F(1, 277) = 3.3, MSE = 1.33, p < .075. Neither on MP, F(1, 277) < 1, nor on MT, F(1, 277) = 1.88, MSE = 1.33, p > .17, the quadratic trend was mediated by a linear trend.

As expected, AC and DA acceptance showed a different pattern with higher span groups tending towards lower acceptance ratings. The trend analysis established the linear trend[1] in the AC, F(1, 277) = 11.97, MSE = .9, p < .001, and DA ratings, F(1, 277) = 3.95, MSE = 1.25, p < .05, while a quadratic trend was absent, both F(1, 277) < 1.

In order to make sure there was indeed a linear trend without quadratic trend mediation on AC/DA on the one hand, and a quadratic trend on MP/MT on the other hand, we ran an analysis on the combined MP/MT and DA/AC acceptance ratings. This more powerful analysis confirmed the findings [linear trend, F(1, 277) = 8.87, MSE = 1.71, p < .005, without quadratic trend mediation, F(1, 277) < 1, on AC/DA, and a quadratic trend, F(1, 277) = 4.46, MSE = 1.30, p < .04, without linear trend, F(1, 277) = 1.67, MSE = 1.30, p > .2, on the combined MP/MT ratings].

Discussion

The present results establish that MP and MT acceptance ratings follow a quadratic, U-shaped trend in function of WM-capacity, while the AC and DA acceptance rating patterns follow a negative linear trend. MP and MT acceptance ratings were lowest for the 'medium spans' in the middle quintile of the WM-capacity distribution. This pattern is precisely what one would expect if the disabler inhibition during conditional reasoning occurs only for people highest in WM-capacity. In the absence of a disabler inhibition process, medium spans can allocate more WM-resources to the disabler retrieval than lower spans. Because high spans will inhibit the disabler retrieval, the search will be most successful for the medium spans. Consequently, the middle group shows the lowest MP and MT acceptance ratings.

It is assumed that the basis of high spans' disabler inhibition is a minimal notion of the logical principle that the truth of the antecedent implies the truth of the consequent. While this notion conflicts with the possibility that the consequent does not occur when the antecedent occurs (i.e., a disabler), it does not conflict with the possibility that the consequent occurs in the absence of the antecedent (i.e., an alternative). Thus, the process where alternatives are

[1] The pearson product moment correlation between span group and mean DA and AC acceptance was -.11, n = 282, p < .06, and -.21, n = 282, p < .001, respectively.

retrieved from long-term memory should not be inhibited. Successful retrieval of alternatives results in the rejection of the AC and DA inferences. Since higher WM-capacity allows a more efficient retrieval, higher WM-resources will lead to lower AC and DA acceptance ratings in the reasoning task. Consistent with this framework, contrary to MP and MT, AC and DA acceptance linearly decreased for the successive span groups. These findings further demonstrate the crucial role of WM as mediator of the counterexample retrieval during everyday conditional reasoning.

In an earlier study, De Neys et al. (2002, Experiment 3) measured the efficiency of the disabler retrieval process in a generation task and linked this with performance in a conditional reasoning task. It was observed that a more efficient disabler retrieval process resulted in lower MP and MT acceptance ratings. Later, De Neys et al. (2003) found that although high spans were better at retrieving disablers, they nevertheless accepted MP and MT more than low spans. These findings seem to contradict each other. However, De Neys et al. (2002) only measured retrieval efficiency of 40 randomly selected participants (WM-capacity was not assessed). De Neys et al. (2003) specifically selected participants from the top and bottom quartile of first-year psychology students' WM-capacity distribution. As De Neys et al. (2003) argued, the top WM-spans were probably small in number in the random sample of 40 participants. The present data show that in that case higher WM-capacity (and thus better retrieval) should indeed result in lower MP and MT acceptance. Hence, the trend analysis allows us to reconcile both studies.

The quadratic trend in the MP and MT acceptance ratings implies that, in general, people with the highest and lowest WM-capacity will accept MP and MT to the same extent. Given these findings it makes sense that the effects De Neys et al. (2003) observed when MP and MT acceptance of a sample of students from the bottom and top quartile of the WM-capacity distribution was compared, were only moderate. When the performance of a selected group of high and low spans is contrasted (e.g., see De Neys et al., 2003), ratings may differ depending on the relative position of the participants in both groups on the WM-capacity distribution but the differences will never be strongly pronounced. High spans' MP and MT acceptance can be best contrasted with medium spans' ratings. The trend analysis provides the crucial bigger picture here. This underlines the importance of examining the inference acceptance patterns over the whole WM-capacity distribution.

EXPERIMENT 2

The quadratic MP and MT trends in Experiment 1 supported the claim that reasoners highest in cognitive capacity inhibit the disabler inhibition during everyday conditional reasoning. Experiment 2 and 3 present an additional test of the inhibition hypothesis. As Byrne (1989), we constructed extended MP problems where a possible disabler was explicitly mentioned. We reasoned that the explicit disabler presentation would increase the inhibition demands by a stimulation of the disabler search process. If a possible disabler is added to the MP premises, it will be incorporated in the elementary mental representation of the inference problem. This representation is held in working memory. As suggested by many authors, activation will automatically start to spread from the information stored in WM (or "the focus of attention" see Cowan, 1995) to related long-term memory elements (Anderson, 1993; Cowan, 1995; see also Markovits & Barrouillet, 2002). Thereby, the incorporated disabler will directly prompt the spreading activation towards the stored disablers that will receive more activation. Furthermore, explicitly presenting a disabler can also contribute to a more active stimulation of the search process. It has been argued that mentioning a disabler conveys the message to the participants that taking disabling information into account is the relevant thing to do in the inference evaluation (e.g., Bonnefon & Hilton, 2002; Politzer & Bourmaud, 2002). Therefore, the mere presentation of a disabler with the standard premises might actively trigger people to start searching disablers. Consequently, we can assume that it will be more likely that additional stored disablers will be retrieved (i.e., the activation reaches the critical retrieval threshold). Precisely because the explicit disabler presentation will result in an increased disabler activation, the inhibition demands for the high spans should also increase.

In Experiment 2 participants were given a measure of WM-capacity and extended MP problems that mentioned a possible disabler. If high spans still manage to inhibit the stimulated disabler search to some extent, we expect to replicate the U-shaped, quadratic trend in the acceptance ratings in function of WM-capacity: Both low and medium spans will benefit from the additional search stimulation but overall medium spans should still be more successful at retrieval because they can allocate more WM-resources to the active disabler search. Therefore, medium spans should show lower MP ratings than low spans. Because of the retrieval inhibition, acceptance ratings should increase again for the high spans.

Method

Participants

A total of 105 first-year psychology students from the University of Leuven (Belgium) participated in the experiment in return for course credit. None of the students had had any training in formal logic.

Material

Working memory task. Participants' working memory capacity was measured with the version of the Operation span task (Ospan, La Pointe & Engle, 1990) adapted for group testing (Gospan, for details see De Neys, d'Ydewalle et al., 2002).

Reasoning task. Participants received three extended MP problems. These were Dutch translations of the three Byrne (1989) MP problems (see Dieussaert, Schaeken, Schroyens, & d'Ydewalle, 2000). The following item format was used:

> Rule: If she has an essay to write, then she will stay late in the library.
>
> If the library stays open, then she will stay late in the library.
>
> Fact: She has an essay to write
>
> Conclusion: She will stay late in the library.

All three MP problems were presented on a separate page of a booklet together with a 7-point rating scale ranging from 1 (*Very certain that I cannot draw this conclusion*) to 7 (*Very certain that I can draw this conclusion*) with 4 representing *can't tell.* Participants placed a mark on the number of the scale that best reflected their evaluation of the conclusion.

Procedure

Participants were tested in groups of 21 to 42 at the same time in a large computer room with an individual booth for every participant. All participants started with the Gospan task that was run on computer. After all participants of a group had finished the Gospan-task the extended MP evaluation task was presented. The three items were presented on separate pages of a booklet. The first page of the booklet included the task instructions. 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. The task instructions did not mention to accept the premises as true or to endorse conclusions that followed necessarily. Instead participants were told they could evaluate the conclusions by the criteria they personally judged relevant.

Results and discussion

Three participants were discarded because they did not meet the operation correctness or latency requirements of the WM-task (see De Neys, d'Ydewalle et al., 2002). The remaining 102 participants were split in three span groups of equal n based on the boundaries of the Gospan-score distribution. Mean Gospan-score for the three successive span groups was 23.27 (SD = 4.34, low span), 35.15 (SD = 2.79, medium span), and 45.89 (SD = 4.97, high span).

For every participant we calculated the mean acceptance rating for the three extended MP problems. The means were subjected to an ANOVA with span group as between-subject variable. There was a significant effect of span group, F(2, 99) = 5.55, MSE = 1.23, p < .01. The acceptance rating showed the expected pattern: Medium spans (M = 3.84, SD = 4.67) showed lower MP acceptance ratings than the low (M = 4.56, SD = 1.11) and high spans (M = 4.67, SD = .99). A trend analysis confirmed that there was a significant U-shaped, quadratic trend, F(1, 99) = 10.85, MSE = 1.23, p < .005 without mediation of a linear trend, F(1, 99) < 1.

Thus, even when the disabler search was specifically stimulated high spans showed the highest levels of MP acceptance. This is consistent with the claim that high spans are inhibiting the disabler search and underlines the generality and robustness of the inhibition phenomenon.

EXPERIMENT 3

Experiment 2 showed that the acceptance ratings for the extended MP problems differed for participants of different WM-capacity. In Experiment 3, we compare the acceptance ratings of standard vs. extended MP problems in function of WM-span. This allows us to establish the impact of the explicit disabler presentation per se. For the validity of our framework it is crucial that the acceptance ratings decrease when a disabler is explicitly presented. First, for low spans it is assumed that the disabler search with standard MP problems will not be very successful. Although one can argue that low spans' limited resources will restrict the impact of the extra search stimulation, the extended disabler manipulation does present low spans a disabler they will probably not retrieve in the standard condition. Therefore, low spans' inference acceptance should decrease for the extended MP problems. Second, because of the more efficient retrieval, medium spans in the standard condition will probably retrieve the disabler presented in the extended MP condition themselves. Hence, the disabler per se does not provide new information that would directly affect medium spans' ratings. However, if we are right that the search process is stimulated by the disabler presentation one should expect that in the extended condition additional disablers will be retrieved and this should decrease the MP acceptance (see De Neys, Schaeken, & d'Ydewalle, in press-b, for a study on the effect of the number of retrieved disablers on MP acceptance). High spans are expected to inhibit the search both for the standard and extended problems. However, it is explicitly assumed that the inhibition is not automatic, but draws on WM-resources. Therefore, the inhibition should be less successful when the process is more demanding. De Neys et al. (2003) already observed that the increasing inhibition demands caused by an increasing number of available disablers or a secondary task load resulted in a less efficient disabler inhibition. Hence, although the high spans should overall show a high MP acceptance, their acceptance level should be nevertheless affected by the stimulated disabler search.

In sum, we expected a standard suppression effect for all span groups: Acceptance ratings should be lower for the extended (vs. standard) MP problems. In addition, overall MP acceptance ratings should be affected by WM-span: Extended and standard MP acceptance in the successive span groups should follow the U-shaped trend observed previously.

Method

Design

As standard condition or baseline we used the MP evaluations of the 282 participants in Experiment 1. Remember that these participants were presented a standard conditional inference task with causal conditionals and a measure of WM-capacity. We calculated the mean MP acceptance for different span groups and used this as a baseline to compare the MP acceptance of matched span groups with similar extended causal MP problems.

Participants

All 105 participants of Experiment 2 evaluated the extended MP inferences in the present experiment. The data for the standard MP condition were taken from Experiment 1 where 282 first-year psychology students evaluated standard conditional inferences.

Material

Working memory task. All participants' working memory capacity was measured with the Gospan-task (see De Neys, d'Ydewalle et al., 2002).

