2011

Necessity, Possibility and the Search for Counterexamples in Human Reasoning

Serpell, Sylvia Mary Parnell

http://hdl.handle.net/10026.1/560

University of Plymouth

Necessity, Possibility and the Search for

Counterexamples in Human Reasoning

Sylvia Mary Parnell Serpell

School of Psychology
University of Plymouth

A thesis submitted to the University of Plymouth in
partial fulfilment of the requirements for the degree of
Doctor of Philosophy

6th December 2011

Abstract

Necessity, Possibility and the Search for Counterexamples in Human Reasoning
Sylvia Mary Parnell Serpell

This thesis presents a series of experiments where endorsement rates, latencies and measures of cognitive ability were collected, to investigate the extent to which people search for counterexamples under necessity instructions, and alternative models under possibility instructions. The research was motivated by a syllogistic reasoning study carried out by Evans, Handley, Harper, and Johnson-Laird (1999), and predictions were derived from mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991).

With regard to the endorsement rate data: Experiment 1, using syllogisms, found no evidence to suggest that a search for counterexamples or alternative models took place. In contrast experiment 2 (transitive inference) found some evidence to support the search for alternative models under possibility instructions, and following an improved training session, experiment 3 (transitive inference) produced strong evidence to suggest that people searched for other models; which was mediated by cognitive ability.

There was also strong evidence from experiments 4, 5 and 6 (abstract and everyday conditionals) to support the search for counterexamples and alternative models. Furthermore it was also found that people were more likely to find alternative causes when there were many that could be retrieved from their everyday knowledge, and that people carried out a search for counterexamples with many alternative causes under necessity instructions, and across few and many causal groups under possibility instructions. The evidence from the latency data was limited and inconsistent, although people with higher cognitive ability were generally quicker in completing the tasks.

Contents

List of figures

List of tables

List of appendices

ACKNOWLEDGMENTS

First, I would like to acknowledge the support of my supervisory team. My Director of Studies, Professor Simon Handley, has been supportive and patient throughout this period of study, as well as offering invaluable help and guidance. I also thank my second Supervisor, Emeritus Professor Steve Newstead, for affording me this opportunity, and for always having words of wisdom when most needed. My thanks also go to Dr Alison Bacon, who inspired and introduced me to deductive reasoning when, as an undergraduate, I participated in her study on syllogistic reasoning.

Second, I thank my husband Charlie, for his continued love and support; my daughters Caroline and Vicki for their patience and understanding; my mother, Mary Parnell for her unconditional love throughout this challenging process; and my aunt, Ruth Parnell for being that special person everyone needs in their life. My gratitude also goes to all of my friends for their emotional support, and for believing in me when I didn't believe in myself.

Last, but not least I would like to acknowledge two special men in my life; my dad, Francis Williams Parnell who I loved very much and who is always with me in spirit; and my dearest uncle, Ernest Claude Parnell who sadly did not see me reach the end of my studies.

Author's declaration

At no time during the registration for the degree of Doctor of Philosophy has the author been registered for any other University award without prior agreement of the Graduate Committee.

This programme of research was financed with the aid of a fees paid studentship from the School of Psychology at the University of Plymouth.   The thesis is the result of the author's own investigations, carried out under the guidance of her supervisory team. Other sources are explicitly referenced.

During this period of study, the author has also completed the MSc Certificate in Psychology Research Methods, and achieved the following conference contributions:

Refereed conference contributions:

Serpell, S. M .P., Handley, S. J., Newstead, S. E., (2009). *Do reasoners carry out a search for counterexamples?* Paper presented at the PsyPAG Conference, Cardiff, 29 – 31 July, 2009.

Serpell, S. M. P., Handley, S . J., Newstead, S. E., (2006). *Differences between formal and everyday reasoning.*   Paper presented at the Psychology School Conference, University of Plymouth, 26 April, 2007.

Serpell, S. M. P., Newstead, S. E., Handley, S. J., (2006). *Necessary and possible inference: The role of counter-example search.* Poster presentation at the Experimental Psychology Society Meeting, Plymouth, 10 – 11 July, 2006.

Word count for the main body of this thesis: 55,882

Signed

6th December 2011

Chapter 1

Introduction to Human Reasoning

It has been said that reasoning is fundamental to human intelligence (Evans, Newstead, & Byrne, 1993), and as such is central to science, society, and the solution of all practical problems. However, despite a long history of reasoning research, there are still a number of unanswered theoretical and practical questions about the cognitive processes involved. The aim of this thesis is to add to our knowledge, by exploring the way in which people consider *possibilities* when seeking to find a solution to a problem, which are based upon the knowledge and information available to them.

The majority of studies in the large body of psychological reasoning research generated since the early 1900's, have adopted deductive logic as a standard against which to explore these processes. This dates back to the work of 19th century philosophers Boole (1854) and Mill (1843), who argued that the laws of logic are the laws of thought. The experimental programme of work reported in this thesis uses three deductive reasoning paradigms: syllogistic reasoning, transitive inference, and conditional inference, which are applied to a range of problem types and content (abstract and

everyday) to facilitate the collection of responses and response latencies. A cognitive ability[1] test was administered in all of the experiments, to gain a more sensitive measure of the relationship between general intelligence and deductive competence. This is primarily because past research has found reliable correlations between reasoning performance and cognitive ability (Evans et al., 1983; Newstead et al., 2004, Stanovich and West, 1998a; Torrens et al., 1999).

The approach that has been taken is novel, as previous research studies have tended to focus on only one reasoning paradigm, with abstract *or* everyday content, and few have added a measure of ability. This has therefore not allowed comparisons to be made across paradigms or content types, within a study. Another way in which this research is different from all but a few of the studies reported to date (i.e. Evans, Handley, Harper, & Johnson-Laird, 1999), is that not only were participants asked to judge whether conclusions were *necessary* given some premises, but participants were also asked to judge whether conclusions were *possible* in the light of the given information. The differences between these two types of instruction will be discussed later in this chapter; but the first part is concerned with standard logical concepts of necessity, where a deductive inference is valid if its conclusion must be true, given that its premises are true.

The main rationale and design of the experimental work is motivated by the mental model theory of deductive reasoning (Johnson-Laird, 1983; Johnson-Laird & Byrne,

---

[1] Cognitive ability is defined by Sternberg and Salter (1982) as *goal directed adaptive behaviour;* and *the ability to deal with cognitive complexity* (Gottfredson, 1997).

1991), which supposes that people reason deductively by constructing and manipulating internal representations (mental models) of the information available. Research has shown (i.e. Johnson-Laird & Bara, 1984) that those tasks requiring the manipulation of multiple models prove more difficult to participants than tasks with only one model, reflecting higher error rates and longer latencies. Initially, this chapter will introduce the main deductive reasoning paradigms and findings in the literature, which relate specifically to the experiments presented in the thesis. This will be followed by a review of the most significant general theories of deductive reasoning, which have attempted to explain reasoning processes; with reference to reasoning paradigms where appropriate.

The list of deductive reasoning paradigms and theories is exhaustive, but in the main serves to provide a balanced account of reasoning across domains, in support of the experimental program presented herein[2].

Deductive reasoning

Deduction is a process of thought whereby people start with information such as perceptual observations, memories, statements, beliefs or imagined states of affair, to arrive at a novel conclusion that follows from the given information. In other words, the conclusion is not wholly explicit in the premises, but can be deduced from the content of the preceding statements.

---

[2] A comprehensive review of both theories and paradigms can be found in more general texts such as (Evans et al., 1993), together with specific texts such as Rips (1994) and Johnson-Laird (1983).

Reasoning paradigms and tasks have changed little since the early days of reasoning research, in that individuals are given deductive premises, to which they are required to generate a response, evaluate a given conclusion, or select a response from a number of possible ones made available. Generally past research has been organized around three main questions which look at either; the competency of untrained reasoners in deduction tasks, the kinds of systematic biases influencing their inferences, and the extent to which responses influenced by content and context. One of the most frequently used paradigms, which has generated a large body of published research, is syllogistic reasoning.

Syllogistic reasoning

Syllogisms of the type first devised by the Greek philosopher Aristotle (384 BC – 322 BC) as a tool for teaching logic, are deductive arguments consisting of two premises followed by a conclusion. The premises are made up of 3 terms, A, B and C, which are defined by one of four quantifiers all, none, some, or some …. not; the first premise links terms A and B, the second premise links terms B and C, and the middle term (B) is common to both premises. The conclusion links terms A and C, and content varies depending on the requirements of the research study but is generally abstract, thematic, or content which depends on people's everyday knowledge of the world (see examples below and on the next page).

Abstract content: Some of the A's are B's
                     None of the B's are C's
                     Therefore: Some of the A's are not C's

Thematic content: All of the Actors are Beekeepers

                      All of the Beekeepers are Chemists

                      Therefore: All of the Actors are Chemists

Everyday content: Some police dogs are vicious

                      Some highly trained dogs are vicious

                      Therefore: Some highly trained dogs are not police dogs

There are four possible arrangements of these premise terms, which are traditionally referred to as figures, and although the arrangement varies depending upon the source, this thesis will follow the convention of Johnson-Laird (1983) by describing the four figures of premise arrangements as shown in table 1.1.

As there are four figures, and the premises and conclusion each contain one of four different quantifiers, there are a total of 256 syllogisms; which when extended to include conclusions in the form A - C *and* C - A, provide a total of 512 possible combinations.

Table 1.1
The four Syllogistic Figures (Johnson-Laird, 1983)

Figure 1	Figure 2	Figure 3	Figure 4
A - B	B - A	A - B	B - A
B - C	C - B	C - B	B - C

Furthermore, each of the premises and the conclusion of a syllogism is described by four quantifier moods, which are referred to by the terms A (all), E (no/none), I (some) or O (some …. not); thus, a syllogism containing the quantifier *no* in the first premise, *all*

in the second premise, and *no* in the conclusion, is described as mood EAE. The quantifiers *all* and *no* are *universal* in that they encompass all members of a group, and the quantifiers *some* and *some .... not* are *particular* because they encompass specific members of a group (see table 1.2).

Table 1.2
Mood term, together with quantifier and description

Mood term	Quantifier	Description
A	All	Universal affirmative
E	No	Universal negative
I	Some	Particular affirmative
O	Some .... not	Particular negative

The logical validity of a syllogism is determined by the mood and the figure, but only 27 out of a possible 256 syllogisms (or 512 if the order of the major and minor premises is changed) yield logically valid conclusions. To illustrate this, below is an example of a valid syllogism, which combines the structure of figure 3 with an A - C conclusion, in the mood OAO:

> Some of the A's are not B's
> All of the C's are B's
> Therefore: Some of the A's are not C's

However, when the conclusion is changed to mood I, the problem becomes invalid, because although it can be deduced from the premises that some of the A's are not C's, the premises do not imply that some of the A's are C's (see next page).

> Some of the A's are not B's
> All of the C's are B's
> Therefore: Some of the A's are C's

A number of syllogistic reasoning behaviours and effects have been consistently observed and reported in the literature, which fall broadly into two categories; response generation effects such as the atmosphere effect, figural effect, and matching theory; and linguistic explanations such as conversion theory and conversational implicature. These are discussed below; although given the long history of syllogistic reasoning research that has been carried out, this list is not exhaustive.

Response generation effects

Probably the earliest reported heuristic explanation is the *Atmosphere effect* (Begg & Denny, 1969; Woodworth & Sells, 1935), which suggests that reasoners are influenced by the 'atmosphere' created by the mood (A, E, I or O) of the premises, so:

> When presented with at least one negative premise, *no* (E) or *some .... not* (O) a negative atmosphere is created, and participants are inclined to select a negative conclusion

> When one or more of the premises is particular, *some* (I) or *some .... not* (O) then the preferred conclusion is particular

> If neither of the above is present then an affirmative universal *all* (A) conclusion is chosen

8

For example:

> Some of the A's are B's
> Some of the B's are C's

Given that both premises are *positive* and *particular*, the theory predicts a strong preference for a *particular* conclusion, despite it being an invalid problem:

> Some of the A's are C's

Although the atmosphere effect seems to account for up to 90% of responses (Johnson-Laird & Bara, 1984), the theory only describes patterns of performance, does not allow for differences in difficulty between problems (Evans et al., 1993), and fails to explain why some participants produce a 'no valid conclusion' response.

Matching theory, another heuristic effect that has been found in studies of syllogistic reasoning (Wetherick, 1989; Wetherick & Gilhooly, 1990), is based on the notion that when the logic of a problem is not immediately apparent, reasoners choose conclusions where the quantifier is the same as one of those used in the premises. So, where there is a choice, a preference is shown for the more conservative one[3], or in other words where the quantifier commits the speaker to the smallest possible number of *positive* instances, E > I = O > > A, for instance:

> Some of the A's are B's
> All of the B's are C's

---

[3] The forms are ordered for conservatism from most to least: *no*, *some*, *some .... not*, and *all*; with *some* and *some .... not* being equal.

According to the rule, participants tend to select the conclusion *some of the A's are C's*, because this uses the more conservative of the two quantifiers *some* (I) and *All* (A); but if both premises contain the quantifier a*ll* (A) then the conclusion a*ll* (A) is chosen. Generally, matching theory makes similar predictions to that of the atmosphere effect, but differ with the premise pairs, IE and OE, where the atmosphere effects predicts an O conclusion, while the matching effect predicts an E conclusion. However inspection of data collected by Dickstein (1978) reveals that the *atmosphere* of a syllogism correctly predicts the response to the premises pairs 29% of the time, whereas *matching theory* makes the correct prediction 17% of the time.

Other predictions that can be made from the theory are, when one of the premises contains *some* and the other *some …. not*, the preferred response should be *some*. However this was found not to be the case by Evans et al. (1993); and Johnson-Laird and Byrne (1989) found that participants rarely preferred conclusions containing *only*, even when the both premises used the quantifier referred to. It would therefore appear that there is little to choose between the earlier findings of the *atmosphere effect* and the more recent *matching effect*.

The *figural effect*, which is another of the earlier effects reported in the literature, suggests that the figure of a syllogism influences both accuracy and directionality preferences. Studies using conclusion production tasks (i.e. Johnson-Laird & Bara, 1984) and more recently Stupple and Ball (2007) have shown that A - B, B - C problems yield more correct responses than B - A, C - B problems. Also, directional bias on conclusion production tasks tend to show a preference for A - C conclusions when

presented with terms in the order of A - B, B - C, and C - A conclusions when presented with terms in B - A, C - B order. Notwithstanding this, other studies (Evans, Handley, & Harper, 2001; Morley, Evans, & Handley, 2004), failed to replicate these effects, or at best found weak associations between figure and endorsement rates.

Linguistic theories

The first of the three theories with a linguistic basis is *conversion theory* (Chapman & Chapman, 1959). As one of the earlier theories in the history of syllogistic reasoning, conversion theory claims that people treat the quantifiers *all* and *some .... not* as though they imply their converses, for example, All of the *A's are B's* implies that *All of the B's are A's*. This fallacious inference can perhaps be best illustrated by a statement using realistic content: *All cats* (A) *are animals* (C), does not mean *All animals* (C) *are cats* (A). Support for the conversion theory is consistent across studies (i.e. Dickstein, 1978; Revlis, 1975), although results for the also irreversible problem *Some .... not* are less convincing, and the theory does not extend to *Some* and *No* statements, because *Some of the A's are B;s* can also be correctly interpreted as *Some of the B's are A's*, and *None of the A's are B's* can also be correctly interpreted as *None of the B's are A's.*

The second linguistically based explanation of syllogistic reasoning, is *conversational implicature*, which is rooted in the Maxim of Quantity (Grice, 1975), and states that speakers should be as informative as possible and not deliberately withhold information they know to be true. Therefore if the speaker means *all*, then they should say *all*, rather than *some*. There is a wealth of early reasoning research producing evidence to suggest that *some* is frequently interpreted as *some* but not *all,* rather than

the logical interpretation of *some*, and *possibly all* (Begg & Denny, 1982; Newstead & Griggs, 1983; Newstead, Pollard, & Riezebos, 1987). Later studies of syllogistic reasoning however (i.e. Newstead, 1995), report that the effects diminish as the logical demands of the task increase.

More recently, Schmidt and Thompson (2008) looked at whether pragmatic responses explain some of the errors in syllogistic reasoning, regardless of the logical demands; by replacing the standard particular premises *some* and *some … not* with *at least one* and *at least one … not*. Schmidt and Thompson (2008) found that reasoning performance was significantly improved, which raises interesting questions as to whether standard quantifiers should be clarified in reasoning tasks, and also suggests that Gricean implicature *does* impact on performance. While discussion of this is beyond the scope of the thesis, it is worthy of further investigation at a later date.

The review of effects and behaviours reported in studies of syllogistic reasoning was perhaps made simpler because it is a relatively small field compared to transitive inference and conditional inference. The next section will review some of the transitive inference literature, while at the same time remaining focussed on the topics which are most relevant to this thesis.

Transitive Inference

Transitive inferences are made by all individuals on a daily basis, when required to decide between three or more entities based on their relative attributes; for example, if *textbook A* is easier to understand than *textbook B*, which is in turn is clearer than *textbook C*, then a student may consider that *textbook A* is probably the best book to

purchase. However, it is important to differentiate between these and intransitive relationships which cannot be arranged on a linear scale, such as:

> James is the father of Harry
> Harry is the father of Charlie

As well as transitive and intransitive relationships there are also atransitive relationships, when nothing can be deduced from the premises. Consider the example shown below, when it is unclear whether James, Harry and Charlie are standing in a triangle or in a row:

> James is next to Harry
> Harry is next to Charlie

Within the transitive inference paradigm, negations (not as long as), inverse negations (not as short as) and inverse relational adjectives (shorter than) increase the number of possible combinations, and the speed and ease with which individuals make inferences depends largely on the combination that is presented. Using length as a property, the terms ABC might be expressed as:

> A is longer than B   or as   B is not as long as A
> C is shorter than B   B is longer than C

Typically, transitive inference studies use either 3-term series problems (sometimes referred to as linear syllogisms), which as the name suggests, are constructed from 3 terms, ABC, which can be arranged in a linear sequence according to their relative properties. A number of studies have used 5-term series problems (e.g. Capon,

Handley, & Dennis, 2003; Vandierendonck, Doerckx, & Vooght, 2004), although these tend to be studies designed specifically to investigate working memory, or to test how well people understand spatial descriptions, when the use of 3-term series problems on their own is too simple a task.

Research has reliably shown that performance is affected by the number of arguments in a problem, in line with the Theory of Relational Complexity (Andrews & Halford, 1994, 2002); which argues that the number of interacting variables determines the difficulty in correctly resolving the relationship between entities. Goodwin and Johnson-Laird (2005) also found that complexity affects the ease by which an integrated representation of the premises is formed, for instance A > B, C > D, D > A is difficult because of the need to hold the first two premises (A > B, C > D) in mind, to integrate them with the third premise. The merits of 5-term series problems are discussed in more general texts (e.g. Evans et al., 1993), but for the purpose of this thesis, we will concentrate on 3-term series problems, which are more suited to the aims of the study, both in terms of difficulty and problem complexity

When considering the terms in 3-term series problems, typical tasks contains either abstract terms (A, B and C) or thematic terms (the jug is to the right of the glass), and task requirements are similar to studies of syllogistic reasoning, in that participants are required to either produce a conclusion, or evaluate a given conclusion about the relationship between the terms. Although in deductive reasoning studies it has been found that most people provide the correct answers to simple problems, with correct responses ranging from 81% to 92% (Huttenlocher, 1968) the time taken to reach

these conclusions varies between problems, which suggests variations in difficulty. For instance, take the two 3-term series shown below; the first of these has emerged as being one of the easiest problems, and the second has emerged as being one of the more difficult:

1. B is better than C   2. C is worse than B
   A is better than B   B is worse than A

The two main theories specific to transitive inference, are *spatial array theories* and the belief that deductive reasoning ability on materials with transitive properties is based on the *linguistic representation of the premises*. The theories are contrasting, in that they do not lead to similar predictions about the relative difficulty of the problems.

Spatial array theory

The view put forward by spatial array theorists (De Soto, London, & Handel, 1965; Huttenlocher, 1968) is that reasoners represent the terms as a visual image, and 'read off' the answer by inspecting that image. Take for instance the transitive terms: Ann is taller than Beth, Cath is shorter than Beth. Spatial array theorists would suggest that individuals visualize Ann, Beth and Cath in a spatial array, to reach the correct conclusion that Ann is taller than Cath:
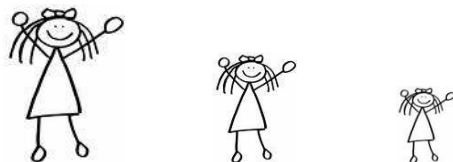
Ann   Beth   Cath

Furthermore, evidence has been found (Clement & Falmagne, 1986; Shaver, Pieron, & Lang, 1975) to suggest that the responses are mediated by the ease in which the given materials can be visualized internally; therefore returning to the above examples, visualizing Ann as taller than Beth, might be easier than visualizing Ann as better than Beth.

Evans et al. (1993) suggest that spatial array theory may be an early precursor to mental model theory (Johnson-Laird, 1983; Johnson-Laird and Byrne, 1991), which proposes that people make inferences by constructing and revising mental models of the premises under consideration. This is a view that is also shared by Knauff, Rauh, Schlieder, & Strube (1998), and is discussed in chapter 3.

Linguistic representation of the premises

On the other hand, those who support a more linguistic explanation for the interpretation of the premises, such as Clark (1969), suggest that reasoners represent the relational meaning of the premises. Therefore rather than integrating the two premises, reasoners represent them by a set of linguistic propositions relating to the underlying meaning of the premise. So, with the premise Ann is better than Beth, reasoners construct a linguistic representation, based on a dimension of goodness:

Ann is more good; Beth is less good

Alternatively, with Beth is worse than Ann, the representation is based in a dimension of badness:

Beth is more bad; Ann is less bad

The differences in difficulty according to the predictions of this theory, are mediated not by how well the materials can be visualized, but by whether or not there are negated propositions; when increased difficulty in constructing negated propositions results in longer response times.

More recently, a number of researchers (Knauff, 1999; Knauff & Johnson-Laird, 2002; Rauh, Hagen, Schlieder, Strube, & Knauff, 2000) have questioned the materials traditionally used in studies of transitivity, such as *left of*, *right of, in front of*, and *behind*, suggesting that they have no clear semantics. This poses the question as to whether the results reported in the literature can be attributed to the inference processes, or whether the ambiguity of the relations also plays a role. Further reference and more detailed discussion takes place in chapter 3, where the development of the experimental materials was heavily influenced by these concerns.

Conditional Inference

The third and final paradigm used in this thesis is that of conditional inference, which is based on *if-then* statements. Conditional inference, which has made a major contribution to our understanding of the processes underlying deduction, is studied in three main ways. The first of these, and the one that was used in the preparation of this thesis, is when people evaluate or generate conclusions relating to four basic conditional inferences (shown in table 1.3), namely Modus Ponens, Modus Tollens, Denial of the Antecedent and Affirmation of the Consequent. The second involves the study of how people understand truth tables: a truth table is based on a mathematical table used in logic to express the truth status of logical connectives as a function of

truth value assigned to its component propositions. The third is the selection task developed by Wason (1966) where participants are shown a set of cards on which there is either a letter or a number. Following this they are given a conditional rule, and asked to decide which of the four cards would need to be turned over in order to decide whether the rule is true or false[4]. The following discussion will focus on studies using the four basic conditional inferences mentioned above.

Over the past decade, the majority of deductive reasoning studies using the conditional inference paradigm, have asked participants to make an inference on the basis of the major premise, if p then q, and the minor premise q. The first term of the major premise (p) is known as the antecedent and the second term (q) is known as the consequent. The four basic inferences that can be made from a major conditional premise are shown in table 1.3.

Table 1.3
The four basic conditional inferences of the form if p then q

Inference	Major Premise	Minor premise	Conclusion
MP: Modus Ponens	if p then q	p	q
MT: Modus Tollens	if p then q	not q	not p
DA: Denial of the antecedent	if p then q	not p	not q
AC: Affirmation of the consequent	if p then q	q	p

In formal logic, MP and MT inferences lead to logically certain conclusions; and the AC and DA forms are logically uncertain or invalid. The number of possible premise

---

[4] See more general texts (Evans et al., 1993; Manktelow, 1999) for discussion of truth table tasks and the Wason selection task.

arrangements can be increased by negating either the antecedent or consequent, or both: *if p then not q*, *if not p then q*, and *if not p then not q* (see table 1.4).