Reasoning tasks. All conditionals were selected from the generation studies of De Neys et al. (2002) and Verschueren, De Neys, Schaeken, & d'Ydewalle (2002). Eight causal conditionals were used for the standard condition and six causal conditionals for the extended condition (see Appendix). Half of the conditionals in each condition were previously classified as having many possible disablers, while the other half had only few possible disablers. The number of possible alternative causes (see Cummins, 1995) of the selected conditionals with few and many disablers was kept constant. The item format for the extended and standard task was similar except that for the extended items a possible disabler was mentioned. We always presented the disabler that was most frequently generated for that conditional in the generation task (e.g., see De Neys et al. in press-a, 2002). Following Byrne (1989), the disablers (e.g., engine broken) were always presented as an additional requirement, embedded in a conditional (e.g., If the engine works, then the car starts). This resulted in the following format:

Rule: If the ignition key is turned, then the car starts.
If the engine works, then the car starts.
Fact: The ignition key is turned.
Conclusion: The car starts.

It should be noted that the set of conditionals in the standard and extended condition was not completely similar. Although both conditions used causal conditionals with a comparable number of possible disablers, the standard condition should therefore not be conceived as a control condition per se. Rather, the standard condition serves as a baseline against which the performance of the different WM-span groups can be compared.

Procedure

Participants were tested in groups of 21 to 48 at the same time in a large computer room with an individual booth for every participant. All participants started with the Gospan task that was run on computer. After all participants of the group had finished the Gospan-task, the extended MP evaluation task or the standard conditional inference task was presented. The standard task was run on computer. Participants evaluated eight standard MP inferences mixed with other conditional inferences. The six items of the extended MP task were presented on separate pages of a booklet. This booklet was presented before the booklet with the items of Experiment 2. Task instructions for the extended MP task were similar to the instructions given in Experiment 2.

Results and discussion

In order to match the span groups in the extended and standard conditions as closely as possible we decided to split both samples up in five span groups each, based on the quintile boundaries of the Gospan-score distribution of the 282 participants in the standard condition. A 5 (span group, between-subjects) x 2 (MP task, between-subjects) ANOVA on the Gospan-scores established that there were no WM-capacity differences for participants in both task conditions [effect of MP task, $F(1, 374) < 1$; interaction MP task x Span-group, $F(4, 374) = 1.84$, $MSE = 10.91$].

Each participant evaluated eight or six MP evaluations. The mean of these ratings was calculated and subjected to a 5 (WM-span, between-subjects) x 2 (MP task, between-subjects) ANOVA.

Explicitly presenting a disabler clearly decreased the MP acceptance, $F(4, 374) = 37.56$, $MSE = .72$, $p < .0001$. Figure 2 shows that, as expected, this effect was present for all WM-span groups, span group x MP task interaction, $F(4, 374) < 1$. There was also a marginal main effect of WM-span, $F(4, 374) = 2.28$, $MSE = .72$, $p < .06$. As Figure 2 indicates, a trend analysis clearly established that the MP ratings followed a U-shaped, quadratic trend in function of WM-span, $F(1, 374) = 6.77$, $MSE = .72$, $p < .01$. There was no sign of a linear trend, $F(1, 374) < 1$, and the quadratic trend did not differ for the standard and extended MP problems, $F(1, 374) = 1.07$, $MSE = .71$, $p > .35$. Thus, as expected, all span groups showed an impact of the explicit disabler presentation, but both on the standard and extended problems the MP acceptance ratings were affected by WM-capacity.

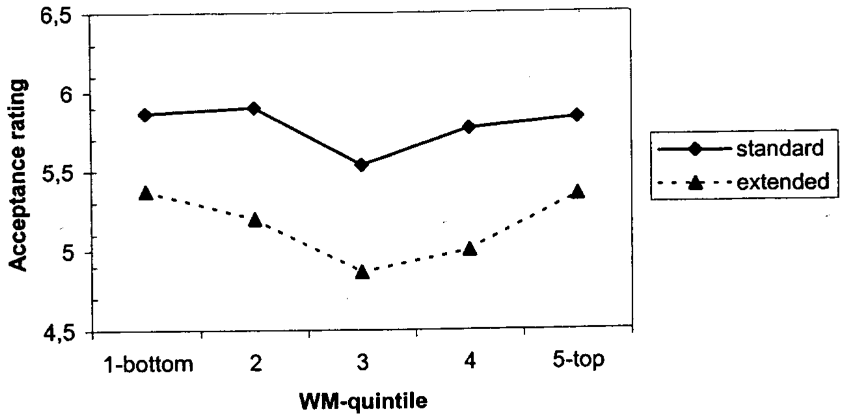


Figure 2. Mean MP acceptance rating in function of WM-capacity with (extended) and without (standard) explicitly presented disabler. The rating scale ranged from 1 (very sure cannot draw this conclusion) to 7 (very sure can draw this conclusion).

The present results confirm the finding that even when a disabler is explicitly presented MP acceptance ratings of the successive working memory (WM)-span groups follow a U-shaped trend. Previous studies already established that the higher MP acceptance ratings of high vs. medium spans, despite high spans' superior retrieval capacities, result from an active inhibition of the disabler search. The fact that in the present study the same pattern is found under conditions that can be assumed to stimulate the search process points to the generality of the inhibition phenomenon.

As in previous work (De Neys et al., 2003) we hypothesized that the disabler inhibition is not occurring in a cognitive vacuum but draws on working memory resources. Therefore, higher inhibition requirements were expected to result in a less efficient inhibition process. The goal of the search stimulation by the explicit disabler presentation was precisely to increase the inhibition demands. Disablers that would be inhibited under less demanding inhibition conditions could 'slip through' the filter and decrease the MP acceptance. Consistent with these hypotheses the results clearly showed that even high spans' MP acceptance decreased for the extended MP problems.

GENERAL DISCUSSION

In the present study we examined trends in the relation between WM-capacity and reasoning performance with everyday conditionals. Experiment 1 established that MP and MT acceptance ratings follow a quadratic, U-shaped trend in function of WM-capacity, while the AC and DA acceptance rating patterns follow a negative linear trend. This pattern supports the claim that people highest in WM-capacity inhibit the disabler search during everyday conditional reasoning. Experiment 2 and 3 replicated the quadratic MP trend findings with extended problems that explicitly mentioned a possible disabler. The explicit disabler presentation was assumed to increase the inhibition demands. Experiment 3 established that the efficiency of high spans' inhibition decreased when the inhibition demands increased. These findings generalize previous inhibition findings and underline the robustness of the phenomenon.

Markovits et al. (2002) calculated linear correlations between reasoning performance and WM-capacity in a conditional reasoning task with the kind of causal conditionals adopted in the present study. Consistent with the findings of Experiment 1, Markovits et al. found significant correlations for AC and DA: Higher WM-capacity resulted in a more frequent rejection of the AC (r = .21) and DA (r = .20) inferences. However, Markovits et al. also reported a smaller but significant linear correlation for MP. In the present study there was clearly no sign of a linear trend on MP. We suspect that the difference with the present findings lies in the task instructions. Markovits et al. were interested in the study of formal, deductive reasoning with realistic conditionals. Therefore participants were explicitly instructed to reason logically (i.e., participants had to assume that the premises were always true). We are interested in the reasoning process people use in everyday life whatever the nature of this process may be (e.g., deductive or probabilistic). Therefore, in our studies people can evaluate the conclusions by the criteria they personally judge relevant. Our data show that reasoners in the top levels of the WM-distribution are able to inhibit the disabler retrieval. The findings thereby indicate that high spans spontaneously (without being instructed) adhere to a standard logical norm in their reasoning. Markovits' data might suggest that when the norm is explicitly presented even medium span will tend to adhere to it. There is some evidence (George, 1995; Vadeboncoeur & Markovits, 1999) suggesting that stressing the logical nature of the task reduces the MP rejection. Thus, when properly instructed even medium spans might to some extent block the disabler retrieval and consequently show a boost in MP acceptance. Because of the larger WM-capacity pool, the resource demanding inhibition will still be more successful for the high spans. Since disabler retrieval is already unlikely for participants lowest in WM-capacity, the instructions should only have minimal impact on low spans' acceptance ratings. Therefore, one could expect a more (positive) linear trend on MP and MT acceptance with standard instructions. The higher WM-capacity is, the more successful the inhibition will be, and the more MP and MT will tend to be accepted.

An important final remark concerns the status of standard, first-order logic as a normative reasoning system. As in most reasoning studies, we always refer to first-order, 'textbook' logic as the logical norm (Evans, 2002). However, note that despite its widespread use in psychological reasoning studies, the status of standard logic as the correct normative system for conditional reasoning is heavily debated (e.g., Edgington, 1995; Evans, 2002; Evans, Handley, & Over, in press; Oaksford & Chater, 1998). Logicians have constructed alternative logical systems with different validity characteristics. For example, Van Lambalgen and Stenning (2002) worked out a nonmonotonic logic where rejecting MP and MT in the light of possible disablers is considered valid. When we claim that participants higher in WM-capacity manage to inhibit the disabler retrieval, no claims are made about the quality of the reasoning process. It is not claimed that high spans are 'better' reasoners. One could argue that medium and low spans adhere to a different normative system where there is simply no need for a disabler inhibition. However, the disabler inhibition phenomenon does suggest that cognitively skilled reasoners have a basic notion of the standard logical principle that a conditional utterance excludes the possibility that the consequent does not occur when the antecedent occurs. Therefore, the findings do question the opposite claim that standard logic would have no bearing on peoples everyday life reasoning (e.g., Oaksford & Chater, 1998). For some people, to some extent, it has.

Appendix

Table A1

Material for the Experiments

Conditionals used for the reasoning task in Experiment 1:

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.
If the match is struck, then it lights.
If the brake is depressed, then the car slows down.
If water is heated to 100°C, then it boils.
If Marry jumps into the swimming pool, then she gets wet.
If the gong is stuck, then it sounds.
If the correct switch is flipped, then the porch light goes on.
If fertilizer is put on plants, then they grow quickly.
If the apples are ripe, then they fall from the tree.

Conditionals used for the standard MP task in Experiment 3:

If the trigger is pulled, then the gun fires.
If the intensity of light increases, then the pupils of the eyes grow smaller.
If the match is struck, then it lights.
If the brake is depressed, then the car slows down.
If water is heated to 100°C, then it boils.
If Marry jumps into the swimming pool, then she gets wet.
If fertilizer is put on plants, then they grow quickly.
If the apples are ripe, then they fall from the tree.

Material for the extended MP task in Experiment 3:

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

If Marry jumps into the swimming pool, then she gets wet.
If there's water in the swimming pool, then she gets wet.

If the ignition key is turned, then the car starts.
If the engine works, then the car starts.

If the correct switch is flipped, then the porch light goes on.
If the electric power is on, then the porch light goes on.

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

If water is heated to 100°C, then it boils.
If the water is pure, then it boils.

Notes on the manuscript

NOTE 1

We mentioned that the number of counterexamples of the selected conditionals in Experiment 1 constituted a 2 (few/many) x 2 (alternatives/disablers) design. In an additional analysis we entered the number of alternatives and disablers as within-subjects factors in the ANOVA. De Neys et al. (2003) found that both high and low spans showed the standard impact of the number of counterexamples. Here we examined whether the findings for high and low spans could be replicated and whether the effects were also clear for the medium spans. For the analyses we divided our sample in tree span groups (about 100 participants in each group) based on the Gospan-score. Participants with a score lower than 28 were classified as 'low spans', participants with scores above 36 were classified as 'high spans', and the remaining participants were classified as 'medium spans'. Each participant evaluated inferences based on two different conditionals within each 2 (number of disablers) 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 (number of alternatives) x 2 (number of alternatives) x 3 (span group) x 4 (inference type) ANOVA.

Once more the effects of number of alternatives, number of disablers, inference type and their interactions replicated the standard findings of De Neys et al. (2003) and De Neys et al. (2002): The number of alternatives primarily affected AC, F(1, 279) = 936.74, MSE = 1.73, p <. 0001, and DA, F(1, 279) = 652.73, MSE = 1.56, p <. 0001. The alternatives had also a similar but smaller impact on MP, F(1, 279) = 19.78, MSE = .63, p < .0001, and MT, F(1, 279) = 75.48, MSE = .1.39, p < .0001. The number of disablers primarily affected MP, F(1, 279) = 167.17, MSE = .91, p < .0001, and MT, F(1, 279) = 275.9, MSE = 1.86, p < .0001 acceptance. There was also a smaller, opposite impact on AC, F(1, 279) = 122.49, MSE = .67, p <. 0001, and DA acceptance, F(1, 279) = 26.25, MSE = .91, p < .0001.

We explored for each inference type separately whether the effect of the number of counterexamples differed for the three span groups. The effect of disablers interacted with span group on MP, F(2, 279) = 3.6, MSE = .91, p < .03, and MT, F(2, 279) = 4.62, MSE = 1.86, p < .015, while there was no interaction on AC and DA, both F(2, 279) < 1. Both on MP and MT the effect of the number of disablers tended to be larger for the higher span groups.

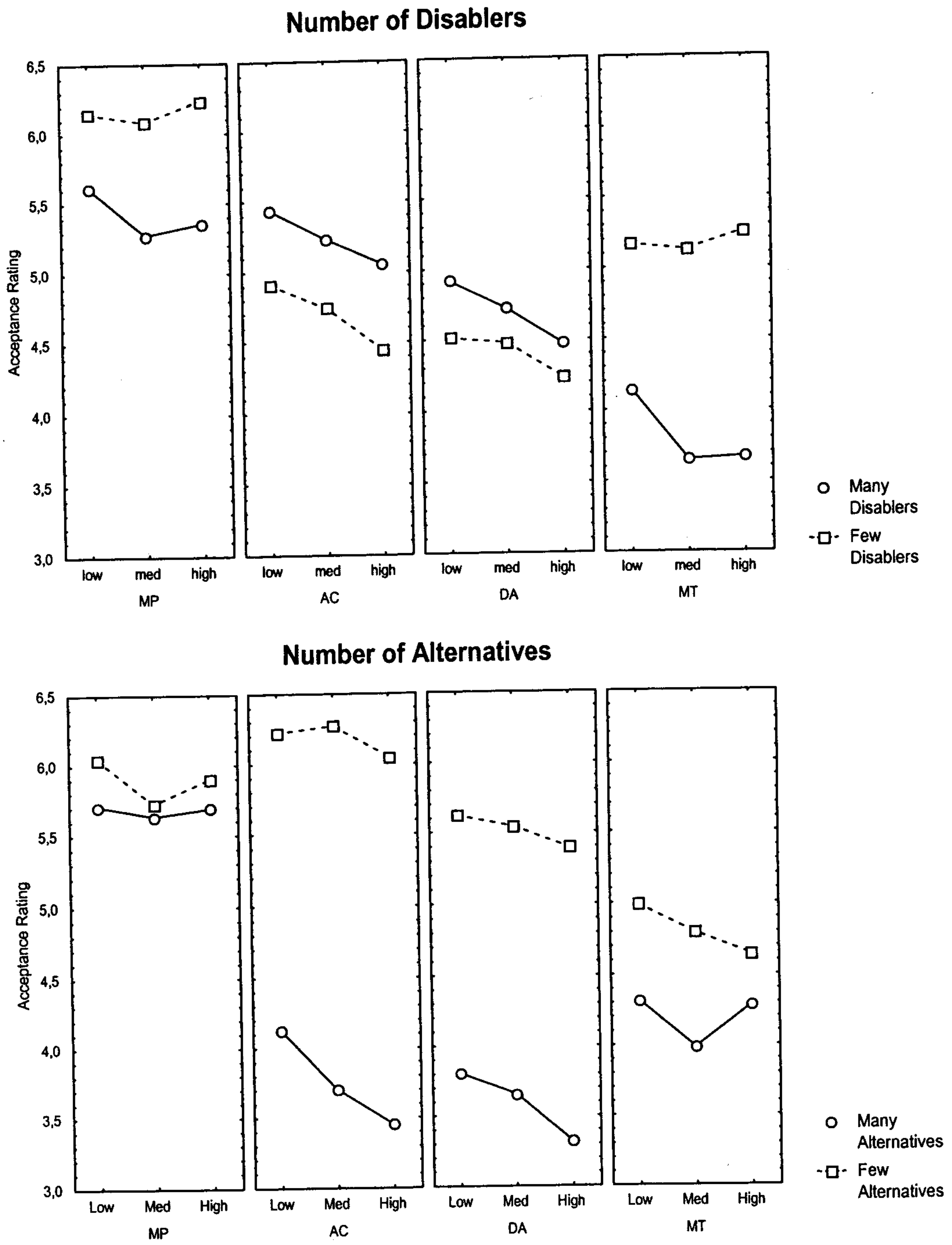


Figure N1. The effect of the number of alternatives and disablers on low, medium, and high spans' inference acceptance.

The effect of alternatives interacted with span group for the AC, F(2, 279) = 4.46, MSE = 1.79, p < .015, and MT, F(2, 279) = 3.71, MSE = 1.39, p < .03, inferences, while the interaction was not significant for DA, F(2, 279) = 1.15, MSE = .1.56, and MP, F(2, 279) = 2.15, MSE = .63. The number of alternatives tended to have a larger impact on the AC ratings of the higher span groups. High spans' MT ratings were somewhat less affected by the number of alternatives compared to the lower span groups.

The crucial finding is that planned contrast test established that for each of the three span groups the impact of the number of alternatives and disablers on the different inferences remained significant (only the impact of alternatives on medium spans' MP acceptance failed to reach significance). This confirms the findings of De Neys et al. (2003). We can conclude that the acceptance ratings of all span groups show the standard impact of the available number of alternatives and disablers of a conditional. Figure N1 gives an overview of the results.

NOTE 2

Together with the acceptance rating our computer program also recorded the time needed for every inference evaluation in Experiment 1. In a similar reasoning task, De Neys et al. (2002) found that AC inferences took more time for conditionals with many (vs. few) alternatives and MP inferences took longer for conditionals with many (vs. few) disablers. We wanted to check whether these basic latency effects could be replicated. We also examined the impact of span group. De Neys et al. argued that the longer latencies reflected a more extended search process in case there were many stored counterexamples. Given the acceptance rating data, we expected that all WM-span groups would have a more extended search process in case there were many alternatives available. Thus, all span groups should show the AC latency effect. We also expected to see the MP effect for all span groups. As with the alternatives, the disabler search should be more extended in case there are many disablers for the low and medium spans. High spans on the other hand will inhibit the disabler retrieval. This inhibition is not automatic, but draws on WM-resources. De Neys et al. (2003) already argued that the inhibition process will be more demanding (more and more strongly associated disablers have to be inhibited) in case there are many available disablers. One can expect that the higher inhibition-processing demands for conditionals with many disablers will result in longer evaluation latencies.

For the ease of interpretation we restricted the analysis to the key effects of the number of alternatives on AC/DA and the effect of disablers on MP/MT acceptance. This resulted in a 3 (span group) x 2 (number of counterexamples) x 4 (inference type) design.

Each participant evaluated inferences based on four different conditionals within each 2 (number of counterexamples) x 4 (inference type) cell of the within-subject part of the design. The mean of these four observations was calculated and subjected to analysis. In order to eliminate biased measures, latency data were trimmed prior to analysis. First, all reaction times longer than 60 s were discarded. Second, any reaction times that were more than 3 SD above a person's mean reaction time were replaced by the 'mean + 3SD' cutoff value. This procedure affected less than 1.3 % of all observations.

Figure N2 shows the results. There was a main effect of number of counterexamples, F(1, 279) = 18.00, MSE = 8.53, $p < .0001$, and inference type, Rao R(3, 277) = 35.66, $p < .001$, and both factors interacted, Rao R(3, 211) = 14.29, $p <.001$. Planned contrast test showed that the number of counterexamples affected MP latencies, F(1, 279) = 29.44, MSE = 6.38, $p < .0001$, and AC latencies, F(1, 279) = 42.19, MSE = 6.06, $p < .0001$, while the effects on MT, F(1, 279) < 1, and DA, F(1, 279) = 2.66, MSE = 7.72, were not significant. This confirms the findings of De Neys et al. (2002).

Neither the main effect of WM-span, nor its interactions with the number of counterexamples and inference type factors were significant, all F or Rao R < 1. Thus, we can conclude that for all span groups the MP and AC latencies show the expected effects of the number of counterexamples (planned contrast tests confirmed this). The patterns in Figure N2 seem to suggest, however, that the effects of WM-span and number of counterexamples might interact for the different inferences. We therefore explored for each inference type separately whether the effect of the number of counterexamples differed for the three span groups (although there was no significant Number of Counterexamples x Inference Type x WM-span interaction). However, none of the interactions reached significance (not even at the .1 level).

De Neys et al. (2002) argued that the additional WM-load caused by the processing of negations for DA and MT inferences overrides the retrieval of additional counterexamples. Consequently, the counterexample search would be less extended when evaluating MT (vs. MP) and DA (vs. AC) inferences (see also De Neys, d'Ydewalle, & Schaeken, in press-b) and therefore latency effects on MT and DA would be less clear. In general, such a mechanism can explain the absence of MT and DA latency effects for the low and medium spans. However, it is assumed that high spans will inhibit the disabler search both on MP and MT. We hypothesized that the MP latency effects for the high spans are caused by the fact that the inhibition is more demanding when many disablers are stored. Now, one could further speculate that if the higher processing requirements for the MT (vs. MP) inference decrease the efficiency of the disabler retrieval, there will be less need for an additional inhibition (since less disablers will be activated). Consequently, it makes sense that the inhibition-dependent latency increase will be less pronounced for MT inferences.

In sum, the only significant latency effect was the overall difference between conditionals with few vs. many disablers on MP and conditionals with few vs. many alternatives on AC. Other trends or contrasts did not reach significance (not even at the liberal .1 level).

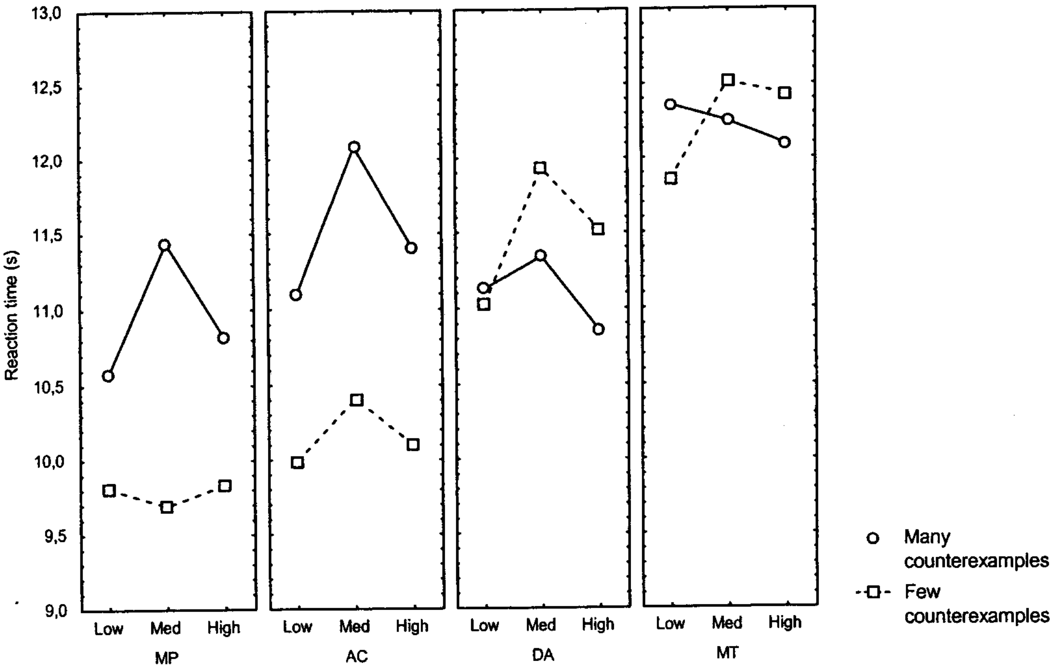


Figure N2. The effect of the number of counterexamples (alternatives for AC/DA, disablers for MP/MT) on low, medium, and high spans' inference latencies.