Table 1.4
Conditional inferences with basic and negated major premises

	MP		MT		AC		DA	
	Given	Conclude	Given	Conclude	Given	Conclude	Given	Conclude
If p, q	p	q	not q	not p	q	p	not p	not q
If p, not q*	p	not q	q	not p	not q	p	not p	q
If not p, q*	not p	q	not q	p	q	not p	p	not q
If not p, not q*	not p	not q	q	p	not q	not p	p	q

*conditionals with negated major premises

Although earlier studies within this paradigm tended to focus on conditional inferences with abstract content, there is growing emphasis in more recent literature to use materials with a more everyday or realistic content, which will be discussed in the following two sections.

Conditional inference with abstract content

Studies which use abstract content, or content where no everyday knowledge can be accessed in order to interpret the premises, tend to be similar to the MP inferences shown in the following examples:

> If the letter is a p then the number is a 2
> The letter is a p
> Therefore: The number is a 2

> If Mary is in Paris then Julia is in London
> Mary is in Paris
> Therefore: Julia is in London

Typically, studies with these types of content, present people with conditional statements in each of the four argument types (MP, MP, AC and DA), using basic premises, or all four forms of the major premise. Performance is generally good on the logically valid MP problems, with some studies reporting 100% correct response rates (Rumain, Connell, & Braine, 1983; Wildman & Fletcher, 1977). On the other hand, correct response rates have been found to be lower for MT problems (also logically valid), ranging from 41% to 81% (Evans et al., 1993). The two fallacies, AC and DA are also quite often endorsed as valid arguments, although there has been found to be more variability, between studies not only in terms of results (endorsements ranging from 21% to 75%), but in methodology, making it less easy to draw clear conclusions from the data (Evans et al., 1993).

Conditional inference with everyday content

While conditional reasoning studies with abstract or context free content might help researchers to understand logical competence; over the past decade the impact of background knowledge and prior knowledge on reasoning processes has become dominant in the literature (Byrne, 1989; Cummins, 1995; Cummins, Lubart, Alksnis, & Rist, 1991; Handley & Evans, 2000; Thompson, 1994). One of the key findings is that it is possible to measure the willingness of a participant to fallaciously endorse a conditional inference problem, by explicitly introducing additional information or changing the content of the inference. This has become known as the *suppression effect* (Rumain et al., 1983), and is particularly relevant to this thesis in terms of the selection of materials for experiment 5.

The *suppression effect* therefore occurs when logically valid inferences are suppressed by the content or context of the premises, resulting in reasoning behaviours based on the content or context rather than the logical structure. This often leads to better reasoning on AC and DA inferences. Byrne (1989) found that instead of just saying *if she meets her friend* (*p*), *then she will go to a play* (*q*), but giving additional information such as *if she has enough money,* fewer AC (she will go to play, therefore she meets her friend) and DA (she does not meet her friend, therefore she will not go to a play) endorsements were made. Furthermore, Byrne (1989) also found that when explicitly presenting other reasons why it might not be possible to go to the play, there was an increased rejection of MP (she meets her friend, therefore she will go to a play) and MT (she will not go to a play, therefore she did not meet her friend) inferences. For example, *if she meets her friend*, *then she will go to a play* followed by the additional information of *if the theatre is open.*

The range of literature relating to psychological research on conditional reasoning is vast (see for instance Evans et al., 1993; Evans & Over, 2004; Manktelow, 1999), much of which is not pertinent to this thesis; but relevant findings will be discussed in the introductory sections for experiments 4, 5 and 6, all of which are rooted in making conditional inferences.

General theories of deductive reasoning

There are a number of general theories which have been developed to explain deductive reasoning, and until recently the two major and opposing schools of thought were the theory that reasoning depends either on the manipulation of mental models,

or on logical rules. Although a great proportion of reasoning research tends to fall into one or other of these theoretical accounts, with model theories arguably gathering the most support amongst researchers; there are other theories which warrant discussion. Therefore as well as the mental model theory and rule based theories; the Verbal Reasoning Hypothesis (Polk & Newell, 1995), the Probability Heuristics Model developed by Chater and Oaksford (1999), and various dual processing theories incorporating hypothetical thinking theory (Evans, 2003, 2004; Kahneman, 2003; Sloman, 1996) will also be considered.

Mental Model Theory of Deduction

The mental model theory of deduction (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991) proposes a semantic approach to deduction, whereby reasoners construct and manipulate mental models representing the possible state of affairs consistent with the premises. One of the main tenets of the theory is that deductive competence is achieved by individuals having the ability and/or the desire to search for counterexamples. These processes can be conceptualized in three stages:

Comprehension and model formation: Reasoners combine their general knowledge and knowledge of language to understand the premises, and then construct an internal model from the state of affairs described.

Conclusion formation: Reasoners try to form a parsimonious conclusion by fleshing out the model they have constructed. This conclusion should assert something that is not explicitly stated in the premises. When no such assertion is found, the conclusion is deemed be invalid.

> *Conclusion validation:* Reasoners search for alternative models (counterexamples) of the premises in which their putative conclusion is false. If no such model is found, then the conclusion is valid.

Although the mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991) was originally devised as an account of syllogistic reasoning, and has been widely used in this paradigm (Johnson-Laird & Bara, 1984; Newstead, Handley, & Buck, 1999; Newstead, Pollard, Evans, & Allen, 1992); it has since been adapted to account for performance patterns in other paradigms, such as conditional inference and transitive inference, and these will be clarified in the appropriate chapters of the thesis. The main principles of the theory are illustrated using the syllogistic reasoning paradigm.

Johnson-Laird and Bara (1984) argue that the difficulty of a syllogism is dependent upon the number of models it is necessary to construct when attempting to validate or produce a conclusion. Out of the 27 valid syllogisms, 10 are single model problems, in as much as the conclusion can be drawn from one initial model with no fleshing[5] out required, and the remaining 17 syllogisms are multi-model, because there are two or three possible models of the premises which need to be considered (Johnson-Laird & Byrne, 1991). The notational system developed by Johnson-Laird & Bara (1984) is most frequently used to illustrate each stage of the reasoning process, and is used below to define the reasoning process of a three model syllogism:

> All of the Beekeepers are Athletes
> None of the Beekeepers are Chemists

---

[5] Constructing more models.

These premises initially elicit the following model, where the square brackets indicate that a token is exhaustively represented in the set of models, and the ellipsis '....' indicates that there are alternative models of the premises that are not initially represented. Each line represents a hypothetical individual possessing the characteristics indicated by the token:

     [a   [b]]
     [a   [b]]
                    [c]
                    [c]
     ....

This initial model supports the putative conclusion; *None of the Chemists are Athletes,* and as such has been found to be one of the most common errors with this problem type.

In order to conclude that the conclusion is not necessitated by the premises, the model shown on the following page refutes the initial conclusion, and the two models together support the conclusion *Some of the Chemists are not Athletes*:

     [a   [b]]
     [a   [b]]
      a   [c]
                    [c]
     ....

Model three is a counterexample to the second model, which suggests the conclusion,

Some of the Athletes are not Chemists:

```
[a    [b]]
[a    [b]]
 a           [c]
 a           [c]
....
```

It is these three models that collectively support the valid inference, and the conclusion can only be drawn with certainty after constructing the full set of models of the premises, as shown above. The mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991) proposes that these models are produced in a specific order determined by the way in which the models of individual quantified assertions are combined (Bucciarelli & Johnson-Laird, 1999). This multi-model problem that has been used as an example is one of the hardest syllogisms, which is perhaps not surprising given the number of stages involved. The role of counterexample search to enable validation, is the third stage of the mental model theory, and is more fully discussed later in this chapter.

Mental Logic Theories

Formal logic proposes that a deductive argument is correct *only if* the conclusion is true in all states of affairs in which the premises are true, and probably the best developed mental logic theories to explain the process are those of Rips (1994) and Braine and O'Brien (1998). These theories assume that reasoning is carried out by applying rules

of inference stored in a mental logic. Problem difficulty is accounted for in terms of the number of rules that must be applied, and problems that require long 'proofs', are predicted to be more difficult than those requiring shorter 'proofs'. Accordingly, people reason by abstracting the underlying logical structure of an argument and then applying the inference rules. Mental logic theories propose separate rules for dealing with every connective or quantifier, and a formal proof is a finite sequence of propositions in which each sentence is either a premise, an axiom of the logical system, or a sentence which follows on from a preceding sentence by one of the system rules.

This can be illustrated by one such commonly used rule, the conditional inference *modus ponens:* according to the principle of mental logic, the proposition *if p then q,* and the proposition *p,* entails the proposition *q.* The following example is taken from Rips (1994):

> If Steve deposits 50 pence, Steve will get a coke
> <u>Steve deposits 50 pence</u>
> Steve will get a coke

However, not all inferences are as simple, and the process for *modus tollens* requires a supposition that is more prone to errors, and requires reasoners to disprove a proposition by showing that it leads to an untenable conclusion. So, for instance given the same *if p then q*, and then *not q*; reasoners suppose *p*, infer *q*, conjoin *q* and *not q,* before concluding *not p.* Therefore, with reference to a similar example to the one above, if it is found that *Steve did not get his coke*, the assumption can be made that *Steve did not deposit his 50 pence.* A full set of inference rules based on classical logic can be found in Rips (1994).

When considering how to apply the rules of mental logic to syllogistic reasoning, which include quantifiers such as *All* and *Some*, it can again be illustrated by an example from Rips, (1994, p. 5):

> All square blocks are green blocks
> <u>Some big blocks are square blocks</u>
> Some big blocks are green blocks

 To determine whether the conclusion to the above problem is correct, a reasoner might take an arbitrary big square block and call it 'b'. Block 'b' must be green since 'b' is square and all square blocks are green. Hence, some big blocks (b, for instance) are green, as stated in the conclusion. The proof proceeds by considering an arbitrary example of the premises, tests whether this examples guarantees properties mentioned in the conclusion, and generalises it to the entire conclusion. Again, a full set of inference rules based on classical logic can be found in Rips (1994, p. 52).

In summary, mental logic theories propose that people reason by applying inference rules to the logical structure of the argument, the focus being the interpretation of the premises using linguistic and pragmatic influences. The deduction process is based upon two cognitive skills: the ability to make suppositions or assumptions and the ability to formulate sub-goals within working memory before linking these to reach a conclusion. A more thorough and unbiased explanation of rule based theories can be found in broader texts (i.e. Evans et al., 1993). The next general theory of deductive reasoning to be reviewed is the Verbal Reasoning Hypothesis (VRH) which includes both representational and rule-based processes.

Verbal Reasoning Hypothesis

The VRH (Polk & Newell, 1995) which centres on the role of coding and encoding, proposes that the linguistic ability of people is deployed adaptively to deductive reasoning tasks, and reasoning processes must occur in a way which reflects the needs of deduction, rather than those of everyday communication. The reasoning process involves repeatedly re-encoding the problem until a conclusion is formed, based on a detailed computational model where the initial stage is to construct a mental model of a situation in which the premise in question is true (see Polk and Newell, 1995 for a detailed explanation). Polk and Newell (1995) present a model of reasoning, when the initial stage is to construct a mental model of a situation in which the premise in question is true. The objects represented in this model are annotated by two additional pieces of information; a *not* flag that the object does not have a specific property, and an *identifying* flag indicating that the object is identified by a specific property.

As was the case with the mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991) the VRH (Polk & Newell, 1995) was first developed within the paradigm of syllogistic reasoning. Table 1.5 shows the default encodings for each of the four standard syllogistic quantifier premises, and how these may be changed or augmented when information from a second premise is introduced. The *identifying* properties correspond to the grammatical subject of the premises and are distinguished from other lesser properties by being more readily available; for instance given *All A are B*, the model distinguishes an *A* (identifying) who is a *B*; from a *B* (identifying) who is an *A*. There are often several ways in which premises can be represented; an annotated model may contain information that is *not* inherent in the original premise, or fail to

encode information which *is* inherent. For example given *Some of the A's are B's*, the model may contain unnecessary information if it also encodes the notion that *Some of the A's are not B's*.

When the encoding process has taken place, the reasoning process produces a conclusion based on the annotated model, and conclusion of the form *All of the A's are B's* or *None of the A's are B's,* will be proposed when there is an object with an *A* as an identifying property, and all objects with the property *A* also possess the property *B* or *not B*. Conclusions of the form *Some A are B* or *Some A are not B* will be proposed when there is an object A as the identifying property and at least one other object with the properties *A* and *B/not B.*

Table 1.5
Default encodings for the four syllogistic quantifiers

Premise	Initial model	Augmented model
All of the A's are B	(A' B)	All (A ...) $\longrightarrow$ (A' ... B)
Some of the A's are B	(A' B) (A')	MR (A ...) $\longrightarrow$ (A' ... B) (A')
None of the A's are B	(A' not Y)	All (A ...) $\longrightarrow$ (A'... not B)
Some of the A's are not B's	(A' not Y) (A')	MR (A ...) $\longrightarrow$ (A' ... not B) (A' ...)

' identifying property flag
not = not flag
... other properties

Although the theory, as has already been mentioned, was initially developed within the syllogistic reasoning paradigm, Polk and Newell (1995) ran a number of studies, which produced evidence to show that their computational model provides a good explanation of behaviour on a variety of deductive reasoning tasks. Furthermore it accounts for all of the major phenomena found in deductive reasoning such as the atmosphere effect and figural bias in syllogistic reasoning, together with no valid conclusion responses across problem types. This is fully discussed in a much cited paper (Polk & Newell, 1995); but the general assertion of the VRH is that behaviour can be explained in terms of standard linguistic processes, without the need to posit reasoning-specific mechanisms.

Dual process theories (including hypothetical thinking theory)

It is thought that dual process theory dates back as far as the American psychologist and philosopher, William James (1842-1920), who believed that there were two different kinds of thinking: associative thinking and true reasoning. This belief was based on the view that associative thinking is used for creative things like art, where things are created from past experience, while true reasoning is used for navigating obstacles that have not previously been experienced.

Within the psychology of reasoning over the past two decades, several researchers have revisited dual process theory (i.e. Evans, 2008; Evans & Over, 1996; Kahneman & Frederick, 2002; Sloman, 1996; Stanovich, 1999)[6]. While there is some variation in

---

[6] See Evans (2008) for a comprehensive discussion and review of the literature relating to dual process theory.

these accounts of dual process theory, they all share a number of general characteristics in that on the one hand there are the fast, automatic and unconscious cognitive processes of system 1, while the system 2 processes are slow deliberative and conscious. In terms of the two different kind of thinking proposed by James (1842-1920), system 1 is akin to associative thinking, and system 2 is analogous to true reasoning. The general characteristics[7] are set out in table 1.6:

Table 1.6
Dual Process Theory typical characteristics relating to deductive reasoning

System 1: Implicit	System 2: Explicit
Unconscious	Conscious
Automatic	Controllable
Independent of language	Related to language
Pragmatic/contextualized	Logical/abstract
High processing capacity, parallel	Constrained by working memory, sequential
Driven by learning/innate modules	Permits hypothetical thinking
Independent of general intelligence	Correlated with general intelligence

Although different proponents of dual process theory have proposed names for these two kinds of thinking, Evans (1989) refers to these systems as heuristic and analytic processes, where heuristic processes (system 1) are pragmatic and preconscious, which act to form selective mental representations of reasoning problems. This is carried out by representing problem features and applying relevant prior knowledge which is retrieved from long term memory. On the other hand analytic thought (system 2) is involved in abstract hypothetical thinking and logical reasoning, before

---

[7] See Evans, 2008 for a comprehensive discussion and review of the literature relating to dual process theory.

subsequently being applied to the selective representations. Dual process theories tend to portray heuristic and analytic elements as *competing*, thus explaining cognitive biases in terms of faulty heuristic processing. This is perhaps best explained by taking an example from the belief bias paradigm, which is where there is a conflict between the believability of the conclusion and its underlying logical status. Consider for instance, the following argument (taken from Sa, West, & Stanovich, 1999):

> All plants need water
> Roses need water
> Therefore, roses are plants

The heuristic-analytic dual process account (Evans, 1989) suggests that the heuristic type 1 response is to endorse the conclusion, because it is consistent with the underlying beliefs that roses are plants; although the logically correct rejection of the conclusion requires a more deliberate analytic type 2 process.

However, although the various accounts of dual process theories have been widely researched leading to a large number of publications, there some controversial aspects of dual-system theories particularly relating to consciousness and evolution. Although discussion of these areas is not within the scope of this thesis, they are discussed in detail by Evans (2008). However, it is these contentious areas that led to the development of a revised version of the heuristic-analytic theory (Evans, 2007) which minimizes these issues, and also provides a more inclusive theoretical framework to explain hypothesis testing, forecasting, decision making, counterfactual thinking, and suppositional reasoning. This revised theory is called hypothetical thinking theory, and is based around three principles:

The singularity principle: People consider a single hypothetical possibility, one at a time. This is because hypothetical thinking requires use of system 2, which is constrained by working memory and sequential in nature. People often consider more than one possibility, but not at the same time.

The relevance principle: People consider the possibility that is most relevant to the current goals (generally the most plausible).

The satisficing principle: These possibilities are evaluated with reference to the current goals and accepted if satisfactory, unless there is a good reason to reject, modify, or replace them.

Hypothetical thinking theory allows that biases will also arise in analytic processing; because with the singularity and satisficing principles, the implication is that one model is considered, and accepted if there is no good reason to reject that model. The dual process account of deductive reasoning is still retained as a processing model, and thinkers are required to imagine possible states of the world. Evans (2007) offers a range of experimental evidence to support hypothetical thinking theory, but perhaps that which is most relevant to this thesis is research carried out by Evans et al. (1999), which is discussed in more detail later in this chapter, and in chapter 2. Evans et al. (1999) concluded that individuals do not search for counterexamples in syllogistic reasoning by default, but merely form a single mental model of the premises and stick with it unless there is reason to search further. If the conclusion that is presented to participants is consistent with the model, then they conclude that it satisfies the current goals, and only when the conclusion cannot be reconciled with the model of the premises is it rejected.

Probabilistic Reasoning

The final theory to be reviewed is a probabilistic account of deductive reasoning (Chater & Oaksford, 1999, 2001), which provides an explanation for reasoning behaviours based on probability theory, rather than on logic. Accordingly, the errors and biases which have been reported in the literature across the reasoning paradigms are thought to occur because people import their everyday uncertain reasoning strategies into the experimental laboratory. Therefore, rather than suggesting that people are trying, but failing to correctly evaluate or produce a conclusion to a logical deductive reasoning problem, it suggests that people are drawing probabilistic inferences, in attempting to choose between probabilistic models of the world. The theory is based upon a number of heuristic processes specific to each individual reasoning paradigm.   It is a complex theory, composed of computational and algorithmic levels; and although it is not proposed to cover all of the paradigms (see Oaksford & Chater, 2001), a simplified form of the probabilistic inferences for syllogistic reasoning is shown below. There are five basic heuristics, of which the first three relate to generating a conclusion, the fourth relates to conclusion order, and the fifth and final one relates to testing the conclusion:

> *Min-heuristic*: the quantifier selected will be the same as the quantifier
> in the least informative premise
>
> *P-entailments* (the next most preferred conclusions to those predicted by
> the min-heuristic): a conclusion will be selected that is probabilistically
> entailed by the min-conclusion so, if *all* C are Y, as long as there are *some*
> X's, it is probable that *some* X are Y.

Max-heuristic: confidence in the min-conclusion is determined by the expected information conveyed by the most informative premise

Attachment-heuristic: If the min-premise has an end-term (A or C) as its subject, this will then become the subject of the conclusion.

O-conclusions: Avoid producing or accepting these (some …. not) as they are uninformative relative to other forms of conclusion.

Take for instance the following example (Chater & Oaksford, 1999):

All Y are X	(max-premise)
Some Z are Y	(min-premise)
I-type conclusion	(by min)
Some Z are X	(by attachment)

First, by the min-heuristic, the conclusion is I (some). The min-premise has an end term (Z) as its subject, therefore by attachment the conclusion will have Z as its subject term and the form some Z are X. If however, the order of terms in both of the premises were reversed, and the min-heuristic also specifies an I conclusion, the I premise does not have an end term (X or Y) as its subject so the conclusion would be some X are Z.

The PHM has been found to provide an accurate account of syllogistic reasoning Chater and Oaksford (1999), following a meta-analysis of data from 5 earlier experiments (Dickstein, 1978; Johnson-Laird & Bara, 1984; Johnson-Laird & Steedman, 1978), using all 64 syllogistic forms. Furthermore, it also provides an accurate account of 'no valid conclusion' responses, which is lacking in other explanations.

Summary of theories and materials

The reasoning theories which have been reviewed each provide a different account of deductive reasoning, and although the majority of them apply the same principles across reasoning paradigms, the PHM (Chater & Oaksford, 1999) is paradigm-specific in that there is a different heuristic process for each paradigm. Furthermore, although the VRH (Polk & Newell, 1995) and hypothetical thinking theory (Evans, 2007a) are both model based theories, the VRH does not include a falsification process, while hypothetical thinking theory (Evans, 2007a) allows that in some instances some individuals carry out a search for counterexamples. .On the other hand, one of the key principles of the mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991), is that the search for counterexamples is part of the default reasoning process when seeking to reach a conclusion as to the validity of a logically invalid conclusion. Evidence to support this belief will be highlighted in the following section.

The search for counterexamples

Over the past two decades it has been argued by many that the search for counterexamples as proposed by mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991) is central to the deductive reasoning process, and a large body of empirical evidence has been produced to support this (Evans, Barston, & Pollard, 1983; Newstead, Handley, Harley, Wright, & Farrelly, 2004; Schroyens, Schaeken, & Handley, 2003; Stanovich & West, 1998a; Torrens, Thompson, & Cramer, 1999). It is proposed that reasoners do not prove conclusions syntactically by applying inference rules or algorithms, or make judgements based on the probability of an event; but merely base

deductions on grasping a semantic principle, namely that a conclusion is valid if there is no model of the premises that excludes it.

However, conflicting evidence has been offered (i.e. Bucciarelli & Johnson-Laird, 1999; Newstead et al., 1999) to suggest that despite having the *ability* to search for counterexamples, many reasoners often fail to do so; instead accepting or offering a conclusion that is consistent with the preferred initial model that is constructed, and rejecting conclusions that are inconsistent with this model. This view is also shared by Evans et al. (1999) who collected a large amount of experimental data using all possible combinations of syllogistic premise and quantifiers; when it was found that under instructions of logical necessity, the first model was frequently accepted by participants, rather than searching for counterexamples to falsify a putative conclusion.

What was a particularly interesting finding in the study carried out by Evans et al. (1999), and in a sense key in facilitating the experiments reported in this thesis, is that while some fallacious syllogisms were endorsed as consistently as valid syllogisms; others were endorsed as infrequently as syllogisms which were presented with an impossible conclusion, or in other words one which does not follow from the premises. Evans et al. (1999) found that reasoning errors were more likely to occur when the conclusion was consistent with the first model of the premises, despite the existence of alternative, falsifying models, and Evans et al. (1999) termed these possible strong (PS) syllogisms. In contrast, syllogisms with conclusions that were not consistent with the first model were termed possible weak (PW) syllogisms, as significantly fewer were incorrectly endorsed. In other words, PS syllogisms are the fallacies that individuals

tend to make, and PW syllogisms are the fallacies they tend to avoid. This also poses another question which is whether this effect is consistent across reasoning paradigms and types of content, such as abstract or everyday content.

In addition to examining the inferences that individual were prepared to make under instructions of necessity, Evans et al. (1999) also posited that in everyday life it is just as important to decide whether a proposition is possibly true. Although the majority of psychological studies of deductive reasoning reported in the literature have asked participants only to decide if a conclusion is *necessary* following a set of given premises, a small number of studies (Bell & Johnson-Laird, 1998; Evans et al., 1999; Galotti, Baron, & Sabini, 1986; Osheron, 1976) introduced a condition in which participants were asked to make judgement on the *possibility* of a conclusion being correct. This will be discussed further in chapter 2.

It would appear therefore that there is evidence to suggest that under some circumstances, some individuals do search for counterexamples; but it may well be that the possibility of reasoners carrying out a search for counterexamples as proposed by mental model theory is dependent upon having the cognitive ability to do so, which is another factor this thesis sets out to explore.

Individual differences in cognitive ability

Traditionally, in psychological studies of deductive reasoning, investigators have focussed on performance between groups, to explore the effects of various experimental manipulations; for example how performance in syllogistic reasoning is

affected by the structural properties of the syllogism, or the effect of content on the inferences that people are prepared to make.

However, Feeney (2007) highlights the need to clarify our understanding of who does what in reasoning experiments, and while many interesting phenomena are reported, the individual differences in cognitive ability is frequently ignored, or is not taken as a starting point for investigations. The relatively small group of researchers (Klaczykski & Daniel, 2005; Newstead et al., 2004; Stanovich & West, 2000; Torrens et al., 1999) involved in individual differences research, have sought to identify factors such as levels of cognitive ability, which facilitate and inhibit logical reasoning performance, although ability has not been the prime motivation behind these studies.

Among the most frequently adopted tests used to explore cognitive ability are the AH4 and AH5 tests of cognitive ability, which are long established, well validated tests of general intelligence developed by Heim (1968). The Scholastic Assessment Test[8]; has also been extensively used in what is arguably one of the largest bodies of individual differences literature, published by Stanovich and West (1999, 2008; 1998a, 1998b, 2000). This program of research reports reliable correlations (.47 and .41, at a probability level of .001) between ability and logically correct performance on a number of reasoning tasks, most notably syllogistic reasoning, suggesting that cognitive ability is a good predictor of performance on syllogistic reasoning tasks.

---

[8] A standardised test for college admission in the University States of America, first introduced in 1901; owned, published and developed by the College Board, and used to assess student's readiness for college.

Research has also shown (i.e. Stanovich & West, 1999) that cognitive ability plays a role in deductive reasoning, when there is a conflict between the believability of the conclusion under evaluation and the logical correctness of the conclusion. Participants were found to be more able to disassociate the content and the logical structure of deductive reasoning problems, in order to provide a logically correct response. There is also evidence from the conditional reasoning literature (Evans, Handley, Neilens, & Over, 2007; Newstead et al., 2004) that performance on MP (if p then q, p, q), AC (if p then q, q, p) and DA (if p then q, not p, not q) is highly correlated with cognitive ability, although this effect has not been found to extend to MT (if p then q, not q, not p) conditionals.

However, there is little research within the transitive inference paradigm investigating the relationship between cognitive ability and correct performance; as the main focus of studies of transitivity tends to be on how the terms are represented; either spatially (De Soto et al., 1965; Huttenlocher, 1968) or linguistically (Clark, 1969).

Each of the paradigms used in the preparation of this thesis, will explore the relationship between cognitive ability and reasoning performance, to enable discussion in the final chapter as to whether the findings are generalisable across paradigms, or whether they are domain specific.

Summary and brief overview of the experimental studies

A long history of research in the field of deductive reasoning has generated a large body of literature, leading to the development of a number of general theories to explain the findings. However, until now, studies have tended to focus on one experimental

paradigm, and the majority of research has asked individuals to evaluate problems for logical necessity. The program of research reported in this thesis not only compares logical responses across a number of paradigms, problem types and content, but it also looks at responses when individuals are asked to make judgements as to whether given conclusions are possible.

The four experimental chapters report studies in which the participants are presented with a range of problem types, under instructions of necessity, and instructions of possibility; following which the results are examined in terms of endorsement rates and reasoning times, with reference to cognitive ability. In a replication and extension of previous research carried out by Evans et al. (1999), experiment 1 adopts syllogistic reasoning tasks, and adds to previous work by including a measure of cognitive ability and the collection of reasoning times. Experiments 2 and 3 extend the methodology to transitive inferences, by considering the importance of training in the relational terms used, with respect to endorsement rates and latencies. Experiment 4 looks at the conditional inferences that individuals are prepared to make when inferences are presented with abstract content, and experiments 5 and 6 adopt causal conditional inferences to look at the impact of the number of other possible causes to a given scenario (experiment 5), and to specific scenarios (experiment 6). Each chapter provides a comprehensive review of relevant research, together with a clear rationale and explanation for the selection of materials.

The research is motivated by the mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991). The general predictions are that people will search for

counterexamples on indeterminate deductive reasoning problems presented under necessity instructions, when the initial model supports the premises; and people will search for alternative models when the first model does not support the premises, under possibility instructions. It is also predicted that higher ability participants will be more likely to successfully carry out this search, which will in turn lead to comparatively longer reasoning times.

Chapter 2

The search for counterexamples & alternative models in syllogistic reasoning

The main theories and effects in syllogistic reasoning, reported in the literature, were reviewed in chapter 1. This chapter will present an experiment where participants were required to evaluate a number of abstract syllogistic reasoning problems, to explore the reasoning processes in terms of whether reasoners searched for counterexamples as claimed by the third stage of the mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991). Reasoning behaviours were also recorded for problems requiring judgements of possibility in the light of given information, which is less common in the literature, but equally important in helping to establish whether reasoners can and do search for other models, when deciding whether to accept a given conclusion on an invalid syllogism. The predictions were based on the mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991), in that people would search for counterexamples on indeterminate syllogisms in order to reject a given conclusion under necessity instructions, and accept a given conclusion under possibility instructions to accept a given conclusion.

2.1 Introduction to experiment 1

The mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991) proposes that the deduction process takes place in three stages: first, reasoners construct a set of models corresponding to possible state of affairs in which the premises are true; second, the models are inspected and an initial conclusion is drawn; and finally, a reasoner carries out a counterexample search, or in other words a search for an alternative model or models in which the premises are true but the conclusion is false. If no such model is found, the conclusion is deemed valid, but if a model is found in which the premises are true but the conclusion is false, the conclusion is judged to be invalid. Consider for instance the following invalid syllogism, which was accepted as valid by all but one of a group of participants in a study by Ford (1995), and from a mental models perspective is due to a failure to search for counterexamples:

> None of the A's are B's
> All of the B's are C's
> Conclusion: None of the A's are C's

The initial model of the premises supports the conclusion *None of the A's are C's*, which is shown below using the notational form discussed in chapter 1:

> a
> a
>   [b]   c
>   [b]   c

Although many reasoners fail to progress past this first model, by carrying out a search for counterexamples, there are alternative models of the premises that falsify this

putative conclusion, in other words the model is consistent with an alternative conclusion when *Some of the A's are C's*:

```
a
a          c
    [b]    c
    [b]    c
```

Furthermore, this conclusion is falsified by a third model in which *All of the A's are C's*:

```
a          c
a          c
    [b]    c
    [b]    c
```

There is however a valid conclusion, which is: Some of the C's are not A's.

Since the prime concern of experiment 1 was to investigate if and under what circumstances, the search for counterexamples or alternative models takes place, past research will be reviewed on the search for counterexamples, before presenting the rationale for the experiment. Although the mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991) is currently the most influential theory in syllogistic reasoning research, evidence to support the search for counterexamples is mixed, with studies reporting conflicting evidence (Bucciarelli & Johnson-Laird, 1999; Evans et al., 1999; Newstead, Thompson, & Handley, 2002).

One of the earlier studies which supports the notion of a search for counterexamples, was carried out by Byrne and Johnson-Laird (1990), who found that participants

fleetingly considered erroneous conclusions based on initial models, only to later reject them as a result of a counterexample search. These assumptions were drawn on the basis that conclusions which were falsely recognised by participants, were consistent with an initial model of the premises, on syllogisms to which they had earlier correctly concluded that nothing followed. While this is a plausible conclusion to draw, it may well be that there is a simpler explanation to account for the findings, and perhaps reasoners were merely re-solving the original syllogism and coming up with a different conclusion.

In a later study by Bucciarelli and Johnson-Laird (1999), their methodology was extended, and participants were video-taped performing a task where they were instructed to make cut-out shapes to represent the different classes of individuals. Bucciarelli and Johnson-Laird (1999) found that people constructed more models for multiple model syllogisms, than for single model syllogisms. While this does not provide firm evidence that participants were constructing alternative models in order to produce the correct response, it does suggest that they were able to construct more than one model if required. One weakness of this study (Bucciarelli & Johnson-Laird, 1999) is that it failed to report whether the number of models was predictive of logical accuracy, and it may well be that some or even many of the responses given by participants were logically incorrect.

The question of whether the number of models produced was predictive of logical accuracy was answered in a study published at around the same time (Newstead et al., 1999), which adopted a similar process-tracing methodology. The results contradicted

those of Bucciarelli & Johnson (1999), in that not only were the number of models constructed by participants not predictive of logical accuracy, but the study failed to provide evidence that the number of models constructed for single model syllogisms varied from the number of models constructed for multi model syllogisms. At the time, it was posited that these findings were more consistent with the VRH (Polk & Newell, 1995) than Johnson-Laird's mental model theory.

Although the VRH is a model based theory, where reasoners evaluate a conclusion by repeatedly re-encoding the problem, there is no assumption that falsification takes place, which of course if one of the assumptions of the mental model theory. Instead it is assumed that reasoners attempt to construct only a single model or representation of the premises, and base their judgements of validity on that one representation. According to the VRH, the default reasoning mechanism is that reasoners do not proceed past the first model to find one that falsifies the conclusion.

A more recent model based theory which may also explain the findings, and one which will be considered in the context of the current experiment, is the hypothetical thinking theory, proposed by Evans (2007a), which claims that reasoners can proceed beyond the first model but do not always do so. Hypothetical thinking theory (Evans, 2007a) allows that in some cases some people search for counterexamples to evaluate or provide a correct conclusion to syllogistic reasoning problems. The theory is based on the belief that when evaluating a putative conclusion, reasoners do not carry out a search for counterexamples if the first model satisfies the requirements of the task (the satisficing principle). In other words, if a model is found that is consistent with the

conclusion, no further reasoning takes place and the erroneous conclusion is accepted as being valid. The theory does however, allow that in some instances further searching does take place.

A frequently cited study which seems to provide overwhelming support for the hypothetical thinking theory was carried out by Evans et al. (1999). Following the presentation of computer generated syllogistic reasoning problems to participants, it was found that reasoners seldom went beyond the initial model. Instead participants chose to accept conclusions that were consistent with the preferred initial model (see Bucciarelli & Johnson-Laird 1999), and reject conclusions that were inconsistent with this model. It was also found that the frequency of errors was higher on invalid syllogisms, when the conclusion was consistent with the initial model when there were alternative falsifying models, which suggests that people can search for counterexamples but do so infrequently, instead preferring to accept a satisfactory solution which is not necessarily the optimum one.

Further evidence to support the notion that some participants do carry out a search for counterexamples was provided by Newstead, Thompson and Handley (2002) who looked at the ability of participants to generate different representations of pairs of syllogistic premises. Newstead, Thompson and Handley (2002) found that while some people failed to proceed beyond the one model, others did, and these differences in reasoning behaviours or 'reasoning styles' as they were referred to, were predictive of whether or not a person searched for alternative representations.

To summarise, although the implications of the studies reviewed so far suggest that at least some reasoners are capable of constructing alternative models, they do not lead to the conclusion that the search for counterexamples is a *compulsory* component of the mental model theory. However it may well be that this ambiguity can be clarified by inspecting reasoning times in addition to endorsement rate data.

2.1.1 Reasoning times

Mental model theory predicts that a reasoner will take longer on syllogisms that require the consideration and processing of multiple models. More particularly, if a search for counterexamples is required in order to produce a logically correct response to a syllogism, the time course of the reasoning process should be longer than for those syllogisms not requiring such a search. There are however few studies which have successfully collected data illustrating the time course of syllogistic reasoning. Evans et al. (1999) for instance collected latency data, but because the experimental design was such that reasoners were asked to evaluate four possible conclusions for each problem, the researchers were unable to isolate the length of time that participants took on each possible conclusion; and for this reason the data was not included in the final analysis.

An early study which did successfully collect latency data was run by Galotti et al. (1986), who found that good reasoners took proportionately longer than poor reasoners[9] on invalid problems requiring the generation of at least two models to falsify the initial conclusion. In contrast, on valid conclusions that did not require the

---

[9] Reasoners were categorised by means of a pre-test condition, and those selected for the main study had scores either in the top third or the bottom third.

generation of extra models, response times were similar for good and poor reasoners. One aspect of this study that is noteworthy is that a training session was carried out prior to presentation of the problems, to ensure that participants fully understood the terms. Also, participants were given a booklet in which to make notes while solving the problems, which is again uncommon in studies of syllogistic reasoning.

However, more recent studies (Stupple & Ball, 2008; Thompson, Striemer, Reikoff, Gunter, & Campbell, 2003) have found that reasoners take significantly longer to process invalid problems than valid problems, suggesting that invalid problems require either more effortful reasoning, or involve more processing stages than valid problems. These findings support the view that valid conclusions can be accepted for logical correctness without spending extra time searching for alternative models, while invalid problems take longer because they require the construction of falsifying models in order to correctly falsify the initial conclusion model that comes to mind.

Another factor which may have a bearing on whether or not there is sufficient evidence to support the search for counterexamples, is whether there are individual differences in cognitive ability.

2.1.2 Individual differences in cognitive ability

Although the relationship between reasoning performance and cognitive ability is well established in the literature (Torrens et al., 1999; Newstead et al., 1992; Stanovich & West, 1998b; Klaczynski, Fauth and Swanger, 1998 Galotti et al., 1986) a surprisingly small number of these studies have used the type of categorical syllogisms referred to in this chapter. Despite being frequently taken as evidence to support a positive

correlational relationship between syllogistic reasoning and cognitive ability, some studies have employed other reasoning tasks. For instance, Torrens et al., (1999) reported reliable correlations between performance and *conditional* syllogisms, and the relationship between reasoning performance and cognitive ability reported by both Klaczynski et al. (1998) and Stanovich and West (1998b), was found on the Wason selection task.

Notwithstanding this, one study which did explore the relationship between cognitive ability and Aristotolean categorical syllogistic reasoning performance was carried out by Newstead et al., (2004), who found a significant positive correlation between logically correct reasoning performance and cognitive ability. One of the claims made by mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991) is that reasoners will search for counterexamples on indeterminate reasoning problems; and failure to do so will result in them incorrectly endorsing a conclusion that is not necessarily true, but merely possibly true. It is therefore fair to assume that those reasoners who are better at constructing alternative models will also score more highly on cognitive ability tests such as Heim's AH4/AH5 cognitive ability tests referred to in the first chapter of this thesis.

 The next section will look at whether the key to reaching a clearer understanding on the role of counterexamples in reasoning, is to focus on judgements not just of necessity, but also on judgements of possibility (Evans et al., 1999). At this point it may be useful to clarify the distinction between the terms 'the search for counterexamples' and 'the search for alternative models', as these are terms which will be used

throughout the thesis. The search for counterexamples is specific to the third stage of the mental model theory, when a search for counterexamples under instructions of logical necessity serves to falsify a given conclusion. The term alternative models, is a more generic term and can also be used with instructions other than necessity; or when not seeking to falsify a given conclusion but merely to explore other possibilities.

2.1.3 Reasoning about Necessity and Possibility

Studies of reasoning behaviours, using the instructions 'is it necessary that' and 'is it possible that', have become more common following a large syllogistic reasoning study which was carried out by Evans et al. (1999). Typically in this type of study, participants are asked whether a conclusion necessarily follows, or whether it possibly follows; with a statement following *necessarily* if it must be true and *possibly* if it may be true. Consider the following three arguments taken from Evans et al. (1999), which are based on universal premises (all or none), and presented with thematic content:

1. All artists are beekeepers,   (Necessary problem)
   Lisa is an artist
   Lisa is a beekeeper (necessarily true)

2. All artists are beekeepers   (Possible problem)
   Lisa is a beekeeper,
   Lisa is an artist (possibly true)

3. All artists are beekeepers   (Impossible problem)
   Lisa is an artist
   Lisa is not a beekeeper (impossible)

When considered under necessity instructions the first argument is a valid inference; if a reasoner assumes that Lisa is an artist, and that all of the artists are beekeepers, it necessarily follows that Lisa is a beekeeper. Argument 2 is invalid because although all artists are beekeepers, there may be beekeepers who are not artists and Lisa may be in that group. Finally, argument 3 is invalid (impossible), as there are no models that hold in which Lisa is not a beekeeper given she is part of the group of artists who are all beekeepers.

However, under possibility instructions, again the first conclusion is both possible and necessary. The conclusion to argument 2 is possible but not necessary, since although Lisa is a beekeeper, there are beekeepers who are artists, and beekeepers who are not artists, so she could be in either group. With argument 3, there are no models that hold in which Lisa is not a beekeeper so it is an impossible conclusion.

Within the framework of mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991), there are three predictions that can be made regarding performance under different types of instruction: First, reasoners should be more willing to judge that a conclusion is possible than it is necessary, based on the notion that only one model will suffice for a possible conclusion. Second, it should be easier to decide that a conclusion is possible if it was also necessary, because necessary conclusions hold in all models of the premises. Third, it should be easier to decide that a conclusion is not necessary if it is also not possible, because there are no models that support the conclusion with impossible problems.

When Evans et al. (1999) presented all 256 syllogisms under both necessity and possibility instructions, they found evidence to support all of the above predictions: participants more frequently endorsed conclusions as possible, as opposed to necessary; there were more endorsements of possibility for statements that were necessarily true, than for statements that were possibly true; and there were more endorsements of problems that were possibly true than impossible. Evans et al. (1999) also found in experiment 3, that some arguments supporting possible conclusions were regularly taken to imply necessary conclusions, and some arguments supporting possible conclusions were rarely taken to imply necessary conclusions; although in both instances the logically correct response was not to endorse the conclusion, for instance:

	None of the A's are B's All of the B's are C's	
is it necessary that	None of the A's are C's	77% of people endorsed the conclusion

	None of the B's are A's All of the B's are C's	
is it necessary that	All of the A's are C's	10% of people endorsed the conclusion

These two types of problems were termed Possible strong (PS) and Possible weak (PW) and results indicated that PS problems were endorsed almost as frequently as Necessary problems, while PW problems were endorsed almost as infrequently as Impossible problems. The two groups were thought to have emerged because many people reason based upon the first model that comes to mind, and PS syllogisms have

an initial model that supports the given conclusion; whereas the PW syllogisms have an initial model that does not. This suggests that people do not go beyond the first model, leading to high endorsement rates when the conclusion is consistent with the first model (PS), and low endorsement rates when the conclusion is not supported by the first model (PW).

2.1.4 Aims and rationale

The main aim of experiment 1 was to investigate whether people spontaneously search for counterexamples as proposed by the mental model theory (Johnson-Laird, 1983), and the extent to which this search is mediated by cognitive ability.

The experiment is a replication and extension of the third experiment carried out by Evans et al. (1999), and the construction of problem types is strongly informed by this work. A similar methodology in terms of problem type, instruction and presentational methods was used, but with the addition of a measure of cognitive ability. The time course of the reasoning process was recorded, but in contrast to Evans et al. (1999) participants were only required to evaluate one conclusion, as opposed to four conclusions; resulting in only one overall reasoning time being recorded. This latency measure enabled the detection of instances where extra processing was required to search for additional models, in order to correctly evaluate a given conclusion. The following four problem types identified by Evans et al. (1999) were Necessary, PS, Impossible and PW, and when presented under necessity (is it necessary that) and possibility (is it possible that) instructions, their properties were (see following page):

Necessary - the conclusion must be true

Possible strong - the conclusion may be true (frequently endorsed)

Impossible - the conclusion must be false

Possible weak - the conclusion may be true (infrequently endorsed)

The key comparisons of interest were Necessary and PS problems under necessity instructions, and Impossible and PW problems under possibility instructions. This is because in order to give the correct response to a PS problem under necessity instructions, a search for counterexamples is needed to find a model that negates the initial conclusion; whereas on Necessary problems a search is not required, since all models support the conclusion. Similarly, on PW problems under possibility instructions, a search for alternatives is necessary in order to produce a correct conclusion, because the first model negates the conclusion; while on Impossible problems no models support the conclusion. In this way it was possible to determine whether the required search for counterexamples or alternatives took place, and the measure of ability allowed comparisons to be made as to whether this was mediated by cognitive ability in terms of higher ability people being more likely to carry out this search. The problem categories are shown in table 2.1; where problems requiring a search for alternative models are marked with an asterisk.

It is important to clarify at this stage, that although there is no evidence to suggest that people know in advance whether they need to search for counterexamples on Necessary syllogisms under necessity instructions, we do know that they only need to confirm that a given conclusion in a conclusion evaluation task is correct. On the other

hand, for syllogisms with a PS structure, people need to take action in terms of searching and finding counterexamples to provide the correct response.

Table 2.1
Problem types and logical definitions for each of the four problem categories

necessity instructions	Necessary *no search required*	PS* *the first model supports the conclusion*	Impossible *no models support the conclusion*	PW* *the first model negates the conclusion*
possibility instructions	Necessary *no search required*	PS *the first model supports the conclusion*	Impossible *no models support the conclusion*	PW* *the first model negates the conclusion*

☐ correct response is 'yes'   ▨ correct response is 'no'

Similarly, under possibility instructions, we know that although people may not know that they do not need to search for alternative models on Impossible problems, we know that if they search for and find alternative models on PW problems, this will allow them to provide the correct response.

The follow on from this is that if people are searching for counterexamples or alternative models; detecting them and making judgements as predicted by the mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991) is more demanding in cognitive resources and time, because finding a counterexample or alternative model, and rejecting a conclusion will take longer. This is strongly supported by work carried out by Stupple and Ball (2008), referred to previously, when conclusion

inspection times increased for invalid syllogisms over valid syllogisms, and conclusion inspection times increased for invalid problems as opposed to valid problems.

Earlier work by Clark and Chase (1974; Clark & Clark, 1977) also supports the claims made by mental model theory (Johnson-Laird & Byrne, 1992, p. 52) that fleshing out and constructing a full set of models takes time. Clark and Chase (1974; Clark & Clark, 1977) employed sentence picture verification tasks, and found that participants took longer to make judgements where the conclusion was false, as opposed to when the conclusion was true. They attributed this to the time it took to detect alternative models.

2.1.5 Predictions

There are a number of specific predictions that can be made within the framework of the mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991), and the work carried out by Evans et al. (1999). These relate to the search for counterexamples or alternatives, in terms of the relationship between performance and cognitive ability, and how these interact with the time course of the reasoning process:

1. In line with previous research (Evans et al., 1999), conclusions will be more frequently endorsed under possibility instructions than under necessity instructions.

2. If participants search for counterexamples: under necessity instructions, there will be fewer endorsements of PS problems than Necessary problems because a search for alternatives will lead to a greater number of logically correct responses. It is expected that this will result in an interaction between

instruction and problem type (Necessary and PS), since under possibility instructions no search is required for either type of problem.

3.   If participants search for alternatives: under possibility instructions there will be a greater number of endorsements of PW problems than Impossible problems, because although the first model does not support the conclusion, a search for alternatives will reveal an instance that will allow the reasoner to accept the conclusion. This will also result in an interaction between instruction and problem type (Impossible and PW), since under necessity instructions no search is required, and the conclusion will be rejected on the basis of the first model.

4.   If participants search for alternatives: PS problems under necessity instructions, and PW problems under possibility instructions, will take longer. This is because a search for alternatives is required in both cases to provide the correct conclusion.

5.   It is anticipated that the effects in predictions 2, 3 and 4 will be mediated by ability, in that higher ability participants will produce more correct responses where a search is required, because of their ability to carry out this search.

6.   On problems that require a search for counterexamples or alternatives (PS problems under necessity instructions and PW problems under possibility instructions), high ability participants will take longer. This is a cautious prediction as these effects may be confounded by general reasoning speed, such as high ability reasoners having faster processing skills.