CHAPTER 7

Where do we stand?
Where do we go from here?

1. WHERE DO WE STAND?

The studies reported in this dissertation specified crucial processing characteristics of the counterexample retrievál during everyday conditional reasoning. We clearly established that the counterexample search is affected by the associative strength of the stored counterexamples. We pinpointed the time course and "stopping" characteristics of the retrieval. We showed that the retrieval is not purely automatic, but draws on WM-resources. Lastly, we clarified that the retrieval can be spontaneously inhibited when the nature of a counterexample conflicts with the logical validity of a reasoning problem.

The final chapter sketches a model of the counterexample retrieval process based on the established specifications of the processing characteristics. I will suggest some ideas and guidelines for further research and discuss some broader considerations.

In general, our findings fit quite well with Markovits' counterexample retrieval model and Rosen and Engle's (1997) component model of memory retrieval. Although we specified extensions and revisions, we also found explicit support for the models' basic assumptions. I believe that the core of these models provides a solid ground for a characterization of the counterexample search process during everyday conditional reasoning. In the next section I present a sketch of the counterexample search process based on the established specification of the retrieval characteristics.

1.1 A sketch of the counterexample search process

When people are confronted with a conditional, they will start constructing a basic mental representation of the elementary information the conditional contains. The elementary information concerns the antecedent and consequent of the conditional and the fact that the occurrence of the antecedent is associated with the occurrence of the consequent. This representation is maintained in working memory. Next, memory structures in long-term (semantic) memory storing relevant information for the evaluation of the conditional will be automatically activated. As suggested by many authors (Anderson, 1993; Cowan, 2001; Logan, 1988), it is assumed that activation will automatically start to spread from the elements in working memory or the 'focus of attention' towards related elements in long-term memory. The stored counterexamples are conceived as nodes in a semantic network (Anderson, 1983). A counterexample will be retrieved when a node's activation level crosses a critical threshold. The height of the retrieval threshold is reflected in the associative strength of a counterexample. More strongly associated counterexamples have lower retrieval thresholds and will be more easily retrieved (De Neys, Schaeken, & d'Ydewalle, in press-a, Chapter 2; Quinn & Markovits, 1998).

When a counterexample is successfully retrieved it will be incorporated in the elementary representation of the reasoning problem that is maintained in working memory. Only those elements that are sufficiently activated will enter working memory and will be available for further inferential processing.

In addition to the passive, spreading of activation, executive working memory (WM) resources will be recruited to monitor the automatic retrieval in order to prevent errors and re-access of previously retrieved counterexamples. Finally, available WM-resources will be used for an active, strategic search (Rosen & Engle, 1997) to access new counterexamples.

Although the automatic search can suffice to activate strongly associated counterexamples, in general, the spreading of activation will not be very successful for causal conditionals. The active search will allow a much more efficient retrieval than the automatic spreading of activation (De Neys, Schaeken, & d'Ydewalle, 2003; Chapter 5).

After successful retrieval of a counterexample, the search process will not stop. If possible, additional stored counterexamples will be accessed and the number of retrieved counterexamples will determine the degree to which inferences are accepted (De Neys, Schaeken, & d'Ydewalle, in press-b, 2002; Chapter 3 and 4).

The extent of the search (i.e., the number of retrieved counterexamples) will be determined by the available WM-resources. The higher one's WM-resources, the more resources can be allocated to the active search, and the more counterexamples will be retrieved (De Neys et al., 2003; De Neys, Schaeken, Dieussaert, & d'Ydewalle, 2003; Chapter 5 and 6). In addition, the search will be affected by inference complexity. As Markovits and Barrouillet (2002), we assume that the additional complexity of the denial inferences (i.e., DA and MT) lies in the required retrieval of a 'complementary' class instance (i.e., cases different from p associated with not-q). In line with Markovits and Barrouillet (see also Schroyens, Schaeken, & d'Ydewalle, 2001), our findings suggest that retrieval of complementary instances for DA and MT has priority over the counterexample retrieval (De Neys et al., 2003; Chapter 5). The complementary class retrieval will burden the available WM-resources. Thus, fewer resources will be available for the counterexample search. Therefore, the search will be less extended and additional counterexample retrieval will be less likely (De Neys et al., in press-b, 2002; Chapter 3 and 4). Furthermore, because of a priority of disabler retrieval over alternative retrieval, WM-resources will be primarily allocated to disabler retrieval and thereby the retrieval of disablers might hinder subsequent retrieval of alternatives (De Neys et al., 2002; Chapter 3).

The retrieval mechanism can also be inhibited . When it concerns retrieving disablers, people from the top of the WM-capacity distribution will not use their WM-resources for an active search but for an inhibition of automatically activated disablers (De Neys et al., 2003; De Neys, Schaeken, Dieussaert, et al., 2003; Chapter 5 and 6). This inhibition might be targeted at the prevention of retrieval (i.e., preventing that the activation level crosses the critical retrieval threshold) or at the discarding of a retrieved disabler at a later stage. The inhibition draws on limited WM-resources. When the inhibition demands increase because more and/or more strongly activated disablers have to be inhibited, this affects the efficiency of the process. Automatically activated disablers that are inhibited under less demanding circumstances will not be sufficiently filtered out and affect the inference acceptance (De Neys et al., 2003; De Neys, Schaeken, Dieussaert, et al., 2003; Chapter 5 and 6).

The framework states that a retrieved counterexample will be incorporated in the elementary representation of the reasoning problem maintained in working memory. We made no claims about the nature of this elementary representation and the further inferential processes that will operate on it (e.g., these might amount to additional mental model construction, adjusting probabilistic parameters, or specific rule selection, see Chapter 1). Our research respects a reasoning theory neutrality. It should be explicitly stressed that this should not be conceived as a shortcoming of the present research. Our work simply stops where current reasoning theories begin. Mental models theory, mental logic, and the probabilistic approach all specify (to some degree) what happens with a retrieved counterexample. The presentation of Markovits' model in Chapter 1 exemplified a possible MMT incorporation. We also clarified how this incorporation could be revised in order to account for our findings (see De Neys et al., in press-b; Chapter 4). Thus, if necessary, we can easily provide an account at the traditional level. However, by leaving the traditional playing field and changing the focus to the neglected and more general search issue, a far wider range of researchers will benefit from our work.

2. WHERE DO WE GO FROM HERE?

2.1 How to take a leap forward?

The studies reported in this dissertation started filling the crucial 'search' gap in traditional reasoning studies. However, I will be the first one to argue that although promising, the work is not yet finished. There are literally dozens of issues that I want to test and specify further. I hope that I will have the opportunity to address these in the future. Below I point to four issues that are of primary importance in my opinion. Anyone interested in the further study of the counterexample search during conditional reasoning should start here.

2.1.1 Types of conditionals and counterexamples

Our studies focused on a specification of the counterexample search with causal conditionals and did not include other possible conditional contents such as promises (e.g., 'If you do your homework, you get some candy'), warnings (e.g., 'If you forget it, I will be mad'), or class-based information (e.g., 'If an animal is a cow, then it has four legs'). Furthermore, we did not distinguish different types of alternatives and conditionals. Indeed, within the disabler category one can, for example, theoretically distinguish 'real disablers' (e.g., 'If the trigger was pulled and the gun had *no bullets*, then the gun would not fire') from 'missing enablers' (e.g., 'If Joe cuts his ginger and *the cut is deep enough*, then it will bleed', see Elio, 1998). The retrieval characteristics for these different materials may differ. From the outset, our research bore the 'try to walk before you try to run' principle in mind. Therefore, we decided to restrict the scope of the research initially. In the meantime, our lab did start exploring reasoning with these different materials (e.g., Dieussaert, Schaeken, & d'Ydewalle, in press, 2002; Verschueren, De Neys, Schaeken, & d'Ydewalle, 2002). This will allow us to extend and generalize the framework. Thereby, the present work makes it possible to start with a direct, model-based prediction testing (instead of starting from scratch with a purely explorative approach).

2.1.2 Retrieval priority

Our findings pointed to a 'priority order' in the retrieval process. For the denial inferences, resources would be primarily allocated to the search for a complementary class instance. In addition, the disabler search would also have priority over the alternative search.

Although our results were consistent with such an ordering and related suggestions have been proposed by others, we lack a direct and explicit test of the assumption.

Furthermore, it is not clear whether the priority should be conceived as a temporal (i.e., first one searches for complementary instances and afterwards for counterexamples) ordering or as a resource allocation ordering (i.e., all types of elements are searched in parallel, but more resources are allocated to specific elements). Cummins (personal communication, May 2002), for example, has suggested a temporal explanation for the disablers/alternatives search priority. According to Cummins, disablers would be primarily associated with the cause of an effect, while alternative causes would be associated with the effect of a cause. The antecedent of a causal conditional typically specifies a cause and the consequent an associated effect. Since people will start reading and processing the antecedent first, activation would therefore primarily start to spread towards the disablers. Such a mechanism makes sense, but it is clear that further testing is necessary here.

2.1.3 Medium spans

Our dual task studies (De Neys et al., 2003; Chapter 5) tested only participants of two extreme capacity groups: Students from the top (high spans) and bottom quartile (low spans) of the WM-capacity distribution. The subsequent experiments (De Neys, Schaeken, Dieussaert et al., 2003; Chapter 6) indicated that in conditions without dual task load, medium spans showed the lowest level of MP acceptance. In future studies, medium spans' reasoning performance should also be tested under dual task load. When low spans' central executive was burdened by a dual task, the MP acceptance level (because of a less efficient retrieval) slightly increased, while high spans' MP acceptance level (because of a less efficient inhibition) decreased under load. The study of De Neys, Schaeken, Diessaert et al. (2003; Chapter 6) indicates that medium spans are actually the most informative contrast group. Low spans' retrieval is already rather inefficient without load. Thus, an additional WM-load can never have a large impact on low spans' acceptance ratings. The medium spans should exhibit a much stronger load impact (i.e., a larger increase in MP/MT acceptance ratings). Therefore, it would be a good idea to compare the load impact for the three different capacity groups directly.

2.1.4 Logic instructions

In our studies participants never received explicit instructions to reason deductively or logically. It would be interesting to assess the impact of instruction manipulations (e.g., George, 1997; Vadeboncoeur & Markovits, 1999) in further experiments. In the instructions or in a short training, participants could be reminded that searching disablers is not in line with standard logic and should be avoided. In the present studies (see Chapter 6) there was no evidence for a disabler inhibition by the medium spans. It is possible that this results from the fact that these people have no intuitive notion of the logical norm and consequently there is no conflict and no reason to inhibit. Alternatively, it is possible that high and medium spans have the same intuitive norm, but medium spans simply lack the additional WM-resources to perform the inhibition. The instructions/training will convey the norm, while the pure inhibition cost will remain the same. Comparing the instruction effects for the different capacity groups (and inference types) should establish why the medium spans do not spontaneously inhibit. Crossing the instruction and dual task manipulation will allow an even more specific test. If medium spans would indeed inhibit the disabler retrieval with explicit instructions, then a dual task load with instructions should result in decreased MP ratings (while the load effect without instructions should show the opposite pattern). These kinds of experiments will take the framework to a further, more fine-grained specification level.

2.2 The logic pendulum

The history of cognitive reasoning research displays a remarkable pendulum swing concerning the role of logic. At the start of the modern psychological reasoning era in the 1950s and 1960s, scholars were heavily affected by the philosophical and psychological tradition of logicism (e.g., Henle, 1962), the doctrine that logic provides the basis for human thought. The influence of Piagetian theory in psychology was enormous. Piaget incorporated the logicist tradition into his theory of cognitive development, proposing that adults eventually developed formal operational thinking on the basis of abstract logical structures (e.g., Inhelder & Piaget, 1958). Typically, the role of logic was not limited to that of a normative standard, but human reasoning itself was conceived as invariably and inevitably logical (e.g., Henle, 1962; Smedslund, 1970). Reasoning was equated with logic. Reasoning competence was conceived as an inherent logicality built into the mind (see Evans, 2002).