2.2 Method

Design

This experimental study was carried out using a within-subjects design, when initially participants completed an AH4 cognitive ability test. This was followed by the syllogistic reasoning task, where participants were presented with one block of 32 randomised abstract syllogistic reasoning problems under necessity instructions, and one block of 32 abstract syllogistic reasoning problems under possibility instructions, the order of which was counterbalanced to minimise order effects.

Participants

A total of 60 undergraduate students at the University of Plymouth took part in the study, in return for either payment or course credit. The sample consisted of 26 males and 34 females with a mean age of 20 years, and they were all native English speakers. No participants were dyslexic, or had received formal training in logic.

Materials and procedure

Participants were run in groups of between 4 and 6 in a laboratory containing several computers. Each participant was seated at their own workstation, to avoid distraction.

Cognitive Ability Test

Initially participants completed Parts I and II of the AH4 cognitive ability test. This pencil and paper test, which was developed by Heim (1968) as a measure of general intelligence for use with a cross-section of the adult population, is widely used in studies of deductive reasoning (e.g. Newstead et al., 2004). Test-retest reliability has

been recorded at 0.919, with retesting after one month (Alexopoulos, 1997). The test consists of two 65 item sections, each of which is presented to participants in separate 10 minute sessions; and in both parts the test items become increasingly difficult. Correlations between scores on parts I and II have been reported to range between 0.60 and 0.81 (Alexopoulos, 1997; Heim, 1968). Part I is made up of verbal items concerning direction, verbal opposites, numerical series, verbal analogies, simple arithmetic computations, and synonyms, for example:

Part II contains diagrammatic items requiring judgments about analogies, sames, subtractions, series, and superimpositions, for example:

The test was administered in accordance with the test instructions, and question booklets and answer sheets were collected by the experimenter before moving on to the reasoning task.

Syllogistic reasoning task

A set of 32 abstract syllogisms was selected from a list of all 512 distinct problems based upon their endorsement rates under necessity instructions, recorded in previous research (Evans et al., 1999). The problem set was made up of 8 problems in each of four categories: Necessary, PS, Impossible, and PW. Problems were selected on the basis of endorsement rates reported under necessity instructions (Evans et al., 1999), as shown on the following page.

Necessary	≥ 70%
PS	≥ 70%
Impossible	≤ 30%
PW	≤ 30%

Examples of the problems, with their logical definitions, are shown in table 2.2.

Table 2.2
Examples and logical definitions for each of the four problem categories

Category	Example	Logical definition
Necessary	All of the A's are B's None of the C's are B's None of the A's are C's	The conclusion statement must be true given that the premises are true
PS	All of the A's are B's All of the B's are C's All of the C's are A's	The conclusion might be true given that the premises are true (frequently endorsed)
Impossible	All of the B's are A's None of the B's are C's All of the A's are C's	The conclusion statement cannot be true given that the premises are true
PW	Some of the B's are A's All of the C's are B's None of the C's are A's	The conclusion might be true given that the premises are true (less frequently endorsed)

In order to present a range of problem types and difficulty, each category had two problems from each of the four syllogistic figures; and within each figure one of the two syllogisms had a conclusion in direction a - c, and the other had a conclusion in direction c - a.   In all but once instance (Impossible problems; c - a direction) every problem type had one problem of each mood in a - c direction syllogisms, and one in c - a direction syllogisms. Randomly chosen letters of the alphabet (excluding I and O) were used for the premise terms. A complete set of the problems used in this experiment, together with figure, conclusion, and percentage endorsement rates previously recorded (Evans et al., 1999) under both necessity and possibility instructions is presented in appendix 2A. A list of all 512 problems and endorsement rates used in the selection process can be found in Evans et al. (1999).

A computer with a 15" monitor screen was used to present the problems, with a computer program written in visual basic. The keyboard was adapted to include *yes* and *no* keys, which were systematically counterbalanced, so that half the participants had the *yes* key on the left of the keyboard and the *no* key on the right, while the other half had these positions reversed.

The two sets of written task instructions modelled on the instructions used by Evans et al. (1999) were printed on A4 paper, included examples of the screen layouts, and were related to whether problems were being evaluated for either necessity or possibility correctness (see table 2.3 and table 2.4). A complete set of instructions is presented in appendix 2B and appendix 2C.

Table 2.3
Written instructions presented to participants (necessity)

Necessity instructions

The purpose of this experiment is to investigate how people solve logical reasoning problems. A number of problems will be presented on the screen one at a time. Each problem consists of two statements which describe the relationship between three letters, followed by a conclusion. Your task is to indicate whether the conclusion necessarily follows from the sentence that precedes it. A necessary conclusion is one that must be true given the truth of the preceding premises. Below are examples of the screen layouts.

Given that
    All of the M's are F's
    None of the D's are F's

press space bar to continue

Given that

    All of the M's are F's
    None of the D's are F's

Is it necessary that

    None of the D's are F's

First you will be shown two statements, and you should press the space bar to indicate your understanding of these. A conclusion will then be added, and your task is to decide whether this conclusion must be true. Using the keyboard, you should press 'yes' if you think the conclusion necessarily follows and 'no' if you think the conclusion does not necessarily follow. You will then be asked to press the space bar when you are ready to continue to the next problem.

Table 2.4
Written instructions presented to participants (possibility)

Possibility instructions

The purpose of this experiment is to investigate how people solve logical reasoning problems. A number of problems will be presented on the screen one at a time. Each problem consists of two statements which describe the relationship between three letters, followed by a conclusion. Your task is to indicate whether the conclusion possibly follows from the sentence that precedes it. A possible conclusion is one that could be true given the truth of the preceding premises. Below are examples of the screen layouts.

Given that
All of the P's and D's
All of the D's are T's
press space bar to continue

Given that
All of the P's are D's
All of the D's are T's
Is it possible that
All of the P's are T's

First you will be shown two statements, and you should press the space bar to indicate your understanding of these. A conclusion will then be added, and your task is to decide whether this conclusion is possible. Using the keyboard, you should press 'yes' if you think the conclusion is possible and 'no' if you think the conclusion is not possible. You will then be asked to press the space bar when you are ready to continue to the next problem.

The instructions were distributed (necessity or possibility) for the first block of problems and after a short reading period, participants were given the opportunity to ask questions on any points about which they were unclear. The participants were also told that they must ask the experimenter for the second set of instructions (necessity or possibility) as soon as a message appeared on the screen, and reminded that the start of each block there were two practice questions.

Participant responses, *yes* or *no*, were recorded by the program, together with the time taken to indicate understanding of the problem (screen 1) and the time taken to complete the reasoning process (screen 2). These were saved to disc.

2.3 Results

The AH4 test sheets were scored in accordance with the test instructions, when one mark was given for each correct answer. There was a significant positive correlation between Parts I and II (r = .49, p < .01), and in line with previous research (Newstead et al., 2004) the scores from both parts were totalled to give an overall general ability score for each participant. The observed mean for participants was 98.30 (SD = 12.90), which was slightly higher than the available norm of 96.36 (SD = 15.01) for university students (Heim, 1968). The sample was divided into high and low cognitive ability groups, on the basis of a median split on the AH4 test scores; cases below the median of 100.5 were classified as low ability and those above the median were classified as high ability.

All participants evaluated conclusions under both necessity instructions and possibility instructions. The first dependent variable was the mean percentage endorsement rates

for each problem category, i.e. the number of *yes* responses. A breakdown of endorsement rates into syllogistic figures and conclusion direction can be found in appendix 2D. The second dependent variable was the time course of the reasoning process; that is to say both premise processing and response times together. These were totalled for each problem type and instruction group to produce a mean reasoning time (in milliseconds)[10]. The approach taken was that which was adopted by Thompson et al. (2003) when the reasoning time was taken to be from presentation of the problem, to the generation of a response (in this case by hitting a key); and this approach will be adopted throughout the thesis. The results from the endorsement rate data are reported first; followed by the results from the reasoning time data. All ANOVA tables for experiment 1 are shown in appendix 2E.

2.3.1 Conclusion endorsement rates

The mean percentage endorsement rates for each of the four problem types are shown in table 2.5; broken down by instruction, problem type and ability. The cells for the low ability group and the high ability group each represent the mean percentage endorsement rates for responses from 30 participants. A breakdown of mean percentage endorsement rates for a - c and c - a conclusions can be found in appendix 2F.

---

[10] The pattern of responding was identical when two individual analyses were carried out, on both the time taken to understand the problem, and the time taken to complete the reasoning process.

Table 2.5
Mean percentage endorsement rates for all problem types (N = 60, SD in brackets)

	Necessary		PS		Impossible		PW	
Necessary								
Low	78	*(21.94)*	74	*(24.64)*	22	*(23.60)*	32	*(24.93)*
High	78	*(27.35)*	68	*(31.93)*	12	*(14.28)*	25	*(18.42)*
M	78	*(25.58)*	71	*(28.45)*	17	*(20.04)*	29	*(21.99)*
Possibility								
Low	85	*(22.67)*	82	*(22.90)*	22	*(22.19)*	35	*(20.26)*
High	84	*(22.73)*	79	*(23.06)*	28	*(28.69)*	46	*(25.67)*
M	84	*(22.51)*	81	*(22.83)*	25	*(25.58)*	41	*(23.57)*

Necessary and PS problems

The predictions were that there would be more endorsements of possibility than of necessity problems, more endorsements of Necessary problems than PS problems; and if reasoners searched for counterexamples there would be an interaction between instruction and problem type. It was also predicted that these results would be mediated by ability. A 2 (instruction) x 2 (problem type) x 2 (ability) mixed Analysis of Variance (ANOVA) test revealed a main effect of instruction [$F(1,58) = 9.02$, $p < .005$, $\eta_p^2 = .14$], reflecting higher endorsement rates under possibility instructions than necessity instructions. There was also a main effect of problem type [$F(1,58) = 7.14$, $p < .05$, $\eta_p^2 = .11$], whereby Necessary problems were more frequently endorsed than PS problems; however, the main effect of ability [$F(1,58) = .24$, $p = .63$] was not significant. The interaction between instruction and problem

type was not significant [$F(1,58) = 1.00$, $p = .32$, $\eta_p^2 = .02$], and there were no other significant interactions.

The main effect of instruction confirmed previous research (Evans et al., 1999), but more importantly it suggests that there was at least some understanding of the differences between necessity and possibility instructions. The differences in endorsement rates for Necessary and PS problems indicate that some participants were able to distinguish between problem types, but the lack of interaction with instruction or ability does not allow us to draw any other conclusions, particularly in terms of providing evidence to support the search for counterexamples.

Impossible and PW problems

It was predicted that there would be more endorsements of possibility than of necessity, more endorsements of PW problems than of Impossible problem; and if reasoners carried out a search for counterexamples there would be an interaction between instruction and problem type. It was also predicted that these results would be mediated by ability. A 2 (instruction) x 2 (problem type) x 2 (ability) mixed ANOVA test revealed a main effect of instruction [$F(1,58) = 12.08$, $p < .005$, $\eta_p^2 = .17$], reflecting higher endorsement rates under possibility instructions, and a main effect of problem type [$F(1,58) = 48.40$, $p < .001$, $\eta_p^2 = .46$], when PW problems were endorsed more frequently than Impossible problems. There was no main effect of ability [$F(1,58) = .02$, $p = .96$].
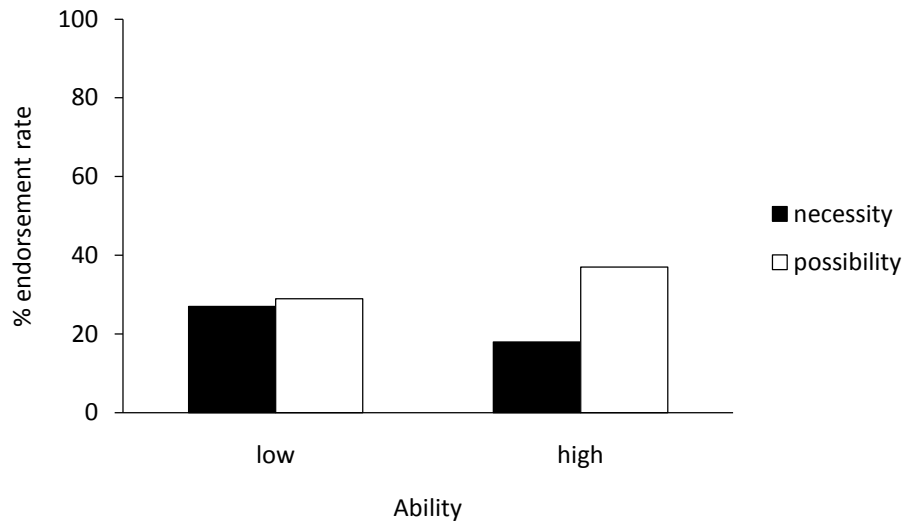
Figure 2.1. Mean percentage endorsement rates for both ability groups, under necessity and possibility instructions

There was a significant interaction between instruction and ability [$F(1,58) = 8.39$, $p <$ .005, $\eta_p^2 = .13$], which suggests that participants in the high ability group were more able to differentiate between necessity and possibility instructions (see figure 2.1).

This was confirmed by follow up within subjects *t*-tests, when there was found to be a significant difference of instruction for the high ability group [$t(29) = 4.18$, $p < .001$], but not for the low ability group [$t(29) = .45$, $p = .66$]. The interactions between instruction and problem type [$F(1,58) = 1.35$, $p = .25$] and problem type and ability [$F(1,58) = 1.30$, $p = .26$] were not significant.

Again, the main effect of instruction was consistent with previous research (Evans et al., 1999), and indicated that there was some understanding of the differences between necessity and possibility instructions. The main effect of problem type suggests that participants understood the differences between the two types of problem, but there is no evidence to suggest that they were more able to discriminate between PW and

Impossible problems under possibility instructions, than under necessity instruction. The lack of interaction between instruction and problem type failed to provide evidence for a search for alternatives on PW problems under possibility instructions.

2.3.2 Reasoning times

The mean reasoning times for all four types of problem are shown in table 2.6; broken down by instruction, problem type and ability. The cells for the low ability group and the high ability group each represent the mean reasoning times for responses from 30 participants, and are shown in milliseconds.

Table 2.6
Mean reasoning times (milliseconds) for all problem types (N = 60, SD in brackets)

	Necessary		PS		Impossible		PW	
Necessary								
Low	14558	*(7962)*	14574	*(8422)*	13449	*(5535)*	14562	*(8218)*
High	11918	*(4476)*	11890	*(5242)*	10641	*(3106)*	10858	*(3486)*
M	13238	*(6541)*	13232	*(7085)*	12045	*(4670)*	12710	*(6531)*
Possibility								
Low	14570	*(7323)*	14985	*(7470)*	14885	*(7771)*	14309	*(7178)*
High	12093	*(5242)*	11975	*(5058)*	11605	*(4647)*	12067	*(4813)*
M	13331	*(6436)*	13480	*(6504)*	13245	*(6560)*	13188	*(6164)*

Necessary and PS problems

It was predicted that reasoning times would be quicker under possibility instructions, and on Necessary problems; and if reasoners carried out a search for counterexamples this would result in an interaction between instruction and problem type. It was also predicted that these results would be mediated by ability. A 2 (instruction) x 2 (problem type) x 2 (ability) mixed ANOVA test was carried out, but there were no main

effects of instruction [$F(1,58) = .06$, $p = .81$], problem type [$F(1,58) = .07$, $p = .80$], or ability [$F(1,58) = 3.33$, $p = .07$], and no significant interactions[11]. There was no evidence to provide support for the prediction that problems requiring a search for counterexamples (PS under necessity instructions) would take longer, and the lack of main effect on instruction suggests that participants did not engage in more complex processing when asked to make judgements of necessity.

Impossible and PW problems

It was predicted that reasoning times would be quicker under necessity instructions; and on Impossible problems; also if reasoners were carrying out a search for counterexamples there would be an interaction between instruction and problem type. It was also predicted that these results would be mediated by ability. A 2 (instruction) x 2 (problem type) 2 x (ability) mixed ANOVA test revealed a main effect of ability [$F(1,58) = 5.09$, $p < .05$, $\eta_p^2 = .08$], suggesting that high ability participants were generally quicker reasoners than low ability participants. There were no main effects of instruction [$F(1,58) = 2.00$, $p = .16$] or problem type [$F(1,58) = 0.88$, $p < .35$] and no significant interactions[12]. Again, the data did not support the prediction that problems requiring a search for alternatives (PW under possibility instructions instruction) would take longer, and the lack of a main effect of instruction suggests that participants

---

[11] An equivalent analysis was repeated with log-transformed reasoning times, because of the number of outliers, but there were no significant effects.

[12] An equivalent analysis was repeated with log-transformed reasoning times, because of the number of outliers, but in line with the untransformed data, the only significant result was a main effect of ability [$F(1.58) = 4.50$, $p < .05$, $\eta_p^2 = .07$].

were not engaging in more complex processing when asked to make judgements of necessity.

2.4 Discussion

The primary aim of this experiment was to evaluate the claim that syllogistic reasoning involves a search for counterexamples, as proposed by the third stage of the mental model theory, and to investigate whether the likelihood of reasoners carrying out this search can be predicted by cognitive ability. This was done by first asking participants to complete an AH4 Cognitive Ability test, which enabled them to be categorised as either low ability or high ability. Following this they evaluated four different types of syllogistic reasoning problems (Necessary, PS, Impossible, and PW), under both necessity and possibility instructions. The analysis was directed at making comparisons between Necessary and PS problems, and Impossible and PW problems, under both types of instruction. Endorsement rates and reasoning times were recorded to detect where there was evidence of extra processing on items which required a search for additional models. The syllogisms were selected so that there was a range of problem types, figures, and direction of conclusion, to ensure as far as possible that the results were not due to biases such as the figural effect or conclusion direction bias.

It was predicted that there would be fewer endorsements of PS than Necessary problems under necessity instructions; together with an interaction between problem type and instruction, and that there would be a greater number of endorsements of PW than Impossible problems under possibility instructions, and an interaction between

problem type and instruction. It was also predicted that those problems requiring a search for counterexamples or alternatives would take longer in terms of problem processing times, because of the extra time required for the search process, and that participants in the high cognitive ability groups would be more accurate and would take proportionately extra time on problems requiring a search for counterexamples or alternatives.

However, despite the above predictions based upon previous research (Evans et al., 1999; Galotti et al., 1986; Newstead et al., 2004; Stupple & Ball, 2008; Thompson et al., 2003), and the assumptions of the mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991), analysis of the endorsement rate data failed to support the predictions which related specifically to the search for counterexamples.

There was a main effect of instruction for both sets of analysis carried out, which suggests that participants understood the differences between instructions of necessity and instructions of possibility, and the different types of problem. The interaction between instruction and ability on Impossible and PW problems suggests that it was the higher ability participants who had a better understanding of the differences in instruction types. The results implied that generally, when the initial model supported the conclusion (PS problems under necessity instructions), no further reasoning took place. Similarly, when the first model failed to support the conclusion, as was the case with PW problems under possibility instructions, reasoners did not look past the initial model to find one that validated the conclusion.

 With regard to the latency data, there was no evidence to support the predictions that a search for counterexamples of alternative models took place; and no evidence that participants understood the differences between necessity and possibility instructions, or of the different problem types. However, it may be that because people were not discriminating between necessity and possibility instructions, they were treating PS problems under necessity instructions and PS problems under possibility instructions in the same way. This may also be the case with PW problems under both types of instructions.

Given the lack of evidence to support the search for counterexamples or alternative models, it may well be that reasoners were settling for what is 'good enough' unless there was good reason to reject, modify or replace it, which is consistent with the satisficing principle of hypothetical thinking theory (Evans, 2007a). Hypothetical thinking theory (Evans, 2007a) defines the general characteristics of hypothetical thought such that, while the search for counterexamples in invalid problems is not spontaneous, more effortful processing can be encouraged by manipulating the task instructions. This is clear in the analysis of endorsement rates, which showed main effects of problem types for both Necessary and PS problems, and Impossible and PW problems.

Despite the lack of support for the search for counterexamples, the findings do suggest that at least some people are more sensitive to instructions, in terms of modifying and reducing their threshold for endorsing conclusions under possibility instructions. However given that previous research (i.e. Newstead et al., 2004) has reported ability

to be a good predictor of performance on syllogistic reasoning problems, we might also have expected these response rates to be mediated by ability on Necessary and PS problems, with a bigger difference for the high ability group. Nevertheless, there was an interaction between instruction and ability on endorsement rates for Impossible and PW problems, suggesting that people in the high ability group were generally more conservative when asked to evaluate conclusions under necessity instructions.

While it may be that a general increase in endorsement rates for problems presented under possibility instructions is due to response bias, such as a caution effect, this would seem unlikely given the interaction that was found. Instead it suggests that it is the higher ability participants who consciously modify their response threshold according to the instructions.

In conclusion, although there is no evidence to support the search for counterexamples in terms of it being the default setting; syllogisms are complex reasoning problems, and the first solution that comes to mind is perhaps more attractive than searching for alternatives. It may also be the case that due to the structural complexities of syllogisms, and the suggestion put forward by Grice (1975) that quantifiers which have multiple meanings may be hard because of linguistic complexity, participants were willing to put more effort into integrating the premises, leaving little in respect of cognitive resources to carry out a search for alternative models.

Furthermore, Bell and Johnson-Laird (1998) also suggest that the ability to search for counterexamples may well be influenced by the nature of the task; some paradigms such as syllogistic reasoning using complex arguments which involve understanding

the meaning of quantifiers and integration of premises, leaving few resources available for considering alternative representations of the premises. On the other hand, if the lack of support for the search for counterexamples was due to the structural and linguistic complexity of syllogisms, we might have expected there to have been some effects of cognitive ability, as suggested in the literature (Evans et al., 1983; Newstead et al., 2004; Newstead et al., 1992; Stanovich & West, 1998b; Torrens et al., 1999).

The following chapter presents two experiments adopting a similar methodology, which is applied to a range of transitive inference problems; to explore whether the findings in experiment 1 remain specific to abstract syllogistic reasoning tasks, or whether the absence of a search for counterexamples or alternatives is present in other paradigms.

78

Chapter 3

The search for counterexamples & alternative models in spatial transitive inference tasks

Experiment 1 revealed clear evidence to suggest that participants with higher cognitive ability were more able to modify their response when evaluating syllogistic reasoning problems, dependant on to whether the problems were presented under necessity instructions or under possibility instructions. This effect was found when reasoners were required to evaluate PW problems where the first model did not support the conclusion, and Impossible problems where no models supported the premises. There was however, no evidence to suggest that reasoners were searching for counterexamples or alternatives.

One possible explanation for the lack of evidence to support the search for counterexamples or alternatives, may be because reasoners settled for the first model that came to mind, or in other words a conclusion that is 'good enough', without making an effort to amend their initial conclusion. This view is consistent with hypothetical thinking theory (Evans, 2007a), which claims that when we think hypothetically, we

consider only one possible model at a time and use a heuristic or pragmatic process relevant to content and context given the goals of the task. Unless there is good reason to reject, modify or replace it, the theory claims that the decision is then accepted (satisficing principle).

However, an alternative explanation is that these findings are specific to syllogistic reasoning, predominantly because of problem complexity and linguistic structure. It is widely acknowledged that syllogisms are complicated reasoning problems with two premises which may or may not lead to a logically valid conclusion, and it is consistently reported in the literature (e.g. Johnson-Laird & Byrne, 1991) that some problems yield as few as 15% correct responses. The structure of syllogisms is such that the storage and manipulation of the 3 terms (A, B and C) is required, together with the application of two out of four quantifiers (all, none, some, or some …. not), so that a conclusion may be produced which may or may not include one of those already mentioned in the premises. This processes places a high demand on cognitive resources.

The motivation behind the use of the transitive inference paradigm in the two experiments reported in this chapter, is primarily to consider whether the absence of evidence to support the search for counterexamples or alternatives is unique to syllogistic reasoning, or whether it extends to other reasoning paradigms, which are not only less structurally complex, but are also not affected by linguistic ambiguity.