In the following decades, studies pointing to the pervasive impact of background knowledge and content and context factors started bringing the original paradigm down (e.g., Cummins, Lubart, Alksnis, & Rist, 1991; Markovits, 1986; Rumain, Connell, & Braine, 1983; Staudenmayer, 1975; Thompson, 1994). These studies progressively deemphasized the importance of logic in human reasoning. Researcher became increasingly convinced that reasoning was in essence a "contextualization" process where accessing background knowledge determined the inference behavior. The logic pendulum started swinging to the other side. Over the course of 40 years the pendulum even reached the opposing extreme with some researcher claiming that logic has no bearing on reasoning whatsoever (e.g., Oaksford & Chater, 1998, 2001).

Recently, under the influence of the studies that looked at individual differences in reasoning (e.g., Klaczynski, 2001a, 2001b; Stanovich & West, 2000), the pendulum started swinging back to the original position. These individual difference studies showed that a select group of people with the highest cognitive capacities do adhere to the logical standards and block the impact of conflicting background knowledge. The studies proposed a dual processing framework where a logical "decontextualization" process would override the standard "contextualisation" tendency.

Looking backwards, it is stunning to notice that the historical pendulum swing also resonated through this dissertation. Our findings typically pointed to the massive impact of the outcome of the counterexample search. In part, the specification of the counterexample retrieval process even provided a solid ground for the characterization of everyday reasoning as an illogical, memory-based process. In this sense, our research helped swinging the pendulum to the "contextualization" extreme. However, our latest studies also provided clear evidence for a spontaneous retrieval inhibition process in case the counterexamples conflicted with the logical standards. This indicated that, to some extent, knowledge of a standard logical norm does have a bearing on everyday conditional reasoning. In other words, the logic pendulum started swinging back.

However, it should be clear that there is at least one fundamental difference between the more traditional reasoning studies and the present research. Traditional reasoning research and the dual processing frameworks have only provided a general description of the "contextualisation" and "decontextualization" processes. The studies reported in this dissertation did not merely focus on the outcome of the "contextualization" or memory search process, but attempted to characterize the underlying processing mechanism. Likewise, the "decontextualization" or logical ability was not simply conceived as a small, black box in the mind which presence somehow results in correct, logical reasoning. We specified that the decontextualization is in fact a WM-dependent inhibition process. We still assume that high spans have an intuitive notion of a basic logical principle. This notion would amount to the knowledge that an 'if, then' utterance excludes the possibility that the consequent does not occur when the antecedent occurs. Nothing more, nothing less. One could, for example, argue that high spans acquire this notion through an abstraction process of continued induction of contingency experiences (e.g., Holland, Holyoak, Nisbett, & Thagard, 1986). However, the crucial point is that our studies started to disentangle the mechanism behind the decontextualization. At the heart of the decontextualization lies an inhibition process. The inhibition is mediated by WM and draws on WM-resources. Although high spans may have a basic logical competence, the actual reasoning performance will depend on the burden the inhibition puts on the available WM-resources.

3. CONCLUSION

The work reported in this dissertation brings us an important step closer to a specification of what happens in the mind when people engage in reasoning. We started to sketch how the crucial counterexample retrieval process during everyday conditional reasoning operates. We discovered some weighty and promising mechanisms. The presented work also pointed to the potential that a further fostering of the approach holds. A continuation of the work advocated here will be of ultimate importance for all those wanting to understand how man manages to do what has been characterized as the essence of his being: Reasoning.

Dutch summary

Stel dat je verteld wordt: 'Als de contactsleutel wordt omgedraaid, dan start de wagen'. Als je dan verneemt dat de wagen start, zal je waarschijnlijk afleiden dat de contactsleutel werd omgedraaid. Echter, stel dat je er op gewezen wordt dat de wagen ook gestart kan worden met een drukknop of door een autodief die de contactdraadjes met elkaar verbindt, dan zal je wellicht minder geneigd zijn om te concluderen dat de contactsleutel werd omgedraaid. Zo zou je uit het feit dat je weet dat de contactsleutel wordt omgedraaid in eerste instantie waarschijnlijk ook besluiten dat de wagen zal starten. Echter, als je zou bedenken dat de benzinetank misschien leeg is of de motor kapot, dan zou je diezelfde conclusie waarschijnlijk niet trekken.

De bekwaamheid om te redeneren met conditionele, 'als, dan' zinnen wordt beschouwd als een van dé bouwstenen van onze mentale capaciteit. Zoals Edgington (1995, p. 235) het stelt, "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". Op dezelfde manier zal het bijvoorbeeld aangewezen zijn om een conditionele inferentie te maken op het moment dat iemand je waarschuwt "Als je niet stopt met me te plagen, word ik agressief". Conditioneel redeneren speelt een centrale rol in ons causaal kennissysteem en in onze sociale interacties. Het is daarom ook niet verwonderlijk dat het conditioneel redeneren is uitgegroeid tot het meest intens bestudeerde domein binnen het onderzoek naar het menselijk redeneren (e.g., Evans, 2002; Evans, Newstead, & Byrne, 1993).

In het dagelijkse leven redeneren mensen meestal met betekenisvolle conditionele zinnen (bv. 'Als de contactsleutel wordt omgedraaid, dan start de wagen'). Zoals het inleidende voorbeeld illustreert, bevat ons lange-termijn geheugen steeds relevante achtergrondkennis over deze conditionele zinnen. Het opzoeken en terugvinden van deze kennis zal beïnvloeden welke conclusie mensen gaan trekken. Als je je herinnert dat wagens niet starten als de brandstoftank leeg is of dat een dief een wagen kan starten door de contactdraadjes te verbinden, zal dat ervoor zorgen dat je bepaalde inferenties niet gaat maken.

Onderzoekers zijn zich al lange tijd bewust van de mogelijke impact van opgeslagen achtergrondkennis op het redeneren (e.g., Matalon, 1962; Staudenmayer, 1975). Hedendaagse redeneertheorieën kunnen reeds verklaren hoe bepaalde teruggevonden achtergrondinformatie het redeneerproces zal beïnvloeden. Echter, de meer fundamentele vraag hoe de informatie eigenlijk gezocht en teruggevonden wordt, is nog niet beantwoord. Dit probleem vormt de kern van deze verhandeling. Het doel is een elementaire specificatie van de karakteristieken van het zoekproces naar opgeslagen achtergrondkennis tijdens alledaags conditioneel redeneren. Het onderzoek zal zich toespitsen op het zoekproces naar twee specifieke types van opgeslagen achtergrondkennis die 'tegenvoorbeelden' genoemd worden: alternatieve en verhinderende omstandigheden.

INLEIDING

Het onderzoek naar de impact van achtergrondkennis op het redeneren is de laatste jaren uitgegroeid tot dé centrale onderzoeksproblematiek binnen het conditioneel redeneerveld (Evans, 2002; Manktelow, 1999). Vooral de rol van de beschikbaarheid van mogelijke alternatieve ('alternative causes') en verhinderende omstandigheden ('disabling conditions') heeft veel aandacht opgeëist.

Een alternatieve omstandigheid is een mogelijke oorzaak die ook kan leiden tot het effect dat in de conditionele zin vermeld wordt (bv. contactdraadjes verbinden in het inleidend voorbeeld). Een verhinderende omstandigheid is een conditie die verhindert dat het vermelde effect zal optreden ook al is de oorzaak aanwezig (bv. een kapotte motor in het inleidend voorbeeld). Bekijk de volgende conditionele zin:

Als Jan de airco aanzet, krijgt hij het koel.

Mogelijke alternatieve omstandigheden zijn hier:

Kleren uitdoen, kouder weer, gaan zwemmen, ...

Dergelijke omstandigheden zullen eveneens maken dat Jan het koel krijgt. De alternatieve omstandigheden maken duidelijk dat de airco opzetten niet noodzakelijk is voor het koel krijgen. Er zijn nog andere mogelijke oorzaken.

Mogelijke verhinderende omstandigheden zijn:

Koorts hebben, de airco is kapot, er staat een raam open, …

Als zulke verhinderende omstandigheden optreden, zal de airco aanzetten er niet voor zorgen dat Jan het koel krijgt. De verhinderende omstandigheden maken duidelijk dat het opzetten van de airco niet voldoende is voor het koel krijgen. Er moet nog aan bijkomende voorwaarden worden voldaan.

Het onderzoek naar conditioneel onderzoek spitst zich toe op de prestatie op vier verschillende conditionele inferenties. In hun abstracte vorm zien die er als volgt uit:

Modus Ponens (MP)	Als p dan q, p dus q
Modus Tollens (MT)	Als p dan q, niet q dus niet p
Negatie van de Antecedent (DA)	Als p dan q, niet p dus niet q
Affirmatie van de Consequent (AC)	Als p dan q, q dus p

Het eerste (*p*) deel van de conditionele zin wordt de antecedent genoemd, het tweede (*q*) deel noemt men de consequent. In de standaard logica wordt het maken van de MP en MT inferenties als valide gezien, terwijl DA en AC inferenties maken een logische fout impliceert. Dus, wanneer je verteld wordt 'Als de contactsleutel wordt omgedraaid, dan start de wagen' en je verder te horen krijgt dat *de contactsleutel omgedraaid werd*, zegt de standaard logica je om te concluderen dat *de wagen zal starten* (een MP inferentie). Op dezelfde manier zou je wanneer je te horen krijgt dat *de wagen niet start*, moeten afleiden dat *de contactsleutel niet werd omgedraaid* (een MT inferentie). Anderzijds zou je volgens de logica op basis van de informatie dat *de wagen start* niet mogen besluiten dat *de contactsleutel werd omgedraaid* (een AC inferentie). Tenslotte zou je volgens de logica uit het feit dat je weet dat de *contactsleutel niet werd omgedraaid* ook niet mogen concluderen dat *de wagen niet zal starten* (een DA inferentie).

In een baanbrekende studie toonden Rumain, Connell, en Braine (1983) aan dat als er expliciet een mogelijke alternatieve omstandigheid aangeboden werd de AC en DA inferenties minder gemaakt werden. Byrne (1989) repliceerde dit effect en stelde vast dat het aanbieden van een verhinderende omstandigheid een gelijkaardig effect had op de MP en MT inferenties. Dus, bijvoorbeeld voor de zin:

> Als ze een essay moet schrijven, dan zit ze tot 's avonds laat in de bib.

leidde het toevoegen van de alternatieve omstandigheid

> Als ze een essay moet schrijven, dan zit ze tot 's avonds laat in de bib.
> Als ze boeken moet lezen, dan zit ze tot 's avonds laat in de bib.

tot minder AC ('Ze zit tot 's avonds laat in de bib. Dus, ze moet een essay schrijven') en DA ('Ze moet geen essay schrijven. Dus, zit ze niet tot 's avonds laat in de bib') inferenties.

Het aanbieden van een verhinderende omstandigheid als:

> Als ze een essay moet schrijven, dan zit ze tot 's avonds laat in de bib.
> Als de bib 's avonds open blijft, dan zit ze tot 's avonds laat in de bib.

zorgde voor minder MP ('Ze moet een essay schrijven. Dus, zit ze tot 's avonds laat in de bib') en MT ('Ze zit niet tot 's avonds laat in de bib. Dus, ze moet geen essay schrijven') inferenties. Deze observaties zijn bekend geworden als het suppressie-effect (zie Dieussaert, Schaeken, Schroyens, & d'Ydewalle, 2000, voor een uitvoerige discussie). In het verdere verloop wordt Byrnes (1989) terminologie overgenomen en zal er naar alternatieve en verhinderende omstandigheden verwezen worden als tegenvoorbeelden ('counterexamples').

Verder onderzoek heeft aangetoond dat het suppressie-effect ook optreedt zonder expliciete aanbieding van de tegenvoorbeelden (zie Cummins, 1995; Cummins, Lubart, Alksnis, & Rist, 1991; Markovits, 1986). Cummins onderzocht de rol van het vinden van opgeslagen tegenvoorbeelden door te kijken naar het effect van het aantal tegenvoorbeelden dat beschikbaar is voor een conditionele zin. In een pretest werden de proefpersonen gevraagd om steeds zoveel mogelijk tegenvoorbeelden te bedenken voor een reeks conditionele zinnen. Aan de hand van deze gegevens werden de zinnen dan ingedeeld in een groep met veel of weinig mogelijke tegenvoorbeelden. De zo geclassificeerde zinnen werden dan gebruikt voor een conditionele redeneertaak met een tweede groep proefpersonen.

Cummins' (1995) resultaten toonden aan dat voor zinnen met veel mogelijke alternatieve omstandigheden (bv. 'Als planten bemest worden, dan groeien ze goed') mensen de AC en DA inferenties minder aanvaarden dan voor zinnen met weinig mogelijke alternatieven (bv. 'Als Piet het glas met zijn blote handen vastneemt, dan staan zijn vingerafdrukken er op'). Terwijl de valide MP en MT inferenties in vrij hoge mate aanvaard werden voor zinnen met weinig mogelijke verhinderende omstandigheden (bv.'Als je water verwarmt tot 100°C, dan kookt het'), zorgde een groot aantal beschikbare verhinderende omstandigheden (bv. 'Als Jan hard studeert, dan doet hij het goed op de test') ook hier voor minder MP en MT inferenties. Deze bevindingen suggereerden dat mensen tijdens een conditionele redeneertaak hun geheugen doorzoeken naar opgeslagen tegenvoorbeelden. Het aantal beschikbare tegenvoorbeelden, het feit dus of er al of niet gemakkelijk tegenvoorbeelden kunnen gevonden worden, bepaalt welke conclusie mensen zullen trekken.

Het wordt algemeen erkend dat een volledige specificatie van het tegenvoorbeeld zoekproces tijdens conditioneel redeneren cruciaal is voor om het even welke redeneertheorie (bv. Johnson-Laird & Byrne, 1994; Thompson, 1994, 2000). De massale belangstelling voor het suppressie-effect heeft al tot een aantal benaderingen geleid die trachten te verduidelijken hoe de gevonden tegenvoorbeeldinformatie het redeneerproces beïnvloedt (zie Byrne, Espino, & Santamaria, 1999; Johnson-Laird & Byrne, 2002; Oaksford & Chater, 1998, 2001; Politzer, in druk; Thompson, 2000). Al deze benaderingen verduidelijken echter wat er gebeurt eens een tegenvoorbeeld gevonden is. De meer fundamentele vraag hoe een tegenvoorbeeld eigenlijk teruggevonden wordt, is nog steeds onbeantwoord. De karakteristieken van het tegenvoorbeeld zoekproces zijn totaal niet gekend (Johnson-Laird & Byrne, 1994; Oaksford & Chater, 2001). Veelzeggend is hier bijvoorbeeld dat Johnson-Laird en Byrne dit 'kennisgat' zelfs als argument gebruiken voor het feit dat hun redeneertheorie het zoekproces niet specificeert:

> "..., so far, no evidence has revealed anything about the search process itself. The theory accordingly refrains from specifying the process. To refrain from speculation seemed like prudence rather than a major flaw." (Johnson-Laird & Byrne, 1994, p. 776)

De bedoeling van dit doctoraatsonderzoek is te beginnen met de opvulling van dit fundamenteel 'kennisgat' en dus de proceskarakteristieken van het tegenvoorbeeld zoekproces tijdens conditioneel redeneren te specificeren. Het ligt voor de hand dat de kern van het onderzoek een wisselwerking tussen geheugen- en redeneeronderzoek wordt. Hierbij wordt er vertrokken van het werk van de onderzoeksgroep van Markovits. Recent hebben deze mensen een eerste, voorlopige specificatie van het zoekproces voorgesteld (bv. Markovits, 2000; Markovits & Barrouillet, in druk; Markovits, Fleury, Quinn, & Venet, 1998; Quinn & Markovits, 1998). Het innoverende in het werk is dat men zich voor een ruwe schets van het zoekproces gebaseerd heeft op enkele algemene principes en assumpties uit invloedrijke geheugenmodellen (bv. Anderson, 1983, 1993; Anderson & Lebiere, 1998; Cowan, 2001). Het voorgestelde zoekmechanisme werd dan ingepast in een algemeen model voor conditioneel redeneren.

In de volgende sectie zal een algemeen overzicht gegeven worden van de studies die in de verhandeling gepresenteerd worden. De expert lezer wordt verwezen naar de Engelstalige inleiding voor een uiteenzetting van Markovits' model en een specificatie van de meer theoretische invalshoek en doelstellingen van het onderzoek.

OVERZICHT VAN DE STUDIES

Hoofdstuk 2: Opgeslagen verhinderende omstandigheden en associatiesterkte

Cummins' (1995) werk had aangetoond dat het vinden van een tegenvoorbeeld beïnvloed wordt door het aantal opgeslagen tegenvoorbeelden. Quinn en Markovits (1998) stelden echter dat naast het totale aantal elementen in een geheugenstructuur, ook de associatiesterkte van de individuele opgeslagen elementen het zoekproces moet beïnvloeden: hoe groter de associatiesterkte tussen een opgeslagen tegenvoorbeeld in een geheugenstructuur en de conditionele zin (of de elementaire representatie ervan), hoe groter de kans dat het tegenvoorbeeld zal geactiveerd worden. Alhoewel Quinn en Markovits geen formele definitie van associatiesterkte geven kunnen de elementen in een geheugenstructuur daarbij gezien worden als de knooppunten in een semantisch netwerk (Anderson, 1983). Elk knooppunt heeft een bepaalde activeringsdrempel. Een grotere associatiesterkte komt dan neer op een lagere activeringsdrempel.

Om hun hypothese te testen bepaalden Quinn en Markovits (1998) eerst de associatiesterkte van de verschillende tegenvoorbeelden van een conditionele zin. Proefpersonen moesten in 30 s zoveel mogelijk alternatieve omstandigheden opschrijven voor een bepaald effect (bv. 'een hond krabt zich'). De generatiefrequentie (% van de proefpersonen die een bepaald alternatief genereren) werd als index van associatiesterkte gebruikt. Ze zochten dan specifieke effecten met één zeer sterk geassocieerde oorzaak. Door het effect dan zowel met de sterk geassocieerde (bv. Als een hond vlooien heeft, dan krabt hij zich) als met een zwak geassocieerde oorzaak (bv. Als een hond een huidziekte heeft, dan krabt hij zich) te combineren, werden er twee versies van elk zin gemaakt.

Beide zinnen hebben dus een zelfde aantal mogelijke alternatieve omstandigheden. Echter, voor de 'sterk geassocieerde zin' zullen er enkel zwak geassocieerde elementen in de geheugenstructuur met alternatieve omstandigheden zitten. Voor de 'zwak geassocieerde zin' zal er wel een sterk geassocieerd alternatief in de geheugenstructuur aanwezig zijn. Als associatiesterkte de succesprobabiliteit van het zoekproces beïnvloedt, moet er gemakkelijker een alternatieve omstandigheid gevonden worden voor de zwak geassocieerde zin en verwacht je hier dus minder AC en DA inferenties. Quinn en Markovits (1998) vonden dit voorspelde patroon inderdaad terug.

Terwijl Cummins (1995) had aangetoond dat het aantal beschikbare tegenvoorbeelden zowel voor alternatieve als voor verhinderende omstandigheden belangrijk was, beperkten de bevindingen voor de associatiesterkte zich tot alternatieve omstandigheden. In een eerste studie (De Neys, Schaeken, & d'Ydewalle, in druk-a) werd daarom nagegaan of de associatiesterkte ook cruciaal was voor het vinden van opgeslagen verhinderende omstandigheden. In een generatietaak gaven we mensen volledige conditionele zinnen en vroegen hen om hiervoor verhinderende omstandigheden te bedenken. We selecteerden zinnen met één heel sterk geassocieerde verhinderende omstandigheid. Er werden dan redeneerproblemen geconstrueerd waarbij een premisse werd uitgebreid met de negatie van ofwel de sterk ofwel een zwak geassocieerde verhinderende omstandigheid. Er werd dus steeds een mogelijke verhinderende omstandigheid geëlimineerd. Hieronder wordt een voorbeeld van een MP probleem gegeven:

Als de airco wordt aangezet, dan krijgt Bart het koel
De airco wordt aangezet en *de airco is niet kapot* (sterk geassocieerd)
De airco wordt aangezet en *Bart heeft geen koorts* (zwak geassocieerd)

Voor beide vormen van het uitgebreide redeneerprobleem is het aantal mogelijke verhinderende omstandigheden hetzelfde (i.e., het aantal van het originele probleem min één). Echter, bij eliminatie van de sterk geassocieerd verhinderende omstandigheid zullen er enkel zwak geassocieerde elementen in de geheugenstructuur overblijven. Bij de zwak geassocieerde eliminatie zal er echter nog wel een sterk geassocieerde verhinderende omstandigheid in de geheugenstructuur aanwezig zijn. Zoals verwacht bleek dat de MP en MT inferenties minder gemaakt werden wanneer de zwak geassocieerde verhinderende omstandigheid geëlimineerd werd. Dit toonde duidelijk aan dat zowel voor alternatieve als verhinderende omstandigheden de associatiesterkte van de opgeslagen elementen de kans op het vinden van een tegenvoorbeeld beïnvloedt.

Hoofdstuk 3: Verdere tests van het model

De volgende studie (De Neys, Schaeken, & d'Ydewalle, 2002) begon met het in kaart brengen van de relatie tussen de verschillende factoren die een effect hebben op het vinden van een tegenvoorbeeld. We wisten al dat zowel het aantal opgeslagen tegenvoorbeelden als de associatiesterkte cruciaal zijn voor het zoekproces. Echter, de relatie tussen beide was nog niet duidelijk. Het is bijvoorbeeld mogelijk dat de tegenvoorbeelden van zinnen met veel mogelijke tegenvoorbeelden minder sterk geassocieerd zijn dan de tegenvoorbeelden van zinnen met weinig tegenvoorbeelden. Bijgevolg kan men er ook niet zomaar van uitgaan dat de kans dat er een tegenvoorbeeld gevonden wordt inderdaad groter is voor zinnen met veel tegenvoorbeelden. Deze assumptie is nochtans cruciaal voor de verklaring van Cummins' (1995) effecten in het model van Markovits.

Er werd daarom een set causale, conditionele zinnen geselecteerd en vervolgens werd aan proefpersonen gevraagd om hiervoor tegenvoorbeelden te genereren. Hierbij werd voor elke zin het gemiddeld aantal gegenereerde tegenvoorbeelden geregistreerd. Voor elk tegenvoorbeeld werd ook de generatiefrequentie bepaald. Bijkomend lieten we de proefpersonen de plausibiliteit van de tegenvoorbeelden evalueren.

Het experiment toonde aan dat het aantal tegenvoorbeelden, de associatiesterkte van de tegenvoorbeelden en hun plausibiliteit positief gecorreleerd waren: zinnen met veel tegenvoorbeelden hebben sterker geassocieerde en meer plausibele tegenvoorbeelden. Dit bevestigt de assumptie dat het waarschijnlijker is dat er een tegenvoorbeeld gevonden wordt in een geheugenstructuur met veel opgeslagen elementen.