3.1 Introduction to experiments 2 and 3

Decisions based on our ability to make transitive inferences between two or more entities are part of everyday life. Reasoning behaviours are typically studied using 3-term series problems, and experimental studies generally require participants to infer the direction of a relation between two items (A and C), based on the relationship of each to the common term (B), when all differ along a single dimension such as length, size or spatial proximity. Take for example the following statements about Anne, Brian and Colin:

> Anne is taller than Brian
> Brian is taller than Colin

which invite the inference that:

> Anne is taller than Colin

Transitivity is a logical property of some but not all relations, and everyday relationships such is *next to*, are atransitive, in that the premises cannot be arranged on a linear scale. Consider therefore, the following premises James is *next to* Harry, Harry is *next to* Charlie, to which many reasoners would conclude that Harry *is in the middle of* James and Charlie; when the conclusion is in fact erroneous, since despite the fact that James may well be standing next to Charlie, they might be standing in a triangle. Yet another group of relations are intransitive, such as Angela is the mother of Bella, Bella is the mother of Catherine; because no inference can be made on the transitivity of the relationship between Angela and Catherine.

In the same way that some arrangements and combinations of the quantifiers used in syllogisms are easier than others, this is also true for transitive inference problems,

although the difficulty tends to be measured using reasoning times rather than error

rates. For instance, those which include transforming relational terms to their

opposites, or resolving negatives, have been found to take longer. Consider the

following two problems, this time using abstract terms:

B is better than C	and	C is worse than B
A is better than B		B is worse than A

Most people take a relatively short time to provide the correct response to the first of

these problems but other problems such as the second one, generally take longer, and

studies (i.e. Evans et al.,1993), have recorded longer reasoning times, with more

incorrect responses.

As reviewed in chapter 1, the two most popular theories in relational reasoning over

the past ten years, have been Imagery theories and Linguistic theory. Imagery theories

(De Soto et al., 1965; Huttenlocher, 1968) propose that individuals carry out transitive

inferences by constructing a visual image of the terms on a horizontal or vertical axis.

For example, given the relation A is better than B, would put A towards the good end of

the scale, and given the relation B is worse than A, would put B towards the bad end of

the scale; and individuals tend to either represent items on a vertical scale with good at

the top, or on a horizontal scale with good at the left.

The more linguistic explanation offered by Clark (1969), suggests that certain relational

terms are lexically marked, and because of this are harder to understand and

remember. Unmarked comparatives, such as taller than, can be used in a neutral way

to convey the relative degrees of the two items on a scale, but in contrast, marked

comparatives such as shorter than, can be used to refer only to items towards the shorter end of the scale. It is the unmarked terms which give their names to the scale; for example the dimension is called length rather than shortness, and Clark (1969) proposes that inferences should be easier with unmarked relational terms than marked relational terms, and given the statement A is worse than B, reasoners understand that both A and B are bad more quickly than their relative degrees of badness and the congruency of the statements, so if the statements both use the relation is better than, there is incongruity between the response when asked 'who is best'.

Although both Imagery theories and Linguistic theory offer plausible and testable accounts of what the mind computes, imagery or spatial array theories such as those proposed by DeSoto et al., (1965) and Huttenlocher (Huttenlocher, 1968) more readily transfer to the mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991), when terms are represented spatially rather than linguistically. The mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991) proposes that reasoners build an initial model representing the information given in the premises, from which they form a putative conclusion based on this information. If the initial model supports the conclusion, a search for counterexamples is carried out to find a model in which the premises hold, but in which the conclusion is not supported.

The search for counterexamples has not been widely explored using the transitive inference paradigm, perhaps because 3-term series problems are relatively simple compared to syllogistic reasoning; and researchers have concentrated more on how the terms are represented. However a small number of studies (Rauh et al., 2000;

Vandierendonck, 2000; Vandierendonck et al., 2004) have investigated transitive inferences, in relation to whether individuals searched for counterexamples using indeterminate multi-model problems. The results from these studies, led to the conclusion that people do not immediately construct all models in multi-model problems, but merely construct one integrated model, which is annotated in terms of there being a further model or models. This conclusion was reached by collecting latency data, under the premise that multi-model problems would take longer than single model problems if models were represented individually, or less time if they were represented simultaneously.

In order to extend the mental model theory which was initially developed within the syllogistic reasoning paradigm, so that it provided an explanation for experimental findings in transitive inference, Goodwin and Johnson-Laird (2005) introduced a revised model theory, with five main principles:

> *Iconicity*: The structure of the models is iconic in that it is independent from images, while still corresponding to the situation that is represented.

> *Emergent consequences*: The conclusion emerges from models that satisfy their premises.

> *Parsimony*: Individuals tend to construct only a single simple, model.

> *Strategic assembly:* Individuals develop different strategies which reflect the given problem - this assumption was based on the collection of 'think aloud' protocol; which showed that individuals try out various strategies when faced with solving transitive inference problems.

Complexity of integration: The difficulty of relational reasoning depends
on the number of entities that have to be integrated; therefore the ability
to reason correctly is affected by the number of models required.

This extended model theory has been well researched, and the *principle of iconicity* is supported by Knauff and Johnson-Laird (2002) who found that materials eliciting vivid imagery as opposed to spatial representations, served to impede rather than aid reasoning. Similarly, the *principle of strategic assembly* is supported by research (Goodwin & Johnson-Laird, 2006), where the collection of think-aloud protocol suggests that individuals develop different strategies reflecting the task in hand.

However, although Goodwin and Johnson-Laird's (2005) model based theory supports the claims that conclusions are emergent properties of models, it does not support the notion of a search for counterexamples as normative behaviour. Therefore, rather than searching for counterexamples, the theory employs the *satisficing principle*, where individuals meet the criteria for *adequacy* rather than seeking to identify an optimal solution[13]. The experiment reported in this chapter however, will use the original interpretation of the mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991), which includes a search for counterexamples component.

3.1.1 Aims and rationale for experiments 2 and 3

Anderson (1978) argued that the key to understanding our representational system is having an unambiguous understanding of both the content and format of the premises,

---

[13] The term satisfice was a originally coined by Simon (1983), who claims that this is because human beings lack the cognitive resources to maximise or consider all the relevant possibilities with sufficient precision.

and posited that the geometric semantics of the materials used in traditional transitive inference studies are unclear. Recently, some researchers (Knauff, Rauh, & Schlieder, 1995; Knauff et al., 1998; Rauh, Schlieder, & Strube, 1998; Rauh et al., 2000) have suggested that the materials used in older studies of transitive inference and relational reasoning, such as 'to the left-of' and 'to the right-of' may be open to ambiguous interpretation. They suggest that when attempting to represent the terms as visual images, individuals find the spatial relationships semantically unclear. For instance the premise A is to the left of B might be represented spatially as:

A   B
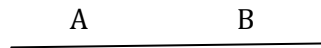
Or alternatively, where A *is* to the left of B, but is also *above* B.

A

                    B
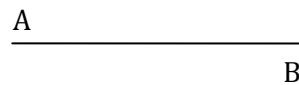
In a bid to overcome possible interpretational problems, such as those illustrated above, Knauff, Rauh, & Schlieder (1995) developed a set of materials with clear spatial relationships, taken from the area of Artificial Intelligence. The aim of their study was to look at differences in model formation, using the mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991) as a framework; and as expected, Knauff et al. (1995) found that there were preferred conclusions on indeterminate spatial transitive inference problems.

The materials used by Knauff et al. (1995) were founded on Allen's (1983) algorithmic interval-based calculus, which consists of 13 interval-based relations with clear start

points and end points, as shown together with their algorithmic symbol and natural language description in table 3.1.
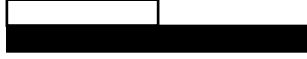
The 13 qualitative relations allow 144 possible[14] 3-term series compositions, of which 72 are determinate combinations, in that there is only a single possible logically correct response; and the other 72 combinations which yield an indeterminate conclusion fall into 4 classes: 42 problems with 3 possible solutions (models), 24 with 5 possible solutions (models), 3 problems with 9 possible solutions (models), and 3 problems with 13 possible solutions (models).
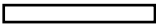
Spatial transitive inference problems similar to those developed by Knauff et al. (1995) were adopted and modified for experiment 2. These were presented to participants using what is a novel methodology within the paradigm of transitive inference, in terms of collecting responses under instructions of necessity and possibility; with the aim of further exploring the role of counterexample search and the search for alternative models.

The aim of experiments 2 and 3 was to test the claim made by Mental Model Theorists (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991), that reasoners search for counterexamples to falsify the first model that comes to mind on indeterminate spatial transitive inference problems. It will also look at whether reasoners search for alternative models on indeterminate problem structures when asked if a conclusion is possible, and where the first model does not support the conclusion.

---

[14] Although equals is included in Allen's calculus, and as such is a possible answer, this was not used in the composition of the 3-term series problems.

Table 3.1

The 13 qualitative interval relations according to Allen (1983), together with natural language description and algorithmic symbol

Relation symbol	Natural language description	Graphical example
X < Y	X lies to the left of Y	
X m Y	X touches Y at the left	
X o Y	X overlaps Y from the left	
X s Y	X lies left justified in Y	
X d Y	X is completely in Y	
X f Y	X lies right justified in Y	
X = Y	X equals Y	
X fi Y	X contains Y right justified	
X di Y	X surrounds Y	
X si Y	X contains Y left justified	
X oi Y	X overlaps Y from the right	
X mi Y	X touches Y at the right	
X > Y	X lies to the right of Y	

X = ▭   Y = ▬

When Knauff et al. (1995) carried out the data collection, a complete set of 3-term term series problems was presented to participants (144 problems, excluding equals), using natural language descriptions, and instructions to provide a conclusion. This enabled the collection of the percentage number of correct responses; a complete set of preferred conclusions for indeterminate problems, together with any subsequent models produced (also with percentage response rates), can be found in Knauff et al. (1995). This data base of preferred models facilitated the construction of PS and PW problems for experiments 2 and 3; where PS problems are those in which the first model supports the conclusion, and PW problems have an initial model that negates the conclusion. Examples of these are shown in table 3.2.

Table 3.2
Examples and logical definitions for each problem category

Problem category	Example
Necessary	The red line is surrounded by the blue line The blue line is to the left of the green line The red line is to the left of the green line
PS	The red line overlaps the green line from the left The green line touches the blue line at the right The red line overlaps the blue line from the right
Impossible	The red line is surrounded by the blue line The blue line is to the left of the green line The red line is to the right of the green line
PW	The red line overlaps the green line from the left The green line touches the blue line at the right The red line surrounds the blue line

When presented under necessity instructions the first problem category (Necessary) is a valid inference; both PS and PW are invalid as more than one conclusion can be drawn from the premises; and Impossible problems have a conclusion that is not possible. Similarly, under possibility instructions, the first problem category (Necessary) is both possible and necessary; this time both PS and PW problems are also possible; and the conclusion given for Impossible problems is again not possible.

In line with experiment 1, a measure of cognitive ability was also taken to look at the influence of individual differences in ability. While the findings to date would generally suggest that the search for counterexamples in transitive inference is not the default mechanism, it may well be that cognitive ability is a determinant of whether or not this mechanism is activated.

3.1.2 Predictions for experiments 2 and 3

The predictions relating to endorsement rates, reasoning times and ability are based on the general assumptions of the third stage of the mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991); and are similar to those more explicitly set out for experiment 1.

1. If participants search for counterexamples, under necessity instructions there will be fewer endorsements of PS problems than Necessary problems; and if participants search for alternatives, under possibility instructions there will be a greater number of endorsements of PW problems than Impossible problems.
2. On problems requiring a search for counterexamples or alternatives, participants will take longer.

3.   The results will be mediated by cognitive ability, in that those participants with higher cognitive ability will produce more correct responses, as well as being quicker reasoners.   Also participants with higher cognitive ability will take proportionately longer on problems requiring a search for counterexamples or alternatives, than participants with lower cognitive ability.

3.2 Pilot study for experiments 2 and 3

A short pilot study was carried out prior to experiment 2, to identify a set of easily understood interval relations with clear geometric semantics, for the relational inference task, which with training avoided interpretational ambiguity. A number of studies (Knauff, 1999; Knauff et al., 1995; Rauh et al., 2000) have used all 13 interval relations, however some of these relations may be less easily understood as they are terms that are not frequently used in everyday English language (e.g. those using the terms left justified and right justified).

Participants

The participants, who were run individually, were 6 undergraduate and postgraduate volunteers from the University of Plymouth; they were all native English speakers, and none had formal training in logic. None of the participants were dyslexic.

Procedure

The 13 interval relations used by Knauff et al. (1995) referred to in the introduction, together with their natural language description, were presented to participants on an A4 sheet of paper (see table 3.3).   The word 'line' was used instead of 'interval' as this

is more commonly used in the English language. After a short reading period, participants were asked for qualitative feedback regarding any interpretational problems they encountered; more specifically if they felt that any of the intervals relations were ambiguous or difficult to understand.

Table 3.3
The 13 qualitative interval relations

1.	The red line lies to the left of the blue line	
2.	The red line touches the blue line at the left	
3.	The red line overlaps the blue line from the left	
4.	The red line lies left justified in the blue line	
5.	The red line is completely in the blue line	
6.	The red line lies right justified in the blue line	
7.	The red line contains the blue line right justified	
8.	The red line surrounds the blue line	
9.	The red line contains the blue line left justified	
10.	The red line overlaps the blue line from the right	
11.	The red line touches the blue line at the right	
12.	The red line lies to the right of the blue line	
13.	The red line equals the blue line	

Qualitative feedback and discussion

There was a general consensus that the natural language definitions which included the words *left justified*, or *right justified*, as in interval relations 4, 6, 7 and 9 (table 3.4) were infrequently used in everyday language, and may be less clearly understood; and also that the term *the red line is completely within the blue line* (interval relation 5) might be best defined using the words *is surrounded by*.

Interval relations 4, 6, 7 and 9 were therefore removed from the list of materials for the main study, and interval relation number 5 was amended to read *the red line is surrounded by the blue line* instead of *the red line is completely in the blue line*.

3.3 Method for experiment 2

Design

This experimental study was carried out using a within-subjects design, when initially participants completed an AH4 cognitive ability test. This was followed by the relational inference task, which consisted of three phases: the definition phase, the learning and practice phase, and finally the inference phase where participants were presented with one block of 32 randomised relational inference problems under necessity instructions, and one block of 32 relational inference problems under possibility instructions, the order of which was counterbalanced to minimize order effects.

Participants

A total of 60 undergraduate students from the University of Plymouth took part in the study, in return for either payment or course credit. The sample consisted of 19 males and 41 females with a mean age of 25 years, and they were native English speakers. No participants were dyslexic, had received formal training in logic, or were colour blind. The exclusion of colour-blind participants is particularly relevant, since participants needed to be able to distinguish between red, blue and green lines, and it has been reported that approximately 99% of people suffering from colour blindness (8% - 12% of males of European origin and about one-half of 1% of females) have problems in distinguishing between red and green.

Materials and procedure

The procedure adopted was broadly similar to the experiment using syllogisms, which was reported in chapter 2. Participants were run in groups of between 4 and 7 in a laboratory containing several computers. Each participant was seated at their own workstation, to avoid distraction.

Cognitive Ability Test

Initially, as a measure of ability, participants completed Parts I and II of the AH4 Test of Cognitive Ability (Heim, 1968), which was administered in accordance with the test instructions and followed the procedure used in experiment 1. Question booklets and answers sheets were collected by the experimenter before moving on to the relational inference task.

Relational inference task:

This task consisted of three phases, and used the materials identified in the pilot study. The first was *the definition phase*, was when pictures depicting the location of 9 red and a blue interval relations were presented to participants, along with a short commentary about the location of the beginnings and endings of these intervals; *in the learning and practice phase*, participants were tested on their understanding of these interval relations; and the *inference phase* involved presenting participants with one block of randomised 32 relational inference problems under necessity instructions, and one block of randomised 32 relational inference problems under possibility instructions.

Table 3.4
The 9 interval relations used for the study, together with an explanation of the semantics relating to the ordering of starting points and ending points

Semantic description	Graphical definition
The red line lies to the left of the blue line	
The red line touches the blue line at the left	
The red line overlaps the blue line from the left	
The red line is surrounded by the blue line	
The red line surrounds the blue line	
The red line overlaps the blue line from the right	
The red line touches the blue line at the right	
The red line lies to the right of the blue line	
The red line equals the blue line	

Definition phase: Participants read descriptions of the semantic relationships between a red and a blue line, in order to familiarise themselves with the terminology that was going to be used in the inference task. These descriptions were printed on A4 paper, when each depiction was accompanied by a graphical representation that matched the relationship between the two lines (see table 3.4). After a period of 2 minutes this information was removed.

Learning and practice phase: At the start of this phase, participants were given an A4 sheet of paper showing the semantic description of the same 9 interval relations used in the definition phase; these were numbered 1 – 9. This was to test how well participants understood the terminology in the relational inference phase. For example:

Graphical definition:　　　5.

Participants were also given a list of the 9 graphical definitions in randomized order and instructed to write the number of the interval relation that correctly depicted the description in the box at the side. An example is shown below, and a full set of the semantic descriptions and definitions can be found in appendix 3A and appendix 3B.

　　Semantic description: The red line is surrounded by the blue line

After confirmation of his/her final choices, each participant was told whether the choices were correct or incorrect. The learning and practice criterion was accomplished when participants had worked through two such lists consecutively, without error; and the graphical definitions were randomized in 8 different ways, so

that even if participants took several attempts to complete the learning and practice phase, the order of the graphical definitions was different.

Inference task: The set of 32 3-term series problems and conclusions was constructed using the 8 (9 including 'equals') interval relations which had been identified in the pilot phase and which participants had become familiar with during the learning and practice phase. The 8 interval combinations in each category (Necessary, PS, Impossible, and PW) were selected from the correct and preferred responses to 64 possible combinations of these interval relations, using the following criteria:

> *Necessary problems*: the percentage correct responses were rank ordered, from which the top 8 were selected, with mean endorsement rates of between 90% and 97%.

> *PS problems*: The correct response percentages for multiple model (indeterminate) problems were rank ordered by the most common response, from which the top 8 were selected, with mean endorsement rates of between 63% and 91%.

> *Impossible problems*: the conclusions for the Necessary problems were reversed to provide 3-term Impossible problems.

> *PW problems*: the same problem structures were used as for PS problems, but one of the other less preferred possible responses was used, as a conclusion for evaluation. The selection of these was arbitrary as the percentage endorsement rates for the other options were not given individually

A computer with a 15" monitor screen was used to present the problems, with the computer program. The keyboard was adapted to include *yes* and *no* keys, which were

systematically counterbalanced, so that half the participants had the *yes* key on the left of the keyboard and the *no* key on the right, while the other half had these positions reversed.

The two sets of written task instructions which included examples of the screen layout, were printed on A4 paper, and were similar to those used in experiment 1. These related to whether problems were being evaluated for either necessity correctness, or possibility correctness. Examples of the screen layouts are shown in table 3.5, and a complete set of instructions is presented in appendix 3C and appendix 3D.

The instructions were distributed (necessity or possibility) for the first block of problems and after a short reading period, participants were given the opportunity to ask questions on any points that they were less clear about. The participants were also told that they should ask the experimenter for the second set of instructions (necessity or possibility) as soon as a message appeared on the screen, and reminded that at the start of each block there were two practice questions. Participant responses, *yes* or *no*, were recorded by the program, together with the time taken to indicate understanding of the problem (screen 1) and the time taken to complete the reasoning process (screen 2). These were saved to disc.

Table 3.5
Screen layouts included in task instructions

Screen 1

> Given that:
>   The red line surrounds the green line
>   The blue line lies to the left of the green line

Screen 2

> Given that:
>   The red line surrounds the green line
>   The blue line lies to the left of the green line
> *Is it necessary that*
>   The red line lies to the left of the blue line

Screen 1

> Given that
>   The red line surrounds the green line
>   The green line touches the blue line at the left

Screen 2

> Given that:
>   The red line surrounds the green line
>   The blue line lies to the left of the green line
> *Is it possible that*
>   The red line lies to the left of the blue line

3.4 Results for experiment 2

The AH4 test sheets were scored in accordance with the test instructions, when one mark was given for each correct answer. There was a significant positive correlation between Parts I and II ($r = .57$, $p < .01$), and in line with previous research (Newstead et al., 2004) the scores from both parts were totalled to give an overall general ability

score for each participant. The observed mean for participants was 89.68 (SD = 15.54), which was considerably lower than the mean reported for experiment 1 (M = 98.30; SD = 12.90) and for the available norm of 96.36 (SD = 15.01) for university students (Heim, 1968). The sample was divided into high and low cognitive ability groups, on the basis of a median split on the AH4 test scores; cases below the median of 91 were classified as low ability and those above the median were classified as high ability. None of the participants recorded a score of 91. The median was also considerably lower than for experiment 1 (median = 100.5).

All participants evaluated conclusions under both necessity instructions and possibility instructions. The first dependent variable was the mean percentage endorsement rates for each problem category, i.e. the number of *yes* responses. The second dependent variable was the time course of the reasoning process; that is to say both premise processing and response times together. These were totalled for each problem type and instruction group to produce a mean reasoning time (in milliseconds). The results from the endorsement rate data are reported first; followed by the results from the reasoning time data. All ANOVA tables for experiment 2 are shown in appendix 3E

3.4.1 Conclusion endorsement rates

The mean percentage endorsement rates for all four types of problem are shown in table 3.6, broken down by instruction, problem type and ability. The cells for the low ability group and the high ability group represent the mean percentage endorsement rates for the responses from 30 participants.

Table 3.6
Mean percentage endorsement rates for experiment 2, on all problem types (N = 60, SD in brackets)

	Necessary		PS		Impossible		PW	
Necessary								
Low	71	*(20.00)*	61	*(22.35)*	32	*(22.43)*	35	*(23.92)*
High	77	*(19.16)*	66	*(21.76)*	26	*(20.86)*	31	*(30.04)*
M	74	*(19.65)*	64	*(21.99)*	29	*(21.68)*	33	*(27.02)*
Possibility								
Low	74	*(20.06)*	63	*(21.67)*	28	*(19.68)*	59	*(23.48)*
High	73	*(18.10)*	71	*(25.29)*	28	*(18.26)*	68	*(28.76)*
M	73	*(18.95)*	67	*(23.68)*	28	*(18.82)*	64	*(26.47)*

Necessary and PS problems

A 2 (instruction) x 2 (problem type) x 2 (ability) mixed ANOVA test revealed a main effect of problem type [$F(1,58) = 11.93$, $p < .001$, $\eta_p^2 = .17$], reflecting higher endorsement rates for Necessary problems than for PS problems. The main effects of instruction [$F(1,58) = .29$, $p = .59$], and ability [$F(1,58) = 1.17$, $p = .28$], were not significant, and there were no significant interactions.

The main effect of problem type confirms that some participants were able to discriminate between problem structures, but there was no evidence to suggest that participants understood the differences between instruction types, and the lack of interaction between instruction and problem type failed to support a search for counterexamples. These results were consistent with experiment 1.

Impossible and PW problems

A 2 (instruction) x 2 (problem type) x 2 (ability) mixed ANOVA test revealed a main effect of instruction [$F(1,58) = 22.75, p < .001, \eta_p^2 = .28$], and problem type [$F(1,58) = 33.79, p < .001, \eta_p^2 = .37$]. This reflected higher endorsement rates under possibility instructions than endorsement rates under necessity instructions, and more endorsements of PW problems than Impossible problems. Both of these results are consistent with experiment 1, and suggest that participants had an understanding between instructions, and problem types. The main effect of ability was not significant [$F(1,58) = .01, p = .94$].
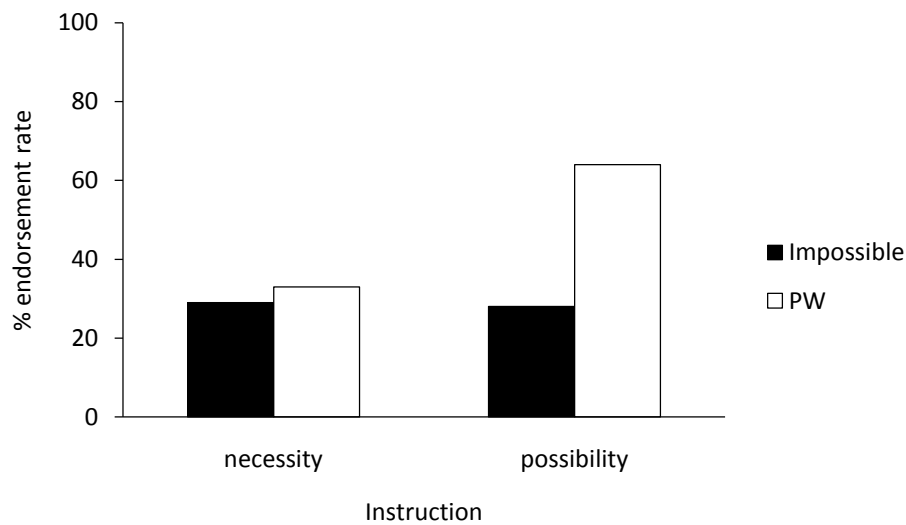
Figure 3.1. Mean percentage endorsement rates for experiment 2, on Impossible and PW problems under necessity and possibility instructions

There was a significant interaction between instruction and problem type [$F(1,58) = 32.44, p < .001, \eta_p^2 = .36$], which supported the search for alternatives, when participants went past the first model on PW problems under possibility instructions (see figure 3.1). This interaction was not found in experiment 1. A repeated measures

t-test confirmed that there was a significant difference between Impossible and PW problems under possibility instructions, but the difference was not significant when problems were presented under instructions of necessity [$t(59) = .99, p = .32$].

3.4.2 Reasoning times

The mean percentage reasoning times for all types of problem are shown in table 3.7, broken down by instruction, problem type and ability. The cells for the low ability group and the high ability group represent the mean reasoning times for the responses from 30 participants, and are shown in milliseconds.

Table 3.7
Mean reasoning times in milliseconds for experiment 2, on all problem types (N = 60, SD in brackets)

	Necessary		PS		Impossible		PW	
Necessary								
Low	20891	*(10024)*	20843	*(8233)*	22845	*(12003)*	21300	*(9868)*
High	16731	*(5160)*	19935	*(6421)*	19479	*(6163)*	19015	*(5722)*
M	18811	*(8178)*	20389	*(7334)*	21162	*(9611)*	20158	*(8080)*
Possibility								
Low	21069	*(9305)*	19666	*(7506)*	20331	*(9866)*	20228	*(10999)*
High	20130	*(5247)*	21100	*(6040)*	19834	*(5394)*	19701	*(4527)*
M	20600	*(7504)*	20383	*(6793)*	20083	*(7887)*	19964	*(8343)*

Necessary and PS problems

A 2 (instruction) x 2 (problem type) x 2 (ability) mixed ANOVA test was carried out,

but there were no main effects of instruction [$F(1,58) = .56$, $p = .46$], problem type [$F(1,58) = 1.59$, $p = .21$], or ability [$F(1,58) = .73$, $p = .40$][15].

There was however a significant interaction between instruction and problem type [$F(1,58) = 4.10$, $p < .05$, $\eta_p^2 = .07$], when under necessity instructions, participants took longer on PS problems than on Necessary problems (see figure 3.2) supporting the search for counterexamples. Follow up repeated measures *t*-tests confirmed there was a significant difference between reasoning times on Necessary and PS problems under necessity instructions [$t(59) = 2.14$, $p < .05$], but the difference was not significant under possibility instruction [$t(59) = .31$, $p = .76$].
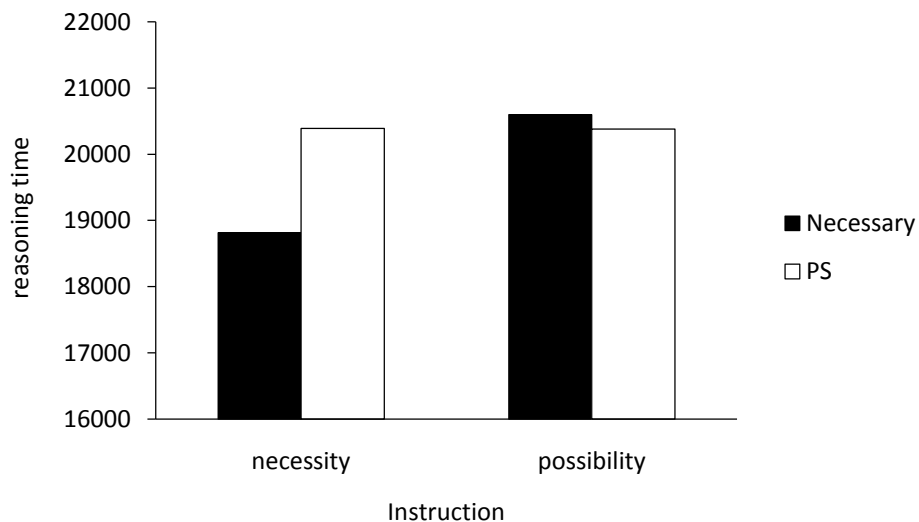
Figure 3.2. Mean reasoning times (in milliseconds) for experiment 2, on Necessary and PS problems under necessity and possibility instructions

---

[15] An equivalent analysis was repeated with log-transformed reasoning times for Necessary and PS inferences, because of the number of outliers, but there were no significant main effects.

There was also a significant interaction between problem type and ability [$F(1,58)$ = 6.77, $p < .05$, η_p^2 = .11], which is illustrated in figure 3.3. The high ability group took significantly longer on PS problems, which is confirmed by a repeated measures t-test [$t(29)$ = 2.75, $p <.01$]; while the low ability group took less time on PS problems [$t(29)$ = .94, $p = .35$].
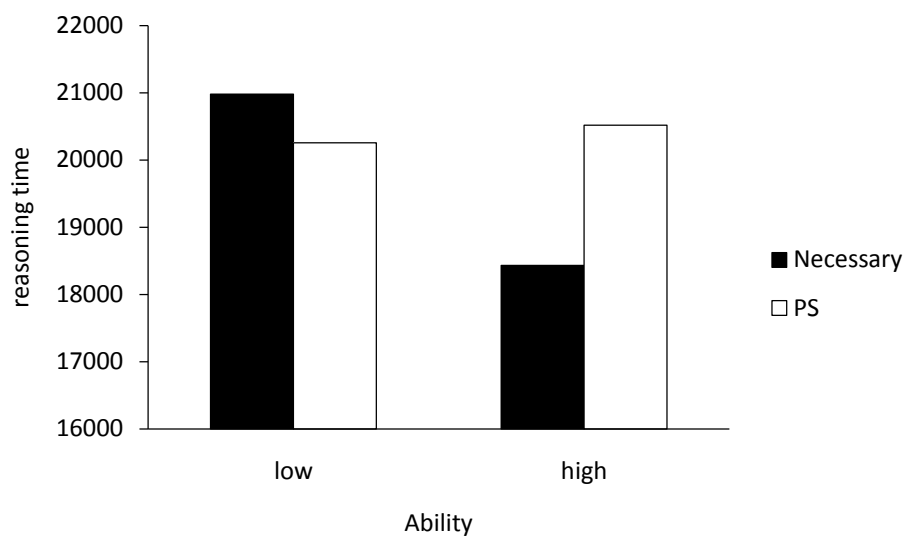
Figure 3.3. Mean reasoning times (in milliseconds) for experiment 2, on Necessary and PS problems for low and high ability groups

The increased reasoning times recorded by the high ability group may be because of a search for counterexamples; although no firm conclusions can be drawn as this was across both types of instruction.

Impossible and PW problems

A 2 (instruction) x 2 (problem type) x 2 (ability) mixed ANOVA test was carried out, but there were no main effects of instruction [$F(1,58)$ = .19, $p = .66$], problem type [$F(1,58)$

= .79, *p* = .38], or ability [*F*(1,58) = .31, *p* = .26]; and there were no significant interactions[16].

3.5 Discussion for experiment 2

The purpose of this experiment was to explore whether the lack of evidence to support the search for counterexamples and alternatives in experiment 1 was because participants were merely satisficing and accepting a 'good enough' conclusion without seeking the optimum solution. Alternatively the lack of evidence may be because the results were specific to syllogistic reasoning, when problem complexity and structure produced results that are uncharacteristic of deductive reasoning processes on other paradigms.

Although the evidence is limited, there is some support from experiment 2 for the search for alternative models in relational reasoning, as envisaged by the third stage of the mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1994), specifically in terms of the endorsement rate data, there was evidence of a search for alternative models on PW problems under instructions of necessity. These findings are in sharp contrast to experiment 1, where there was no evidence to support either the search for counterexamples, or the search for alternatives. However, in line with experiment 1, the endorsement rates for PS problems under necessity instruction did not provide evidence of a search for counterexamples.

---

[16] An equivalent analysis was repeated with log-transformed reasoning times for Impossible and PW inferences, because of the number of outliers, but there were no significant effects.

The endorsement rate data also suggests that participants understood the differences between Necessary, PS, Impossible and PW relational inference problems, because there were significant differences in endorsement rates between Necessary and PS problems, and Impossible and PW problems. The data also suggests that participants understood the difference between being required to make judgements of necessity and judgements of possibility. This support was in line with experiment 1, which in turn confirmed previous research (Evans et al., 1999), and it was particularly important to explore and establish this in experiment 2, because this is a novel experiment in relational inference studies.

The key findings to support the search for counterexamples were provided by the latency data; when reasoners spent extra time on PS problems under necessity instructions, as would be expected if a search was carried out for counterexamples to correctly invalidate the given conclusion. This was further clarified by the interaction between problem type and ability, where there was evidence that it was the high ability group who were more likely to search for counterexamples. Nevertheless, these results were not supported to the endorsement rate data, but this may be because despite a search being carried out, this search was unsuccessful; despite it being reasonable to expect that at least the higher ability group were able to do so.

One possible reason for the lack of effects on endorsement rates for PS problems under necessity instructions, is that all (low and high ability) participants were employing the *satisficing principle* on PS problems, where participants met the criteria for adequacy rather than looking for the optimum solution (Goodwin & Johnson-Laird, 2005). A

problem with this explanation is that the satisficing principle claims human beings lack the cognitive resources to consider all relevant possibilities; when clearly there is some evidence to suggest that the participants had, and were using, additional cognitive resources to search for alternative models on PW problems under possibility instructions. Furthermore, the findings are a poor fit with Goodwin and Johnson-Laird's (2005) principle of parsimony, when the suggestion is that individuals construct only a single simple model; because if this was the case, again there would have been no evidence, from either endorsement rates or latencies, of reasoners carrying out a search for counterexamples or alternatives.

The results are also inconsistent with previous research (Rauh et al., 2000; Vandierendonck, 2000; Vandierendonck et al., 2004), claiming that reasoners do not search for counterexamples on indeterminate problems; based on findings that there were no differences in latencies between multi-model problems and single model problems. While with the latency data it may be fair to conclude from experiment 2 that under possibility instructions, participants did not take longer on PW problems therefore they did not search for alternative models, there was evidence of a search for counterexamples on PS problems suggesting that reasoners did not merely annotate one integrated model as posited by Vandierendonck et al. (2004.)

It is important to note that the observed mean for experiment 2 (M = 89.68) was considerably lower than the available norm of 96.36 (SD = 15.01) and the mean for experiment 1 (M = 98.30; SD = 12.90). Furthermore, it is interesting that previous research (Knauff et al., 1995) reported very high percentage endorsement rates on

determinate problems presented under instructions of necessity (between 90% and 97%), from which the Necessary problems were taken, but the mean percentage endorsement rates recorded in experiment 2 were much lower at 71% for the low ability group and 77% for the high ability group. Therefore, we should not discount the fact that the results may have been influenced simply by the sample being less able to search for counterexamples, but nevertheless able to discriminate between necessity and possibility instructions, and problem types.

Other explanations which may be responsible for the lack of conclusive findings to support the search for counterexamples, and the disparity between the endorsement rate and latency results, is that the success of the learning and practice phase (which includes training) was limited. Anderson (1978) argued that reasoners need to have a clear understanding of the content and format of the spatial compositions in a problem, and if this is not achieved it leads to what Evans (1972) refers to as interpretational problems, where the results deviate from the researchers expectations because reasoners have a different understanding of the semantics of the premises from that of the experimenter, or at least a lack of clarity in the terminology used. Furthermore, Kruger and Dunning (1999) argued that individuals who are unfamiliar with a particular domain may lack the skills to prevent poor performance, and found that individuals, who were trained with the necessary skills to carry out deductive reasoning tasks, increased the accuracy of judgements on a number of Wason selection tasks. Prowse, Turner and Thompson (2009) reported also, that immediate feedback was more effective in remedying misunderstandings found in syllogistic reasoning

tasks, leading to improved performance on a number of determinate and indeterminate syllogistic reasoning problems.

In view of the somewhat inconclusive evidence to support the search for counterexamples or alternatives, concerns about the lower than average ability rates, and the effectiveness of the learning and practice phase which was responsible for training participants to a similar level of understanding of the terms used, it was decided to run a second transitive inference experiment. This was also influenced by observations that during the pencil and paper learning and practice phase in experiment 2, a number of participants did not appear to engage with the task, resulting in a considerable delay for those participants who provided the correct response at the first presentation.

A further consideration in the planning of a second transitive inference experiment, was that although the procedure for our training phase was loosely based on previous work by Knauff (1995), the feedback in experiment 2 was given verbally at the end of each presentation of 2 complete sets of line relationships. On the other hand, Knauff (1995) carried out the testing phase electronically, with immediate feedback after each individual line relationship, and this may have had a bearing on the results.

The next section of this chapter will discuss relevant research on learning and practice, and the need for participants to have a clear understanding of the materials and instructions. Following this experiment 3 will be reported, which is broadly similar to experiment 2, but with a modified learning and practice phase, and it is predicted that this will facilitate the search for counterexamples and alternatives.

3.6 Introduction and rationale for experiment 3

The reasoning times from experiment 2 revealed evidence to suggest that under instructions of necessity, individuals carried out a search for counterexamples on indeterminate problems that are frequently erroneously endorsed as being valid (PS). There was also evidence from the endorsement rate data that under possibility instructions, individuals searched for alternative models when the first model did not support the conclusion (PW). However as discussed in the previous section, the results are inconclusive, and the observed reasoning behaviours may be due to an ineffective learning and practice phase.

Early research into human performance (Fitts & Posner, 1967) suggests that the first stage of acquiring a cognitive skill, thus permitting the learner to generate the desired normative behaviour, is the cognitive stage; during which learners rehearse the information required for the execution of the behaviour. It is at this stage where the learner receives instruction and information about the behaviour in order to be able to perform a cognitive task, which in this study was a spatial relational inference task. Anderson (1982) further breaks this down into the declarative stage and the interpretive stage. First they claim that the initial knowledge is integrated into the system by being encoded declaratively, in other words understanding what should be done rather than how to do it. In order to successfully translate this into normative behaviour, the knowledge then needs to move into the interpretive stage which is when the information is learned. Within the context of experiment 2, it may be that although participants successfully encoded the materials declaratively, the interpretive stage was not successfully accomplished by the learning and practice phase.

Studies looking at the impact of training and instruction on both logical and statistical reasoning tasks (Fong, Krantz, & Nisbett, 1986; Lehman, Lempert, & Nisbett, 1988; Nisbett & Ross, 1980); report significant improvements in the reasoning skills of individuals after training and practice. For example, Fong et al. (1986) found that statistical and conceptual training improved statistical reasoning performance on a variety of reasoning problems, and also that participants were able to successfully apply the training they had been given to a number of problems in different domains. It is not clear if the results reported by Fong et al. (1986) took into account whether all participants understood the concepts involved; and one of the key concerns with experiment 2 is whether all participants reached a similar level of understanding, in order to be able to apply their understanding of the terms to the reasoning problems.

More recently, Neilens (2004) explored the role of individual differences in mediating the effectiveness of training; and reported that participants of higher ability were more able to understand and apply the principles they had been taught to a number of reasoning tasks. Again there is no indication whether all participants reached a similar level of understanding, but this point is relevant given the concerns about the mean AH4 scores which were recorded.

Clearly it is important for all participants to understand the terms used, and to learn the terms, in order to understand how to apply this knowledge. As discussed at the end of experiment 2, the results were inconclusive, and there was disparity between the results from the endorsement rates and the latencies. Therefore, to encourage normative responding on the 3-term series spatial inference problems, the learning and

practice phase was changed to reflect the electronic presentation used by Knauff (1995), which included immediate feedback on the correctness of the judgement made by the participants (Prowse, 2009) .

3.6.1 Predictions for experiment 3

The predictions are based on the general assumptions of the third stage of the mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991), and the success of the revised learning and practice phase. These are set out in experiment 2, and more explicitly explained in experiment 1.

3.7 Design and method for experiment 3

The design of experiment 3 was the same as experiment 2, in that it was an experimental study employing a within-subjects design. Initially participants completed an AH4 cognitive ability test, followed by the relational inference task, which consisted of three phases: the definition phase, the learning and practice phase. The final phase was the inference phase where participants were presented with one block of 32 relational inference problems under necessity instructions, and one block of 32 relational inference problems under possibility instructions, the order of which was counterbalanced to minimize order effects.

Participants

A total of 60 undergraduate students from the University of Plymouth took part in the study, in return for either payment or course credit. The sample consisted of 14 males and 46 females with a mean age of 24 years, and they were all native English speakers.

No participants were dyslexic, had received formal training in logic, or were colour blind (see experiment 2).

Materials and procedure for experiment 3

The procedure was similar to experiment 2, but with changes to the learning and practice phase in order to maximise the effectiveness. Participants were run in groups of between 2 and 5 in a laboratory containing several computers. Each participant was seated at their own workstation, to avoid distraction.

Cognitive Ability Test

Initially, as a measure of ability, participants completed Parts I and II of the AH4 Test of Cognitive Ability (Heim, 1968), which was administered in accordance with the test instructions and followed the procedure used in experiments 1. Question booklets and answers sheets were collected by the experimenter before moving on to the inference task.

Relational inference task

Participants then went on to complete the three phases of the main task: *the definition phase* when the graphical relationships and semantic explanation between two lines were defined; the revised *learning and practice phase* when participants were tested for their understanding of these terms, and the *inference phase*, where a number of inference problems were presented.

Definition phase: This phase remained unchanged, in that participants read descriptions of the semantic relationships between a red and a blue line together with a

graphical description; this was to familiarise them with the terminology that was going to be used in the inference task. After a period of 2 minutes the graphical definitions were removed, and the participants moved on to the improved *learning and practice phase*, followed by the *inference phase*, both of which were presented and executed on a computer with a 15" monitor screen, using the computer program.

Learning and practice phase: This phase was improved from that used in experiment 2 to assist learning, and to practice the relational terms to facilitate the inference task. It consisted of blocks of trials of all 9 relations, in which participants were presented with a one-sentence description of the red and blue line. Participants were given an A4 sheet of paper showing the semantic descriptions of the same 9 interval relations used in the definition phase, numbered 1 – 9, and required to select the appropriate number from the list of graphical representations, by pressing the associated number at the top of the computer keyboard. Feedback was given by computer program as to whether their choice was correct or false. Participants were told that there were no time restrictions; and if the correct response was provided the program would move onto the next graphical definition, but if the response was incorrect an error message would appear on screen and the program would spool back to the first randomized problem. See table 3.8 for examples of the screenshots. The graphical relations were randomised within each trial, and the learning and practice criterion was accomplished as soon as 2 consecutive complete sets (9 definitions) were correctly identified.

Table 3.8
Examples of the screenshots used in the learning and practice phase for experiment 3

The red line is surrounded by the blue line
1   2   3   4   5   6   7   8   9

False
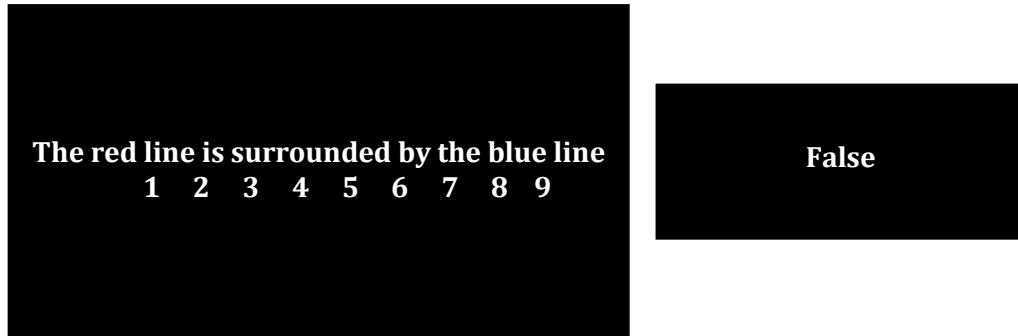
Inference task: The set of 32 3-term series problems and conclusions were the same as the ones used in experiment 2, and identified in the pilot study. The problems consisted of the 8 (9 including equals) interval relations that participants had become familiar with during the learning and practice and training phase. The two sets of task instructions and the procedure were also the same as those used in experiment 2.

The computer keyboard was adapted to include *yes* and *no* keys for the inference phase, which were systematically counterbalanced, so that half the participants had the *yes* key on the left of the keyboard and the *no* key on the right, while the other half had these positions reversed. Participant responses, *yes* or *no*, were recorded by the program, together with the time taken to indicate understanding of the problem and the time taken to complete the reasoning process. These were saved to disc.

3.8 Results for experiment 3

The AH4 test sheets were scored in accordance with the test instructions, when one mark was given for each correct answer. There was a significant positive correlation between Parts I and II ($r = .44$, $p < .01$), and in line with previous research (Newstead et al., 2004) the scores from both parts were totalled to give an overall general ability score for each participant. The observed mean for participants was 95.56 ($SD = 14.28$), which was similar to the available norm of 96.36 ($SD = 15.01$) for university students (Heim, 1968). This was considerably lower than that reported for experiment 1 (102.78), but higher than the mean reported for experiment 2 (89.68). The sample was divided into high and low cognitive ability groups, on the basis of a median split on the AH4 test scores; cases below the median of 95 were classified as low ability and those above the median were classified as high ability. None of the participants recorded a score of 95. The median was lower than recorded for experiment 2 (median = 91), and considerably lower than for experiment 1 (median = 100.5).

All participants evaluated conclusions under both necessity instructions and possibility instructions. The first dependent variable was the mean percentage endorsement rates for each problem category, i.e. the number of *yes* responses. The second dependent variable was the time course of the reasoning process; that is to say both premise processing and response times together. These were totalled for each problem type and instruction group to produce a mean reasoning time (in milliseconds). The results from the endorsement rate data are reported first; followed by the results from the reasoning time data (see appendix 3F for ANOVA tables). One participant was excluded before analysis, because it was observed that he did not engage with the task, the *yes*

response was selected in most instances, and latencies were excessively long due to frequent questions directed at the experimenter.

3.8.1 Conclusion endorsement rates

The mean percentage endorsement rates for all four types of problem are shown in table 3.9, broken down by instruction, problem type and ability. The cells for the low ability group represent the mean percentage endorsement rates for the responses from 30 participants, and the cells for the high ability group represent the mean percentage endorsement rates for the responses from 29 participants.

Necessary and PS problems

A 2 (instruction) x 2 (problem type) x 2 (ability) mixed ANOVA test revealed a main effect of instruction [$F(1,57) = 7.62, p < .001, \eta_p^2 = .12$], reflecting higher endorsement rates when problems were presented under possibility instructions than when they were presented under necessity instructions, and of problem type [$F(1,57) = 23.34$, .001, $\eta_p^2 = .29$], whereby Necessary problems were endorsed more frequently than PS problems. This suggests that not only were participants able to discriminate between problem types, which is consistent with experiments 1 and 2, but also that they were able to distinguish between instructions. There was no main effect of ability [$F(1,57) = 1.66, p = .20$].