De effecten van het aantal opgeslagen tegenvoorbeelden op de mate waarin conditionele inferenties aanvaard worden zijn goed gedocumenteerd. Er zijn echter geen studies die naar het effect op de inferentietijd gekeken hebben. Ons vorige experiment toonde aan dat er voor zinnen met veel tegenvoorbeelden typisch een viertal (sterk geassocieerde) tegenvoorbeelden opgeslagen zijn. Nu, Conway en Engle (1994) toonden aan dat (tot vier opgeslagen elementen) een semantisch zoekproces meer tijd vraagt wanneer het aantal elementen dat gevonden wordt toeneemt. Als we er van uitgaan dat er voor de zinnen met veel tegenvoorbeelden ook meerdere tegenvoorbeelden gezocht en gevonden worden, dan betekent dit dat het zoekproces bij de 'veel' zinnen langer zal duren dan bij de 'weinig' zinnen. Dit langere zoekproces zou tot langere inferentietijden moeten leiden. In een tweede experiment repliceerden we daarom het experiment van Cummins (1995), maar er werd nu ook naar de reactietijd voor de verschillende inferenties gekeken.

Er werd inderdaad geobserveerd dat de AC inferentie langer duurde voor zinnen waarvoor er veel alternatieve omstandigheden beschikbaar zijn dan voor zinnen met slechts weinig alternatieve omstandigheden. De MP inferentie vertoonde een gelijkaardig effect in functie van het aantal opgeslagen verhinderende omstandigheden. Deze bevindingen zijn dus consistent met de assumptie dat het zoekproces langer duurt wanneer er veel tegenvoorbeelden zijn opgeslagen.

De rol van het zoekproces werd verder bestudeerd door in een derde experiment naar het effect van interindividuele verschillen in de efficiëntie van het zoekproces te kijken. Als de uitkomst van het zoekproces inderdaad bepaalt welke conclusie mensen trekken, dan moeten individuele verschillen in de efficiëntie van het zoekproces de redeneerprestatie beïnvloeden. In het experiment werd bij een aantal mensen de capaciteit om verhinderende omstandigheden te vinden gemeten. Na een inferentietaak moesten de proefpersonen voor een reeks zinnen binnen 30 s steeds zoveel mogelijk verhinderende omstandigheden zoeken. Het totaal aantal gevonden tegenvoorbeelden werd dan als maat van de zoekcapaciteit gebruikt. Uit de resultaten bleek dat hoe groter iemands zoekcapaciteit was, hoe minder de MP en MT inferenties gemaakt werden. Deze bevinding illustreert hoe belangrijk de karakteristieken van het zoekproces zijn voor een conditionele redeneertheorie: interindividuele variatie in de efficiëntie van het zoekproces is bepalend voor de soort conclusie die mensen trekken.

Hoofdstuk 4: Elk tegenvoorbeeld telt?!

Een volgende studie (De Neys, Schaeken, & d'Ydewalle, in druk-b) spitste zich toe op de vraag of het zoekproces al dan niet stopt nadat er één enkel tegenvoorbeeld gevonden is. In principe is het inderdaad mogelijk dat het zoeken beëindigd wordt nadat er één tegenvoorbeeld teruggevonden is en dat eventuele resterende opgeslagen tegenvoorbeelden geen effect zullen hebben op de inferenties. In de mentale model theorie (MMT, Johnson-Laird, 1983) en Markovits' eerste (op de MMT geïnspireerde) zoekprocesspecificatie wordt deze assumptie bijvoorbeeld impliciet gemaakt. In onze studie werd de assumptie expliciet getest. De alternatieve specificatie die wij voorstelden is dat het zoekproces niet stopt nadat er één tegenvoorbeeld is gevonden, maar dat er verder zal gezocht worden naar bijkomende tegenvoorbeelden. Het totaal aantal gevonden tegenvoorbeelden zal dan bepalen in welke *mate* een individu een conclusie aanvaardt. Er werd al enige evidentie voor dit 'bijkomend tegenvoorbeelden zoeken' gevonden in de reactietijdeffecten in de vorige studie (De Neys et al., 2002; Hoofdstuk 3).

In een eerste experiment bestudeerden we het effect van bijkomende tegenvoorbeelden op de inferenties door zelf expliciet mogelijke tegenvoorbeelden aan te bieden. Zoals in de traditionele suppressie studies (Byrne, 1989; Byrne et al., 1999; Rumain et al., 1983) simuleerden we met de expliciete aanbieding dus eigenlijk het succesvol terugvinden van een tegenvoorbeeld. De cruciale manipulatie bestond erin dat we het aantal aangeboden tegenvoorbeelden systematisch veranderden. Elke proefpersoon kreeg vijf verschillende conditionele zinnen aangeboden waarbij het aantal aangeboden tegenvoorbeelden varieerde van nul tot vier.

In een tweede experiment werd het zoekproces onderzocht zonder gebruik te maken van een expliciete aanbieding. Er werd in het experiment een voormeting gebruikt. We gingen na hoeveel tegenvoorbeelden elke proefpersoon voor elke conditionele zin in onze basisset kon vinden. Een maand na de pretest kwamen de proefpersonen terug voor het eigenlijke redeneerexperiment. Hiervoor werden dezelfde proefpersonen en dezelfde conditionele zinnen als in de pretest gebruikt. We wisten dus precies hoeveel tegenvoorbeelden er bij elke proefpersoon voor elke conditionele zin opgeslagen waren. Voor elke zin werd er dan gekeken naar de mate (op een 7-punten schaal) waarin een proefpersoon vond dat de verschillende conclusies konden aanvaard worden in functie van het aantal tegenvoorbeelden dat die proefpersoon voor die zin kon bedenken.

Als je kijkt naar de mate waarin een conclusie aanvaard wordt in functie van het aantal beschikbare tegenvoorbeelden dan verwacht de MMT en Markovits een trapsfunctie: tot een zeker aantal opgeslagen tegenvoorbeelden (stel nul of één) zal het zoeken niets opleveren en wordt de conclusie aanvaard. Nadat één enkel tegenvoorbeeld succesvol gevonden is (stel bij twee opgeslagen tegenvoorbeelden) wordt de conclusie verworpen, het zoekproces stopt en bijkomende beschikbare tegenvoorbeelden kunnen geen verdere impact hebben. Onze specificatie verwacht dat er een graduele daling met elk bijkomend beschikbaar tegenvoorbeeld zal te zien zijn. De resultaten steunden onze predictie. De mate waarin AC aanvaard werd, daalde lineair met elke bijkomend beschikbare alternatieve omstandigheid en de MP aanvaarding vertoonde een gelijkaardige negatieve lineaire trend in functie van het aantal beschikbare verhinderende omstandigheden.

Hoofdstuk 5: Werkgeheugen en het terugvinden en inhiberen van opgeslagen tegenvoorbeelden

In een volgende studie (De Neys, Schaeken, & d'Ydewalle, 2003) werd het tegenvoorbeeld zoekproces verder gekarakteriseerd door te kijken naar de rol van het werkgeheugen bij het zoeken.

Het werkgeheugen (WG) wordt vaak gekarakteriseerd als een hiërarchisch georganiseerd systeem waarbij ondergeschikte opslagcomponenten een centrale component ondersteunen die de informatieverwerking controleert (b.v., Baddeley & Hitch, 1974; Engle & Oransky, 1999). De controlecomponent of 'centrale executieve' regelt hoe onze beperkte cognitieve aandachtscapaciteit besteed wordt.

Geheugenonderzoek heeft aangetoond dat sommige geheugenzoekprocessen eerder automatisch en zonder bewuste inspanning verlopen, terwijl voor andere zoekprocessen de executieve controle van het WG nodig is (Kane & Engle, 2000; Moscovitch, 1994, 1995; Rosen & Engle, 1997). Moscovitch (1995) noemde deze twee processen respectievelijk het associatief en strategisch zoeken. Het cruciale kenmerk van het strategisch zoeken is dat het een beroep doet op de executieve WG-capaciteit. Rosen en Engle argumenteerden dat het zoeken in het lange-termijn geheugen steeds start met een associatieve, automatische activatieverspreiding. Bij een strategisch zoekproces zou er dan vervolgens een beroep worden gedaan op het WG voor een actieve "cue" generatie die het mogelijk maakt om minder gemakkelijk terug te vinden opgeslagen elementen te activeren. De actieve "cue" generatie laat daarmee een veel efficiënter zoeken toe dan de passieve activatieverspreiding.

Als het zoeken van opgeslagen tegenvoorbeelden strategisch van aard is en het werkgeheugen dus cruciaal is bij het zoeken, dan verwacht je dat de efficiëntie van het zoekproces zal afhangen van de beschikbare WG-capaciteit. Deze voorspelling werd getest in een eerste experiment. Daarvoor ontwikkelden we eerst een Nederlandstalige en collectief afneembare aanpassing van de OSPAN-test (La Pointe & Engle, 1990) zodat we de WG-capaciteit van de proefpersonen konden meten (zie De Neys, d'Ydewalle, Schaeken, & Vos, 2002). De proefpersonen kregen allemaal de WG-test en een tegenvoorbeeld generatietaak. De resultaten toonden dat een hogere WG-capaciteit inderdaad samenhing met een efficiënter terugvinden van de tegenvoorbeelden.

In een tweede experiment werd dubbeltaak methodologie gebruikt om de causale aard van de bevindingen te toetsen. Als WG-capaciteit inderdaad een rol speelt bij het vinden van de tegenvoorbeelden, dan verwacht je dat het belasten van het werkgeheugen met een secundaire taak een negatieve impact zal hebben op de efficiëntie van het zoekproces. Een puur automatisch zoekproces zal niet beïnvloed worden door de WG-belasting.

Als secundaire taak werd de proefpersonen gevraagd om een vingerpatroon te tokkelen met hun niet-dominante hand (zie Kane & Engle, 2000; Moscovitch, 1994). Vroegere studies toonden aan dat het tokkelen van een complex, nieuw patroon (bv. wijsvinger- ringvinger- middenvinger- pink) de centrale executieve belastte, terwijl het tokkelen van een eenvoudig en vertrouwd patroon (bv. pink- ringvinger- middenvinger- wijsvinger) helemaal niet belastend was voor het werkgeheugen. Eén groep proefpersonen tokkelde het complexe patroon en een andere groep het eenvoudige patroon tijdens de generatietaak.

De resultaten toonden aan dat de zoekefficiëntie inderdaad daalde wanneer het werkgeheugen belast was met de aandachtsvragende complexe tokkeltaak, terwijl tokkelen van het niet-belastende eenvoudige patroon geen impact had op het aantal gegenereerde tegenvoorbeelden.

In een derde experiment vergeleken we de prestatie van een groep laag en hoog spans (proefpersonen uit het bodem en top kwartiel van de WG-capaciteitsdistributie van eerste kandidatuur studenten psychologie) in een eigenlijke redeneertaak. Het eerste experiment van de studie toonde aan dat mensen met een hogere WG-capaciteit meer succesvol zijn in het terugvinden van tegenvoorbeelden. Ons vorig onderzoek toonde aan dat hoe succesvoller het terugvinden van de alternatieve omstandigheden is, hoe minder de AC en DA inferenties aanvaard worden. Daarom verwacht je dat hoog spans (vs. laag spans) minder geneigd zullen zijn om de AC en DA inferenties te aanvaarden. Aangezien het terugvinden van verhinderende omstandigheden leidt tot het verwerpen van MP en MT, kan men ook verwachten dat vanwege het meer efficiënt zoekproces, hoog spans minder geneigd zullen zijn om de MP en MT inferenties te aanvaarden. Men moet hier echter in het achterhoofd houden dat terwijl AC en DA logisch gezien foutief zijn, de MP en MT inferenties valide zijn. AC en DA verwerpen is logisch gezien correct, MP en MT verwerpen niet.