Table 3.9

Mean percentage endorsement rates for experiment, 3 on all problem types (N = 59, SD in brackets)

	Necessary		PS		Impossible		PW	
Necessary								
Low	75	*(16.24)*	64	*(20.33)*	24	*(21.36)*	31	*(26.00)*
High	80	*(17.76)*	51	*(26.80)*	22	*(19.301*	14	*(18.40)*
M	77	*(17.03)*	58	*(24.35)*	23	*(20.21*	23	*(24.06)*
Possibility								
Low	68	*(21.96)*	66	*(23.01)*	26	*(19.24)*	58	*(29.65)*
High	84	*(16.67)*	76	*(16.48)*	24	*(22.89)*	69	*(26.85)*
M	76	*(20.93)*	71	*(20.96)*	25	*(20.96)*	63	*(28.70)*

There were three significant interactions. The first of these was between instruction and problem type [$F(1,57) = 14.64$, $p < .001$, $\eta_p^2 = .20$], which provided support for the search for counterexamples, suggesting that participants went past the first model on PS problems under necessity instructions (see figure 3.4).
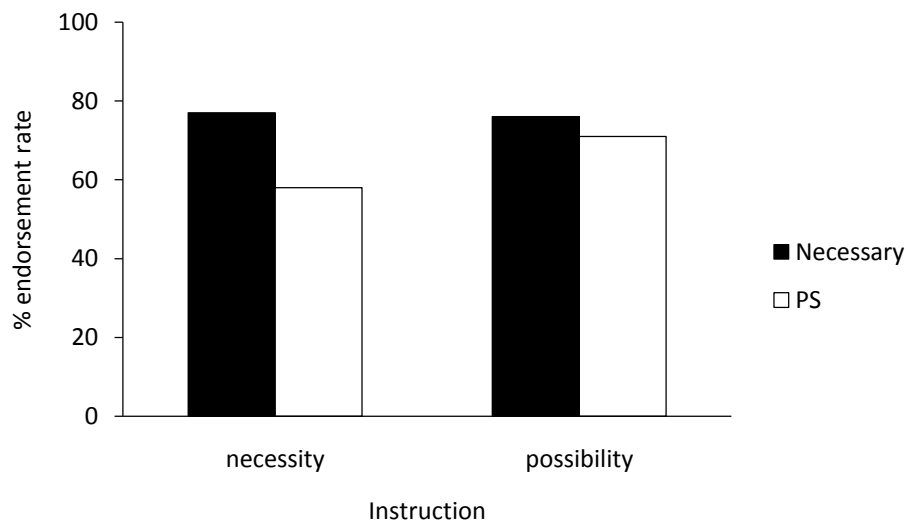
Figure 3.4. Mean percentage endorsement rates for experiment 3, on Necessary and PS problems under necessity and possibility instructions

Follow up between subjects *t*-tests confirmed that there was a significant difference in endorsement rates for Necessary and PS problems under necessity instruction [t(58 = 5.27, p < .001], but not under possibility instructions [t(58 = 1.77, p = .08]. These effects were not found in experiments 1 or 2

The second interaction was between instruction and ability [F(1,57) = 12.65, p < .001, η_p^2 = .18], when the high ability endorsed fewer problems under necessity instructions than the low ability group; but more problems under possibility instructions than the low ability group (see figure 3.5). This was confirmed by follow up between subjects *t*-tests when high ability participants endorsed significantly more problems under possibility instructions than under necessity instructions [t(28) = 4.08, p < .001], but the low ability group did not appear to make this distinction [t(29) = .63, p = .54].
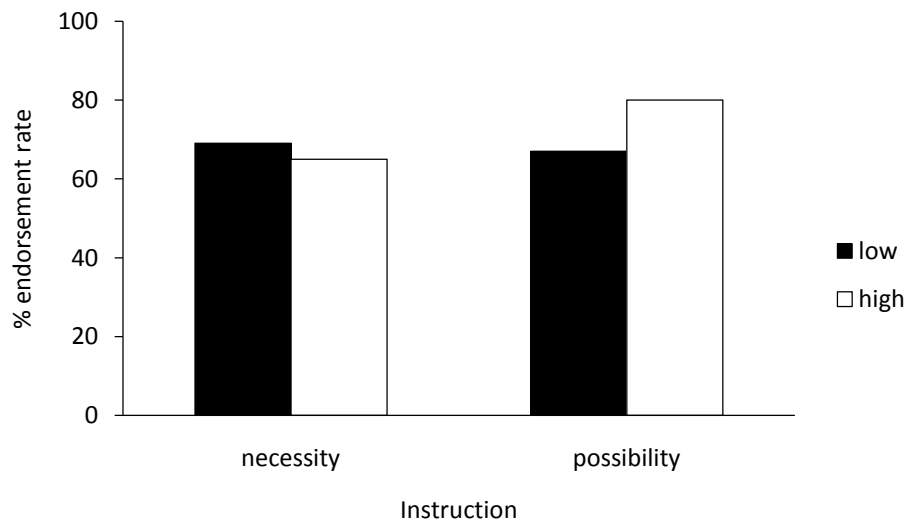
Figure 3.5. Mean percentage endorsement rates for experiment 3, for the low and high ability groups under necessity and possibility instructions
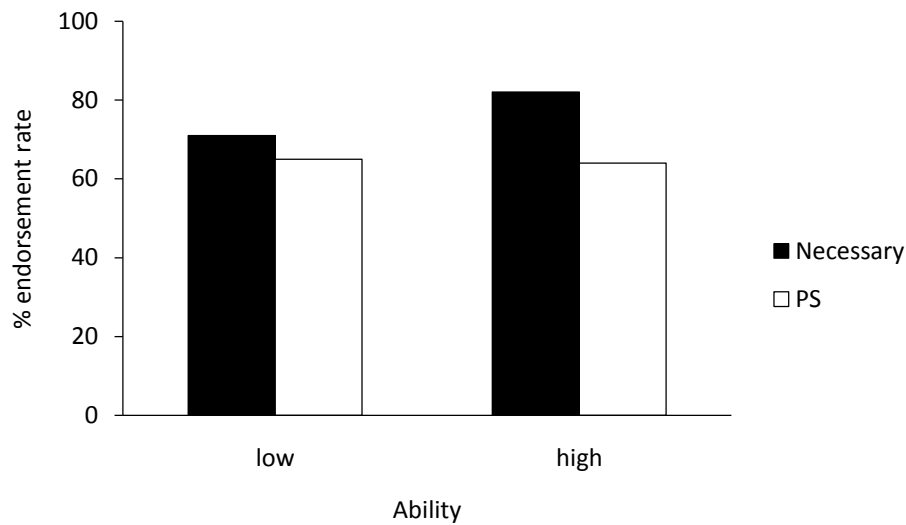
Figure 3.6. Mean percentage endorsement rates for experiment 3, for the low and high ability groups on Necessary and PS problems

The third interaction was between problem type and ability [$F(1,57) = 5.14$, $p < .05$, η_p^2 = .08] suggesting that high ability participants were more able to distinguish between problem types (see figure 3.6). This was confirmed by follow up within subjects *t*-tests when high ability participants endorsed significantly more Necessary problems than PS problems [$t(29) = 5.10$, $p < .001$], but the low ability group treated both types of problem in a similar way [$t(29) = 1.79$, $p = .08$]. This is consistent with experiment 1, but not with experiment 2.

Impossible and PW problems

A 2 (instruction) x 2 (problem type) x 2 (ability) mixed ANOVA test revealed a main effect of instruction [$F(1,57) = 58.33$, $p < .001$, $\eta_p^2 = .51$], and of problem type [$F(1,57) =$ 31.76, $p < .001$, $\eta_p^2 = .36$], reflecting higher endorsement rates when problems were presented under possibility instructions than under necessity instructions, and higher endorsement rates for PW problems than for Impossible problems. This was consistent

with experiments 1 and 2, and suggests that participants understood the differences between both instruction and problem type. There was no main effect of ability [$F(1,57) = .48, p = .48$].

There was a significant two way interaction between instruction and problem type [$F(1,57)$] = 52.53, $p < .001$, $\eta_p^2 = .48$], when endorsement rates were significantly different under possibility instructions but not under necessity instructions (figure 3.7). This was confirmed by repeated measures t-tests between Impossible and PW problems, where the difference was significant under possibility instructions [$t(58) = 7.30, p < .001$], but not under necessity instructions [$t(58) = .13, p = .90$]. This provided support for the search for alternative models, which was in line with experiment 2.
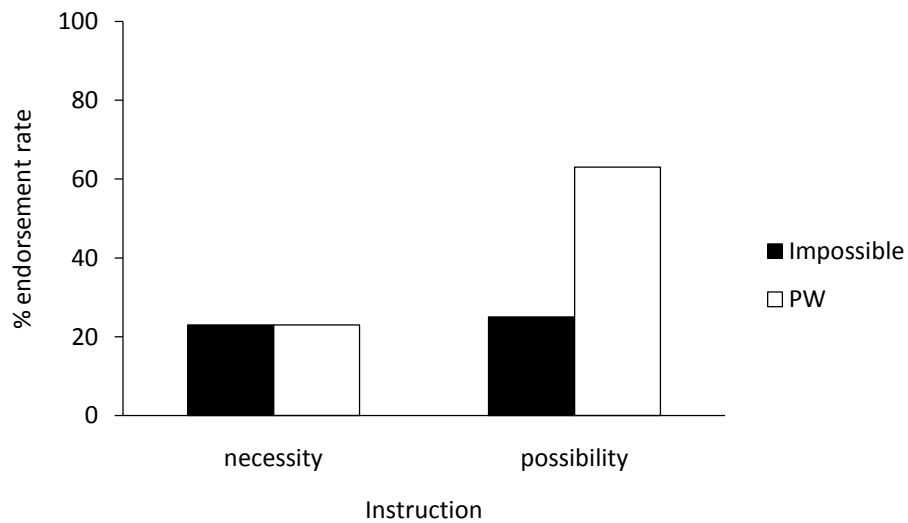
Figure 3.7. Mean percentage endorsement rates for experiment 3, on Impossible and PW problems under necessity and possibility instructions

There was also a significant three way interaction, between instruction, problem type and ability [$F(1,57) = 7.91, p < .05$, $\eta_p^2 = .12$], suggesting that there was a different pattern of responding between ability groups, for problems presented under necessity

instructions (see figure 3.8) and problems presented under possibility instructions (see figure 3.9).

Between subject *t*-tests confirmed that under necessity instructions, the high ability group endorsed significantly less PW problems than the low ability group [$t(28) = 2.97$, $p < .005$], suggesting the high ability participants were more able to understand the meaning of logical necessity; but there was no difference in endorsement rates for Impossible problems [$t(28) = .25$, p .80]. Also there was no difference in the number of problems endorsed by the low ability group for Impossible and PW problems [$t(29) = 1.42$, $p < = .17$], although the high ability group did endorse significantly more Impossible than PW problems [$t(28) = 2.67$, $p < .05$].
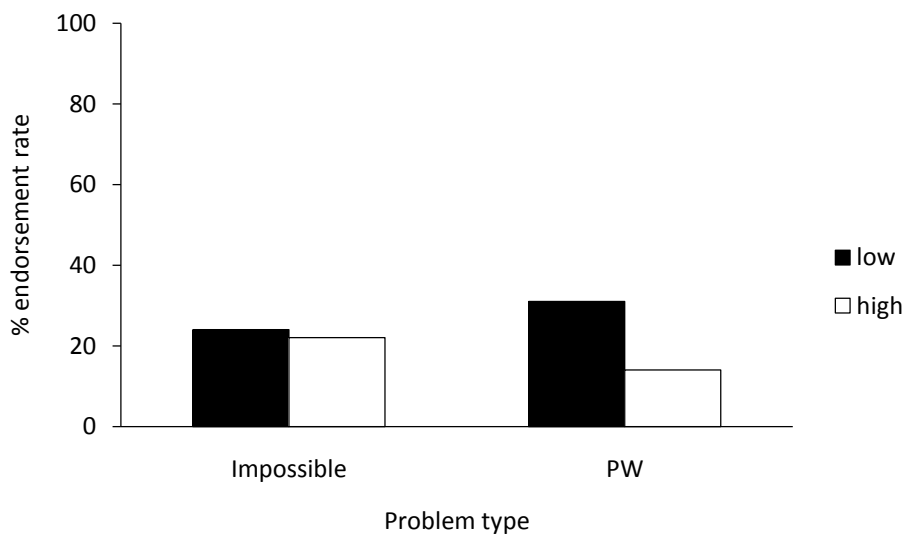
Figure 3.8. Mean percentage endorsement rates for experiment 3, on Impossible and PW problems for both ability groups, under necessity instructions

Within subject *t*-tests confirmed that under possibility instructions, both ability groups endorsed significantly more PW than Impossible problems; low ability [$t(29) = 4.58$, $p <$

.01] and high ability [$t(29)$ = 5.76, p < .01] indicating that they had a good understanding of the differences between Impossible and PW problems. There was no difference in endorsement rates between ability groups for Impossible problems [$t(57)$ = 1.02, p = .38] or PW problems [$t(57)$ = .56, p = .46].
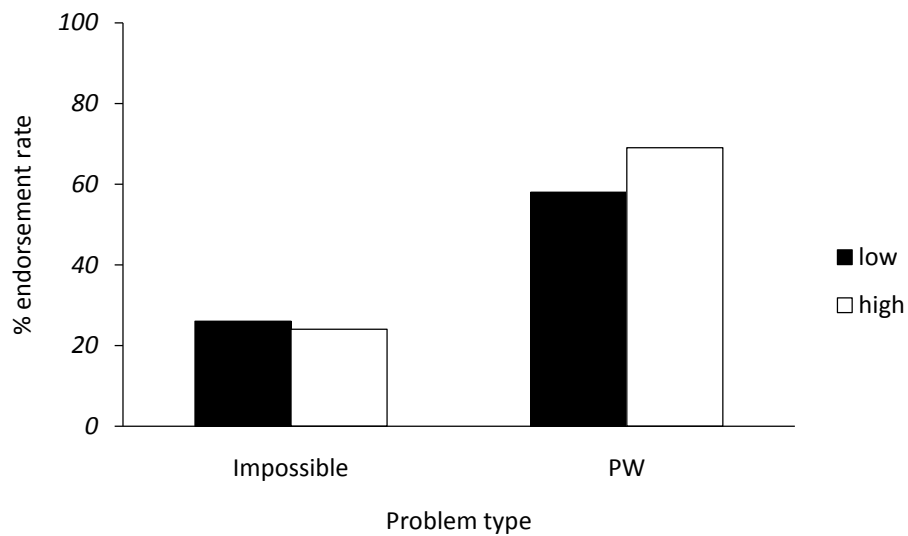
Figure 3.9. Mean percentage endorsement rates for experiment 3, on Impossible and PW problems for both ability groups, under possibility instructions

3.8.2 Reasoning times

The mean reasoning times for all four types of problem are shown in table 3.10, broken down by instruction, problem type and ability. The cells for the low ability group represent the mean reasoning times for the responses from 30 participants, and the cells for the high ability group represent the mean reasoning times for the responses from 29 participants, and are shown in milliseconds.

Table 3.10
Mean reasoning times in milliseconds for experiment 3, on all problem types (N = 59, SD brackets)

	Necessary		PS		Impossible		PW	
Necessary								
Low	20697	*(8369)*	21081	*(8780)*	19565	*(8736)*	20076	*(7458)*
High	16453	*(6224)*	19363	*(7006)*	17250	*(5478)*	17918	*(6335)*
M	18611	*(7636)*	20236	*(7937)*	18427	*(7349)*	19015	*(6955)*
Possibility								
Low	22835	*(12336)*	22303	*(8691)*	21971	*(9178)*	20771	*(78667)*
High	17254	*(7549)*	19143	*(8161)*	17918	*(7972)*	17923	*(6884)*
M	20092	*(10560)*	20749	*(8512)*	19979	*(8774)*	19371	*(7376)*

Necessary and PS problems

A 2 (instruction) x 2 (problem type) x 2 (ability) mixed ANOVA test revealed no main effect of instruction type [$F(1,57) = 1.03$, $p = .32$] which is consistent with experiments 1 and 2. The main effect of problem type just missed statistical significance [$F(1,57) = 3.81$, $p = .056$, $\eta_p^2 = .06$] when Necessary inferences ($M = 19310$) took less time than PS inferences ($M = 20473$); as did the main effect of ability [$F(1,57) = 1.03$, $p = .052$, $\eta_p^2 = .07$], suggesting that high ability participants ($M = 18053$) were quicker reasoners than low ability participants ($M = 21729$). These effects were not found in experiments 1 or 2.

There was a significant interaction between problem type and ability [$F(1,57) = 1.03$, $p < .05$, $\eta_p^2 = .07$], suggesting that high ability participants were quicker reasoners across problem types, and they were more able to discriminate between Necessary and PS problems (figure 3.10).
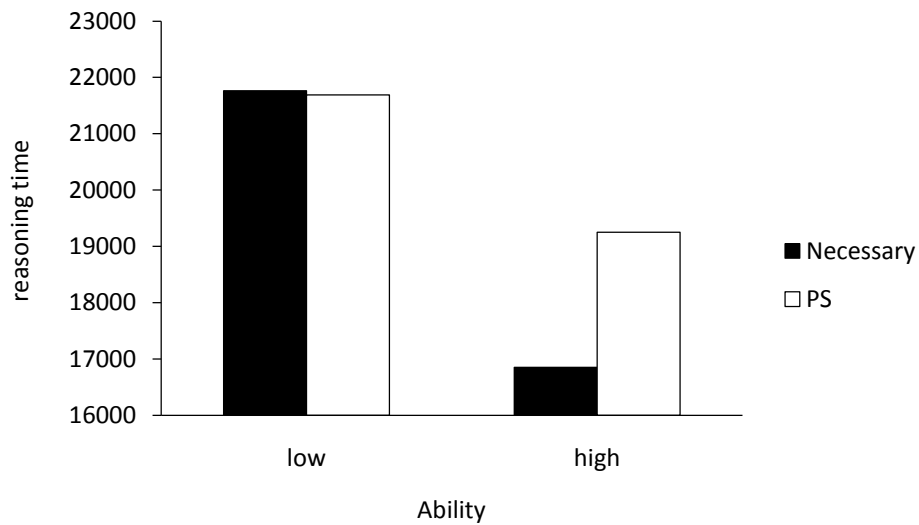
Figure 3.10. Mean reasoning times (in milliseconds) for experiment 3, on Necessary and PS problems for low and high ability groups

A follow up within subjects *t*-test confirmed that those in the high ability group took significantly longer on PS problems than on Necessary problems [$t(28) = 2.94$, $p < .001$]; while the difference for the low ability group was not significant [$t(29) = .09$, $p = .93$]. There was also a significant difference between the ability groups on response times for Necessary problems [$F(1,28) = 5.78$, $p < .05$], when high ability participants were faster, although this did not extend to PS problems [$F(1,28) = 1.75$, $p = .19$]. As discussed in the introduction to experiment 1, we do not have evidence to predict that participants recognise that on Necessary problems a search for other models is not necessary. However we do know that mental model theory predicts that a search for counterexamples is required on PS problems under instructions of necessity. This may have contributed to the longer response times for the high ability group.

Impossible and PW problems

A 2 (instruction) x 2 (problem type) x 2 (ability) mixed ANOVA test revealed no main effect of instruction type [$F(1,57) = 1.36$, $p = .25$], problem type [$F(1,57) = 1.00$, $p = 1.00$], or ability [$F(1,57) = 3.00$, $p = .09$], and there were no interactions. These results were consistent with experiments 1 and 2[17].

3.9 Discussion for experiment 3

The purpose of experiment 3 was to address a number of concerns regarding experiment 2, in terms of lower than norm ability scores, the provision of immediate feedback, and participants failing to engage with the learning and practice phase. Increased emphasis was put on the learning and practice phase to bring it in line with the methodology used by Knauff (1995), when immediate feedback was provided to participants, as to the correctness of judgements made on the spatial relationship between two terms.

Consistent with previous experiments, there was evidence from the endorsement rate data that participants understood the differences between necessity and possibility instructions, and also the differences between problem structures. There was clear support from the endorsement rate data that a search was carried out for counterexamples on PS problems under necessity instructions (in contrast to experiment 2), and that a search for alternatives on PW problems was carried out

---

[17] An equivalent analysis was repeated with log-transformed reasoning times for Impossible and PW inferences, because of the number of outliers, but there were no significant effects.

under possibility instructions (consistent with experiment 2). Also, high ability reasoners were generally quicker on Necessary and PS problems.

In terms of ability effects, the lower endorsement rates for PW problems under necessity instructions suggests that the high ability group understood instructions of logical necessity more than the low ability group. Under possibility instructions there was evidence that both ability groups understood the instructions in a similar way. This was not found in experiment 2, and may be because the ability level of the sample was considerably higher, leading to a better understanding of the differences between necessity and possibility instructions by the high ability group.

Of course a contributing factor may also have been that participants in the higher ability group were more able to apply the skills they had acquired during the training session, which is consistent with conclusions drawn by Neilens (2004), and with the benefit of immediate feedback (Prowse et al., (2009).

Goodwin and Johnson-Laird (2005) proposed that rather than carrying out a search for counterexamples, participants employ the satisficing principle, when the criterion for adequacy is met rather than searching for the optimum solution. However the results from experiment 3 suggest that a significant number of participants routinely went past the first model, and did not merely accept a model that satisficed. Also, when required to evaluate conclusions on problems where the first model did not support the conclusion under instructions of possibility, the high ability reasoners were more inclined to do so.

When considering the latency data; it was suggested (Rauh et al., 2000; Vandierendonck, 2000; Vandierendonck et al., 2004) that the construction of additional models to correctly reject a given conclusion takes longer than when the construction of only one model is necessary, although this was not confirmed by experiments 2 or 3. One point to consider is that it is generally accepted that there is currently no clear explanation as to how response latencies map onto deductive reasoning. In many types of problem solving task, there is a trade off between how fast a task can be performed and how many mistakes are made in performing the task (Evans, Handley, & Bacon, 2009; Thompson et al., 2003). That is, a reasoner can either perform the task quickly with a large number of errors, or slowly with very few errors. Under some testing situations, when people have been instructed to optimize either speed or accuracy, they effectively adopt the appropriate strategy, although results can be difficult to compare and there is a paucity of published research in this area.

The discussion for experiment 2 noted that the mean percentage endorsement rates were substantially lower than those previously recorded by Knauff et al. (1995). This was also the case for experiment 3, when the mean percentage endorsement rate on PS problems under necessity instructions was 77% across both ability groups, which was only slightly higher than the 74% reported in experiment 2. A cautious conclusion for this might be that the participants recruited for experiments 2 and 3 were less able than the sample used by Knauff et al. (1995), although no measure of ability was used in this earlier research.

In view of the increased support for counterexample and alternative search found in experiment 3 compared to experiment 2, there are a number of reasons why this may have been. First the provision of immediate feedback after each individual line relationship in the learning and practice phase may have been effective; second it may well be that the sample was more able, given the 6 point difference in IQ scores; or it may be that the higher ability participants were more receptive to training, which facilitated the provision of normative responding in terms of searching for counterexamples or alternatives. The next section will provide a general discussion of the findings relating to experiment 2 and experiment 3.

3.10 General discussion

The two experiments reported in this chapter used spatial transitive inference tasks to explore whether the lack of evidence to support the search for counterexamples or alternatives in syllogistic reasoning, was primarily because of problem complexity, or because reasoners tended to accept a conclusion that is good enough rather than seeking to find the optimum conclusion. While the first experiment provided *some* support for the search for counterexamples or alternatives, the evidence was limited, and in response to concerns about the learning and practice session being ineffective; changes and improvements were incorporated into a second transitive inference experiment.

With the benefit of the procedural changes to the learning and practice phase; which were to present this phase electronically and give immediate feedback as to the correctness of participant responses; increased evidence was found to support the

notion that reasoners carried out a search for counterexamples on indeterminate problems in order correctly respond *no* when asked if a given conclusions was necessarily true. Also, there was evidence that reasoners carried out a search for alternative models when asked if a given conclusion possibly followed, on indeterminate transitive inference problems. The procedural changes also appeared to be more successfully applied by the high ability group than by the low ability group, as proposed initially by Anderson (1978), so perhaps because this group of reasoners had a better understanding of both how to execute the skill and how to use the skills.

Interestingly, it appears that the choice of materials, based on those previously used by Knauff et al. (1995), which was influenced by the desire to overcome interpretational problems and problems relating to lack of clarity, may have increased the emphasis on the learning and practice phase. It may well be that this problem has not been encountered previously, because most studies exploring the processes involved in, and how reasoners make transitive inferences, have used traditional materials such as A is longer than B, B is longer than C, with much of the focus on whether these are interpreted by visualising the terms on a horizontal or vertical axis, or whether reasoners employ a more linguistic strategy.