Er wordt verondersteld dat alle mensen een “contextualisatie” neiging vertonen om tegenvoorbeelden te zoeken die geassocieerd zijn met het redeneerprobleem. Echter, studies naar interindividuele verschillen tonen aan dat ten minste de mensen met de hoogste cognitieve capaciteiten (bv. hoog spans) ook een elementaire, logische “decontextualisatie” bekwaamheid hebben: een elementaire bekwaamheid om achtergrondkennis opzij te schuiven wanneer deze conflicteert met de logische standaarden (vb. Klaczynski, 2001a; Stanovich & West, 2000). Als hoog spans inderdaad een elementaire notie van logische validiteit hebben, dan zou deze moeten conflicteren met de neiging om verhinderende omstandigheden te zoeken. Op grond van deze assumptie verwachten we dat hoog spans hun WG-capaciteit zullen aanwenden voor een actieve inhibitie van het spontaan zoeken naar opgeslagen verhinderende omstandigheden. Het inhiberen van responsen of cognitieve processen die als ongewenst worden gezien, is inderdaad één van de algemeen erkende basisfuncties van de centrale executieve (vb. Baddeley; Engle, Tuholski, Laughlin, & Conway, 1999; Shallice & Burgess, 1993; Miyake & Shah, 1999). Wanneer het dus aankomt op het zoeken van verhinderende omstandigheden zouden de hoog spans hun WG-capaciteit niet aanwenden voor een actief, strategisch zoeken, maar juist voor een inhibitie van het automatisch zoekproces. Ondanks hoog spans' betere intrinsieke zoekefficiëntie zou de inhibitie dus toch moeten resulteren in hogere MP en MT aanvaarding voor de hoog (vs. laag) spans. Uit de resultaten bleek inderdaad dat AC en DA aanvaarding het hoogst was voor de laag spans, terwijl MP en MT meer aanvaard werden door de hoog spans.

Alhoewel de resultaten van het experiment de voorspellingen ondersteunden bleef de evidentie puur correlationeel en vatbaar voor alternatieve verklaringen. In een volgend experiment voerden we daarom een bijkomende, meer directe test van de hypothese uit. We gingen het effect na van een secundaire WG-belasting (de complexe tokkeltaak uit Experiment 2) op de redeneerprestatie. Als de WG-capaciteit inderdaad gebruikt wordt voor het zoeken en inhiberen van tegenvoorbeelden, dan moet een belasting van het werkgeheugen het goed functioneren van deze processen beletten. De data vertoonde inderdaad het verwachte interferentiepatroon.

Hoofdstuk 6: Werkgeheugen en tegenvoorbeelden zoeken: Een trendanalyse

In de laatste studie (De Neys, Schaeken, Dieussaert, d'Ydewalle, 2003) werd de WG-afhankelijke mediatie van het tegenvoorbeeld zoeken en inhiberen tijdens het redeneren verder bestudeerd. In de vorige studie gingen we er expliciet van uit dat het inhiberen van het zoeken naar verhinderende omstandigheden enkel zou optreden bij mensen met de grootste WG-capaciteit. Er werden echter enkel mensen uit het bodem en top kwartiel van de WG-capaciteitsdistributie vergeleken. Als de assumptie dat de inhibitie enkel optreedt bij de hoog groep correct is, dan moet je verwachten dat mensen met gemiddelde WG-capaciteit (medium spans) juist de laagste mate van MP en MT aanvaarding zullen vertonen. Aan de ene kant verwacht je dat medium spans (vs. hoog spans) het verhinderende omstandigheden zoeken niet zullen inhiberen. Aan de ander kant moet het zo zijn dat medium spans beter verhinderende omstandigheden kunnen vinden dan laag spans omdat ze juist meer WG-capaciteit ter beschikking hebben. Het terugvinden van de verhinderende omstandigheden moet dus het meest succesvol zijn bij de medium spans. Daarom verwachten we eigenlijk dat de MP en MT aanvaarding in functie van de WG-capaciteit een U-vormig verband zal vertonen: door de beperkte capaciteit zal het tegenvoorbeeld zoekproces niet erg succesvol zijn bij de laag spans waardoor zij een vrij hoge mate van MP en MT aanvaarding zullen vertonen. Omwille van het meer efficiënte zoekproces zal de MP en MT aanvaarding dalen bij de medium spans. Omwille van de verhinderende omstandigheden inhibitie moet de MP en MT aanvaarding weer gaan stijgen voor de mensen met de hoogste WG-capaciteit.

Aangezien het terugvinden van alternatieve omstandigheden resulteert in het verwerpen van AC en DA en deze inferenties logisch gezien foutief zijn, is er geen conflict tussen een elementaire logica notie en het zoeken naar alternatieve omstandigheden. Het zoeken naar alternatieve omstandigheden wordt dan ook niet verondersteld geïnhibeerd te worden door de hoog spans. Daarom zal met een toename van de WG-capaciteit het zoeken van alternatieve omstandigheden steeds efficiënter worden en de aanvaarding van de AC en DA inferenties steeds dalen. In tegenstelling tot MP en MT moet de AC en DA aanvaarding daarom een negatief lineaire trend vertonen in functie van WG-capaciteit. Deze voorspelling werd getoetst in een eerste experiment. Driehonderd proefpersonen kregen een redeneertaak met alledaagse conditionele zinnen en een werkgeheugentest. De trend predicties werden bevestigd.

In een tweede en derde experiment werden de bevindingen verder gegeneraliseerd door aan te tonen dat de kwadratische, U-vormige MP trend zelfs kon gerepliceerd worden wanneer een verhinderende omstandigheid expliciet werd aangeboden.

CONCLUSIE

De studies in deze verhandeling specificeerden cruciale proceskarakteristieken van het tegenvoorbeeld zoekproces tijdens het redeneren met dagdagelijks conditionele zinnen. Er kon aangetoond worden dat de uitkomst van het zoekproces afhankelijk is van de associatiesterkte van de opgeslagen tegenvoorbeelden. We specificeerden het tijdsverloop van het zoekproces en kenmerkten hoe het zoekproces stopt. We toonden aan dat het zoekproces niet volledig automatisch verloopt, maar een beroep doet op het werkgeheugen. Tenslotte verduidelijkten we dat het zoekproces ook spontaan geïnhibeerd kan worden wanneer de aard van een tegenvoorbeeld conflicteert met de logische validiteit van een redeneerprobleem. In het afsluitend hoofdstuk wordt op basis van deze gespecificeerde karakteristieken een algemeen model van het tegenvoorbeeld zoekproces geschetst.

References

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

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

Anderson, J. R., & Lebiere, C. (1998). *Atomic components of thought*. Hillsdale, NJ: Erlbaum.

Anderson, M. C., & Bell, T. (2001). Forgetting our facts: The role of inhibitory processes in the loss of prepositional 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, and Cognition, 22*, 1408-1422.

Barrouillet, P., Grosset, N., & Lecas, J. F. (2000). Conditional reasoning by mental models: chronometric and developmental evidence. *Cognition, 75*, 237-266.

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

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

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

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

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

Chan, D., & Chua, F. (1994). Suppression of valid inferences: Syntactic views, mental models, and relative salience. *Cognition, 53,* 217-238.

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

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

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

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

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

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

De Neys, W., Schaeken, W., Dieussaert, K., & d'Ydewalle. (2003). *Working memory span and disabler retrieval inhibition: A trend analysis.* Manuscript submitted for publication.

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

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

De Neys, W., Schaeken, W., & d'Ydewalle, G. (2000). *Causal conditional reasoning, semantic memory retrieval, and mental models: A test of the 'semantic memory framework'.* (Psychological report No.270). Leuven: University of Leuven. Laboratory of Experimental 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,* 908-920.

De Neys, W., Schaeken, W., & d'Ydewalle, G. (2003). *Working memory and everyday conditional reasoning: Retrieval and inhibition of stored counterexamples.* (Tech. Rep. No. 295).. Leuven, Belgium: University of Leuven, Laboratory of Experimental Psychology.

Dieussaert, K., Schaeken, W., & d'Ydewalle, G. (in press). The impact of the nature of disabling conditions on the reasoning process. *Current Psychology Letters.*

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.

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

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

Elio, R. (1997). What to believe when inferences are contradicted: the impact of knowledge type and inference rule. *Proceedings of the Annual Conference of the Cognitive Science Society, 19,* 211-216.

Elio, R. (1998). How to disbelieve p -> q: Resolving contradictions. *Proceedings of the Annual Conference of the Cognitive Science Society, 20,* 315-320.

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

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

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

Evans, J. St. B. T. (2000). What could and could not be a strategy in reasoning. In W. Schaeken, G. De Vooght, A. Vandierendonck, & G. d'Ydewalle (Eds.). *Deductive reasoning and strategies.* Mahwah, NJ: Erlbaum.

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

Evans, J. St. B. T., Handley, S. J., & Over, D. E. (in press). Conditionals and conditional probability. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

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

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

Galotti, K. M. (1989). Approaches to studying formal and everyday reasoning. *Psychological Bulletin, 105,* 331-351.

George, C. (1995). The endorsement of the premises: Assumption-based or belief-based reasoning. *British Journal of Psychology, 86,* 93-111.

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

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

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

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, MA: MIT Press.

Henle, M. (1962). On the relation between logic and thinking. *Psychological Review, 69,* 366-378.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986*). Induction: Processes of inference, learning, and discovery*. Cambridge, MA: MIT Press.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking*. New York: Basic Books.

Jacobs, J. E., & Klaczynski, P. A. (2002). The development of judgment and decision making during childhood and adolescence. *Current Directions in Psychological Science, 11,* 145-149.

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

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

Johnson-Laird, P. N., & Byrne, R. M. J. (1993). Précis of deduction. *Behavioral and Brain Sciences, 16,* 323-380.

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.

Johnson-Laird, P. N., Byrne, R. M. J., & Schaeken, W. (1992). Propositional reasoning by model. *Psychological Review, 99*, 418-439.

Johnson-Laird, P. N., Byrne, R. M. J., & Schaeken, W. (1994). Why models rather than rules give a better account of propositional reasoning: A reply to Bonatti and to O'Brien, Braine, and Yang. *Psychological-Review, 101*, 734-739.

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

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.

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

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgement under uncertainty: Heuristics and biases*. Cambridge, MA: Cambridge University Press.

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. (2001a). Analytic and heuristic processing influences on adolescent reasoning and decision making. *Child Development, 72*, 844-861.

Klaczynski, P. A. (2001b). Framing effects on adolescent task representation, analytic and heuristic processing, and decision making: Implications for the normative/descriptive gap. *Applied Developmental Psychology, 22*, 289-309.

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. *Trends in Cognitive Sciences, 6*, 299-305.

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

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review, 95*, 492-527.

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

Manktelow, K. I., & Fairley, N. (2000). Superordinate principles in reasoning with causal and deontic conditionals. *Thinking and Reasoning, 6*, 41-65.

Markovits, H. (1984). Awareness of the 'possible' as a mediator of formal thinking in conditional reasoning problems. *British Journal of Psychology, 75*, 367-376.

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

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

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

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

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

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

Markovits, H., & Potvin, F. (2001). Suppression of valid inferences and knowledge structures: The curious effect of producing alternative antecedents on reasoning with causal conditionals. *Memory & Cognition, 29*, 736-744.

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

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

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., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "Frontal Lobe" tasks: A latent variable analysis. *Cognitive Psychology, 41*, 49-100.

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.

Newstead, S. E., Ellis, C. E., Evans, , J. St. B. T., & Dennis, I. (1997). Conditional reasoning with realistic material. *Thinking and Reasoning, 3*, 49-96.

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

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

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

Oaksford, M., Morris, F., Grainger, B., & Williams, J. M. G. (1996). Mood, reasoning, and central executive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 476-492.

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

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.

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

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

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

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.

Schaeken, W., De Vooght, G., Vandierendonck, A., & d'Ydewalle, G. (Eds.). (2000). *Deductive reasoning and strategies*. Mahwah, NJ: Erlbaum.

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

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 an 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.

Smedslund, J. (1970). Circular relation between understanding and logic. *Scandinavian Journal of Psychology, 11*, 217-219.

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, NJ: Erlbaum.

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

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

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

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

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

Tulving, E. (1983). *Elements of episodic memory*. Oxford: Clarendon Press.

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

Van Lambalgen, M., & Stenning, K. (2002). *'Suppression effect' and nonmonotonicity*. Unpublished manuscript.

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.

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

Watterson, B. (1990). *Calvin and Hobbes: Weirdos from another planet*. Kansas City, Missouri: Andrews & McMeel.

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