In conclusion, there was evidence from both experiments to support the search for counterexamples and alternatives, and also that people understood the differences between necessity and possibility instructions, and the different problem structures. The support was more conclusive in experiment 3, where there was an increased emphasis for participants to have a clear understanding of the terms they were using,

together with the provision on of immediate feedback. It should also be noted that the IQ scores for the sample were considerably higher for experiment 3. The next experiment will extend the range of materials to look at how reasoners make conditional inferences, which is another reasoning paradigm that has been extensively used in deductive reasoning research.

Chapter 4

The search for counterexamples & alternative models on conditional inference tasks with abstract content

The two experimental paradigms which have been used to investigate whether reasoners routinely search for counterexamples or alternatives, produced conflicting evidence. The first paradigm, syllogistic reasoning, used abstract reasoning problems, but failed to produce evidence to support the search for counterexamples as proposed by the mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991), or alternatives. On the other hand there was evidence that high ability participants were more able to discriminate between instructions of necessity and instructions of possibility.

However the results from experiments 2 and 3, which both used the transitive inference paradigm to test the assumptions of the third stage of the mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991), told a different story. Although the endorsement rate results from experiment 2 failed to provide evidence to

support the search for counterexamples, there was evidence that a search for alternatives was carried out from both the endorsement rate data and the latencies. The second transitive inference experiment found that the high ability group were better at detecting differences between both instruction and problem types in terms of making more correct inferences, and there was evidence to support the search for counterexamples under necessity instructions. There was also evidence to support the search for alternatives under possibility instructions, which was mediated by ability in that the high ability individuals were more likely to carry out this search.

A number of reasons were considered for the increased support found for mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991) provided by experiment 3, including a 6 point difference in IQ scores, and the provision of immediate feedback during the learning and practice phase.

The experiment reported in this chapter extends the range of research paradigms to conditional reasoning, to investigate whether the results from the second transitive inference experiment can be generalised to conditional inference tasks with abstract content.

4.1 Introduction to experiment 4

Conditional reasoning involves drawing inferences about situations in which the occurrence of one event is conditional or contingent upon the occurrence of another event. The large body of research that has been carried out in this area over the past forty years falls broadly into 2 areas: first, where the emphasis is on reasoning with abstract or knowledge lean relations; and second, when the focus is on observing

reasoning behaviours using materials which draw on our everyday knowledge. This chapter will concentrate on the former; conditionals with abstract content, which are not affected by semantics, or open to interpretation with reference to our knowledge of the world.

The conditional inference task, which is frequently used to investigate conditional reasoning[18], requires participants to draw inferences from the truth or falsity of one component of the conditional to the truth or falsity of the other. The task involves making an inference on the basis of a major premise 'if p then q', where 'p' is referred to as the antecedent and 'q' as the consequent. The major premise is followed by a minor premise as shown in table 4.1, and the four traditionally studied inferences, based on standard logic, are Modus Ponens (MP), Modus Tollens (MT), Affirmation of the Consequent (AC) and Denial of the Antecedent (DA) (see table 4.1). As discussed in chapter 1, the basic form of the major premise can be negated to produce a total of sixteen different premise combinations. Typically, studies using abstract materials, present the major and minor premises with letters and numbers (if there is an A there is a 3) or colours and shapes (if it is a square then it is blue).

Logically speaking, the validity of an argument is determined by its syntactic form. The inferences MP and MT are valid, in that the conclusions are necessitated by the premises; whereas AC and DA are fallacies because the conclusion does not logically follow from the premises, or in other words it is not the only conclusion that is possible given the truth of the premises. Endorsement rates for studies using the basic form of

---

[18] Other tasks include the Wason selection task and the truth table task - see general texts such as Evans et al. (1993).

conditionals and abstract content (if p then q) vary considerably, although MP inferences are generally high, at between 89% (Kern, Mirrels & Hinshaw, 1983) and 100% (Evans, 1977). The more difficult MT inferences are lower and more variable, ranging from 41% (Kern et al., 1983) to 81% (Rumain et al., (1983). Endorsement rates for invalid AC and DA conditionals are also more variable, and tend to fall somewhere between 27% (Kern et al, 1983) and 75% (Evans, 1977) for AC conditionals, and 28% (Kern et al., 1983) and 69% (Evans, 1977) for DA conditionals.

Table 4.1
Conditional inferences with basic major premises

	MP		MT		AC		DA	
	Given	Conclude	Given	Conclude	Given	Conclude	Given	Conclude
If p then q	p	q	not q	not p	q	p	not p	not q

Previous chapters in this thesis have described the mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991) in relation to both syllogistic reasoning and transitive inference, where the third stage proposes that reasoners carry out a search for counterexamples where the premises are true but the conclusion is false. In the context of conditional inference, the mental model theory assumes that people begin by forming an initial representation of the premises that is incomplete. Some situations are explicitly modelled, and others are left implicit. For example, the conditional *if p then q* might be represented initially as shown below, where all cases of *p* are exhausted, and other unspecified models are indicated by ellipses.

[p]   q

...

Therefore, given *p*, the reasoner can immediately infer *q* (MP). The MT (if p then q, not q, therefore not p) inference however is more difficult in terms of model formation, and given the premise *not q*, no inference immediately follows because the negated form of *q* is not represented in the model. Hence, the model needs to be fleshed out beyond the initial representation:

    [p]   q

    not p   [not q]

    …

When considering the high number of error rates reported in the conditional reasoning literature, for both AC and DA conditionals; within the framework of the mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991) perhaps the most credible explanation is that reasoners fail to flesh out the full set of models. This process is necessary to indicate that *not p* may occur in the presence of *q* (disconfirming AC), or *q* may occur in the presence of *not p* (disconfirming DA), before rejecting AC/DA inferences, (see next page):

    p   q

    not p   not q

    not p   q

Another possible explanation for error rates on AC and DA inferences is that reasoners interpret the premises biconditionally, where they wrongly assume that *if p then q*

means the same as *if q then q.* Therefore if there are *p's* there are *q's* and if there are *q's* there are *p's*, and if there are *no p's* there are *no q's* and if there are *no q's* there are *no p's*. There is however convincing developmental evidence against a biconditional explanation (Barrouillet, Grosset, & Lecas, 2000; Barrouillet & Lecas, 1998); suggesting that although young reasoners go through a *stage* when they display a tendency to represent the terms biconditionally, leading to all four inferences being made (MP, MT, AC, DA); on approaching adulthood a basic *if p then q* conditional is represented with the models *pq, not p, q, not q.* It may also be, as suggested by Evans and Over (2004) that some experimental procedures encourage biconditional reasoning more than others, depending upon the context and aims of the research.

A large body of literature exists on conditional reasoning with abstract materials , where focus has been predominantly on either content or structure, rather than exploring error rates within the context of whether they construct models in the form of counterexamples. However a frequently cited study, Schroyens, Schaeken and Handley (2003), explored the factors affecting the likelihood with which people engage in a search for counterexamples, by introducing a timing constraint. When this constraint was in place, reasoners were more likely to accept the first model that supported the conclusion, leading to errors on both AC and DA conditionals. This suggests that reasoners need time to search for counterexamples to test inferences, in order to produce a logically correct response.

A second experiment (Schroyens et al., 2003), and later research (Schroyens & Schaeken, 2008), elaborated on the instructions, by presenting inference problems

under either necessity instructions (is it necessary that) or instructions that did not specifically mention the word necessary (does it follow that), when there was found to be an increased tendency for participants to look for falsification when the emphasis on logical necessity was amplified. This is turn, produced more logically correct evaluations.

Schroyens, Schaeken and Handley (2003) suggest the likelihood with which people engage in a search for counter examples is affected by temporal and motivational constraints Therefore, when given time to search, together with being primed with clear logical instructions, reasoners can and do search for counterexamples, leading to the observed acceptance rates for logical fallacies (AC/DA) to be lower.

This pattern of results provides support for the notion that individuals carry out a search for counterexamples as proposed by the third stage of the mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991), but whether this is the default mechanism is open to interpretation. There is also a suggestion (Schroyens & Schaeken, 2008; Schroyens et al., 2003) that this is consistent with the model base theory proposed by Schroyens et al. (2001), known as the Syntactic-Semantic Counterexample Prompted Probabilistic Thinking and Reasoning Engine (SSCEPPTRE).

SSCEPPTRE implies that evaluation of a model involves active reasoning, which has a probability component leading to a revised specification of the mental model theory originally proposed by Johnson-Laird (1983; Johnson-Laird & Byrne, 1991). The likelihood that people will endorse standard inferences is captured by an equation, which in a simplified form, has three distinct components: first, reasoners must be

motivated to search for counterexamples; second even if they are motivated they must be able to construct the models, and third they must be able to evaluate the model in order to accept or reject the model. Perhaps one advantage of SSCEPPTRE is that it can be applied to confirmations of provisional conclusions, such as alternatives, and not just the validation-by-falsification which is the hallmark of the mental model theory.

4.1.1 Aims and rationale

The aim of experiment 4 was to investigate whether reasoners search for counterexamples on abstract conditional inference problems, and alternative models under instructions of possibility. Consistent with experiments 1, 2, and 3, the relationship between cognitive ability and carrying out a search for counterexamples or alternative models was also explored, to look at whether ability is a good predictor of whether this successfully takes place. There were limited ability effects from the previous experiments, when there was evidence that high ability participants were more likely to search for alternative models on PW problems under possibility instructions. Measures of cognitive ability in studies of conditional reasoning have not been widely used, but the small number of studies that have done so, report that performance on logically invalid AC and DA inferences correlates highly with cognitive ability (Evans et al., 2007; Newstead et al., 2004), so it may well be that higher ability participants will be more likely to search for counterexamples and alternative models. To date, research (Schroyens & Schaeken, 2008; Schroyens et al., 2003) suggests that reasoners are able to search for counterexamples; but there is no evidence to suggest that this is a compulsory part of the process unless reasoners are motivated to do so.

The four problem types employed for this experiment were similar to those used in experiments 1, 2 and 3: Necessary (the conclusion must be true), PS (the conclusion may be true as the first model supports the conclusion), PW (the conclusion may be true, although the first model negates the conclusion) and, Impossible (the conclusion must be false (Necessary, PS, Impossible and PW).

The basis for problem selection particularly when finding PS and PW problems was less straightforward than for the experiments 1 - 3. The reasoning problems are composed of conditional inferences with both standard and opposite conclusions. First consider the conditional inferences shown in table 4.2 with standard conclusions, which lead to the composition of Necessary and PS problem types:

Table 4.2
Conditionals with standard conclusions

	Necessary		PS
MP	if p then q p therefore, q	AC	if p then q q therefore, p
MT	if p then q not q therefore, not p	DA	if p then q not p therefore, not q

Under necessity instructions the MP and MT arguments are valid; given that the premises are true and given *p* or *not q*, it is *necessary* that there is a *q* (MT) or *not p* (MT). The second arguments (AC and DA) are invalid, as there may or may not be *p*'s in the case of AC arguments, and similarly there may or may not be *q*'s in DA arguments.

Under possibility instructions the conclusions for all four arguments are *possible*, and as we have seen, AC and DA are endorsed at very high rates (Evans et al., 1995), suggesting that they are 'strong' conclusions, i.e. based upon the first models that come to mind. The write up of this experiment will refer to AC and DA from standard premises as PS conclusions and MP and MT as Necessary conclusions.

Next consider conditional inferences with opposite conclusions under necessity and possibility instructions; where for MP and MT problems the conclusion is inconsistent with all models and is therefore *impossible*, and referred to as Impossible problems (table 4.3). For AC and DA arguments the conclusion is *possible*, but not *necessary*, and is also inconsistent with an initial representation of the premises; so to judge these conclusions as *possible* a search for alternative models must take place. These will be referred to as PW conclusions, because they are *possible* conclusions that are rarely endorsed (Evans et al., 1999).

Table 4.3
Conditionals with opposite conclusions

	Impossible		PW
MP	if p then q p therefore, not q	AC	if p then q q therefore, not p
MT	if p then q not q therefore, p	DA	if p then q not p therefore, q

4.1.2 Predictions

The predictions relating to endorsement rates, reasoning times and ability are based on the general assumptions of the third stage of the mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991); and are similar to those explicitly set out for experiment 1.

1. If participants search for counterexamples or alternatives; there will be fewer endorsements of PS problems than Necessary problems under necessity instructions, and more endorsements of PW problems than Impossible problems under possibility instructions.

2. On problems requiring a search for counterexamples or alternatives, participants will take longer.

3. Participants with higher cognitive ability will be quicker reasoners, produce more logically correct responses, and take proportionately longer on problems requiring a search for other models in order to provide a logically correct evaluation of the given conclusion.

4.2 Method

Design

This experimental study was carried out using a within-subjects design, when initially participants completed an AH4 cognitive ability test. This was followed by the conditional reasoning task, where participants were presented with one block of 32 randomised conditional inference problems under necessity instructions, and one

block of 32 randomised conditional inference problems under possibility instructions, the order of which was counterbalanced to minimize order effects.

Participants

A total of 60 undergraduate students from the University of Plymouth took part in the study, in return for either payment or course credit. The sample consisted of 24 males and 36 females with a mean age of 22 years, and they were all native English speakers. No participants were dyslexic, or had received formal training in logic.

Materials and procedure

The procedure adopted was similar to that used in the previous three experiments. Participants were run in groups of between 4 and 8 in a laboratory containing several computers. Each participant was seated at their own workstation, to avoid distraction.

Cognitive Ability Test

Initially, as a measure of ability, participants completed Parts I and II of the AH4 Test of Cognitive Ability (Heim, 1968), which was administered in accordance with the test instructions and followed the procedure used in experiment 1. Question booklets and answers sheets were collected by the experimenter before moving on to the conditional inference task.

Conditional inference task

The set of 32 conditional inference statements consisted of 8 problems in each of the four categories: Necessary, Impossible, PS, and PW; and in order to provide a balanced mix of problem structures (MP, MT, AC, and DA), there were two different argument

forms within each category (see table 4.4). Randomly chosen letters of the alphabet (excluding I and O) and numbers (excluding 0 and 1) were used for the premise terms, for example: if the letter is a B then the number is a 6, the letter is a B, therefore the number is a 6. A complete set of the problems can be found in appendix 4A.

A computer with a 15" monitor screen was used to present the problems with the computer program. The keyboard was adapted to include *yes* and *no* keys, which were systematically counterbalanced, so that half the participants had the *yes* key on the left of the keyboard and the *no* key on the right, while the other half had these positions reversed.

Table 4.4
The four problem categories, together with premises, conclusions and inference type

	Premises & conclusion	Inference	Logical definition
Necessary	If p then q, p, q	MP	The conclusion must be true given that the premises are true
	If p then q, not q, not p	MT	
Impossible	If p then q, p, not q*	MP	The conclusion cannot be true given that the premises are true
	If p then q, not q, p*	MT	
PS	If p then q, q, p	AC	The conclusion might be true given that the premises are true
	If p then q, not p, not q	DA	
PW	If p then q, q, not p*	AC	The conclusion might be true given that the premises are true
	If p then q, not p, q*	DA	

*conclusion presented in opposite direction

The two sets of written task instructions which included examples of the screen layout, were printed on A4 paper, and were similar to those used in the previous experiments.

These related to whether problems were being evaluated for either necessity correctness or possibility correctness. Examples of the screen layouts are shown in table 4.5 and a complete set of instructions is presented in appendix 4B and appendix 4C.

Table 4.5
Screen layouts included in task instructions

Screen 1

Given:
          If the letter is a T then the number is a 4
                    The number is a 4

Screen 2

Given:
          If the letter is a T then the number is a 4
                    The number is a 4
Is it necessary that
                    The letter is a T

Screen 1

Given:
          If the letter is a B then the number is a 6
                    The number is a 6

Screen 2

Given:
          If the letter is a B then the number is a 6
                    The number is a 6
Is it possible that
                    The number is a 6

The instructions were distributed (necessity or possibility) for the first block of problems and after a short reading period, participants were given the opportunity to ask questions on any points that they were less clear about. The participants were also told that they must ask the experimenter for the second set of instructions (necessity or possibility) as soon as a message appeared on the screen, and reminded that at the start of each block there were two practice questions.

Participant responses, *yes* or *no*, were recorded by the program, together with the time taken to indicate understanding of the problem (screen 1) and the time taken to complete the reasoning process (screen 2). These were saved to disc.

4.3 Results

The AH4 test sheets were scored in accordance with the test instructions, when one mark was given for each correct answer. There was a significant positive correlation between Parts I and II ($r = .62$, $p < .01$), and the scores from both parts were added together to give an overall general ability score for each participant. The observed mean for participants was 102.27 ($SD = 16.35$), which was higher than for the previous experiments, and higher than the available norm of 96.36 ($SD = 15.01$) for university students (Heim, 1968). The sample was divided into high and low cognitive ability groups, on the basis of a median split on the AH4 test scores; cases below the median of 104 were classified as low ability and those above the median were classified as high ability. None of the participants recorded a score of 104.

All participants evaluated conclusions under both necessity instructions and possibility instructions. The first dependent variable was the mean percentage endorsement rates

for each problem category, i.e. the number of *yes* responses. The second dependent variable was the time course of the reasoning process; that is to say premise processing and response times together. These were totalled for each problem type and instruction group to produce a mean reasoning time (in milliseconds). Although for the purposes of this study the analysis was carried out on problem types (Necessary, PS, Impossible and PW), a breakdown of endorsement rates into argument forms can be found in appendix 4D. The results from the endorsement rate data are reported first; followed by the results from the reasoning time data. All ANOVA tables for experiment 4 are shown in appendix 4E.

4.3.1 Conclusion endorsement rates

The mean percentage endorsement rates for all four types of problem are shown in table 4.6, broken down by instruction, problem type and ability. The cells for the low ability group and the high ability group represent the mean percentage endorsement rates for the responses from 30 participants.

Table 4.6
Mean percentage endorsement rates for all problem types (N = 60, SD in brackets)

	Necessary		PS		Impossible		PW	
Necessary								
Low	70	*(19.61)*	66	*(27.88)*	17	*(14.11)*	17	*(18.12)*
High	70	*(17.56)*	50	*(30.88)*	13	*(15.72)*	10	*(17.40)*
M	70	*(18.45)*	58	*(30.24)*	15	*(14.93)*	14	*(17.93)*
Possibility								
Low	83	*(16.24)*	83	*(14.53)*	24	*(15.72)*	37	*(31.65)*
High	89	*(12.60)*	93	*(10.69)*	32	*(18.48)*	55	*(33.10)*
M	86	*(14.75)*	88	*(13.62)*	28	*(17.47)*	46	*(33.14)*

Necessary and PS problems

A 2 (instruction) x 2 (problem type) x 2 (ability) mixed ANOVA test revealed a main effect of instruction [$F(1,58) = 47.09, p < .001, \eta_p^2 = .45$], reflecting higher endorsement rates when problems were presented under possibility instructions; and of problem type [$F(1,58) = 8.35, p < .005, \eta_p^2 = .13$], whereby Necessary problems were endorsed more frequently than PS problems. The main effect of problem type reflects a tendency to endorse more valid (MP/MT) conclusions than invalid (AC/DA) conclusions, in line with the literature. There was no main effect of ability [$F(1,58) = .01, p = .94$]. The main effect of instruction and problem type was consistent with the previous experiments, suggesting that participants understood the differences between the instructions, and also problem structures.

There was a significant two way interaction between instruction and problem type [$F(1,58)] = 9.84, p < .005, \eta_p^2 = .15$], mainly because of the difference in endorsement rates under necessity instructions. This was confirmed by repeated measures *t*-tests where the difference was significant under necessity instructions [$t(59) = 3.33, p < .005$], but not under possibility instructions [$t(59) = .78, p = .90$]; this is consistent with experiment 3. There was also a significant three-way interaction between instruction, problem type and ability [$F(1,58) = 5.90, p < .005, \eta_p^2 = .09$], reflecting a different pattern of responding for the low ability group than the high ability group (see figure 4.1 and figure 4.2). The high ability group were able to discriminate between Necessary and PS problems under necessity instructions, which was confirmed by a within subjects *t*-test [$t(29) = 4.15, p < .001$]; but this effect was not present for the low ability group [$t(29) = .72, p = .48$]. These findings provide strong support for the

prediction that the high ability participants would be better performers, due to a search

for counterexamples on PS problems under necessity instructions.
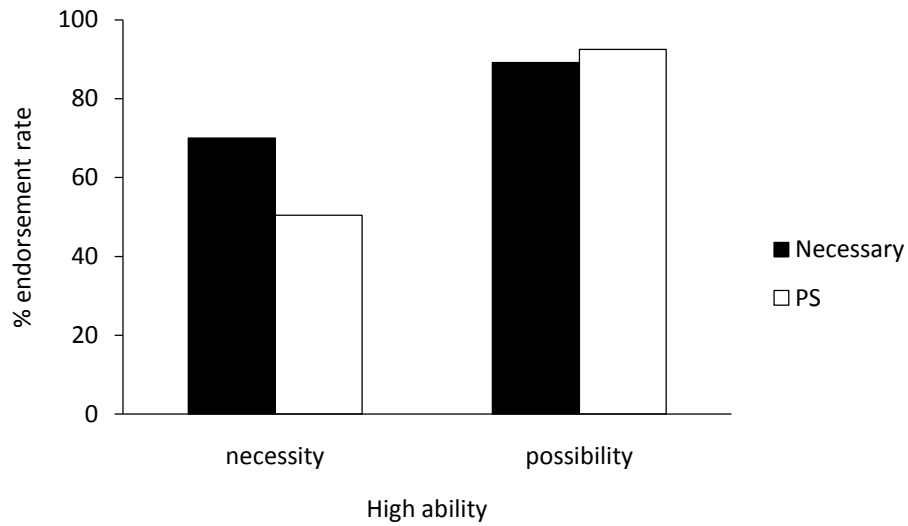
Figure 4.1. Mean percentage endorsement rates for Necessary and PS problems under necessity and possibility instructions for the high ability group
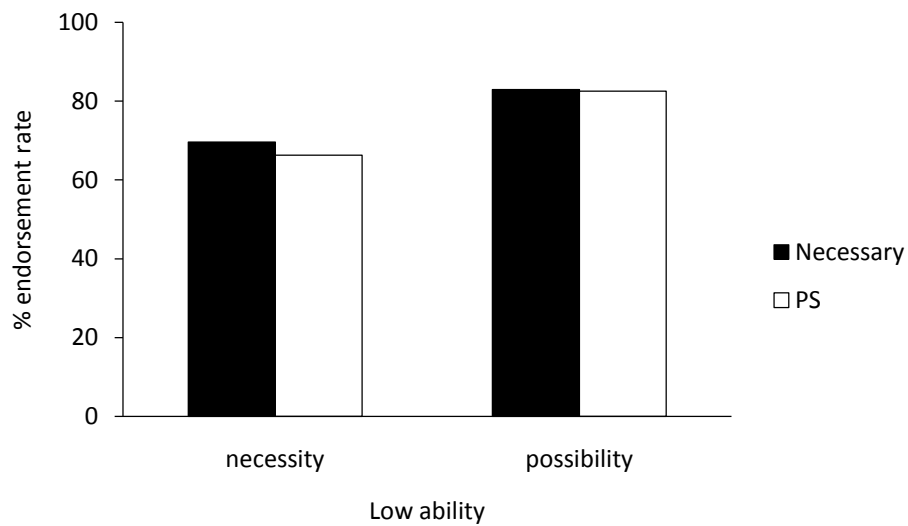
Figure 4.2. Mean percentage endorsement rates for Necessary and PS problems under necessity and possibility instructions, for the low ability group

Impossible and PW problems

A 2 (instruction) x 2 (problem type) x 2 (ability) mixed ANOVA test revealed a main effect of instruction [$F(1,58) = 44.39$, $p < .001$, $\eta_p^2 = .43$], reflecting higher endorsement rates under possibility instructions; and of problem type [$F(1,58) = 14.73$, $p < .001$, $\eta_p^2 = .20$], whereby PW problems were more frequently endorsed than Impossible problems. There was no main effect of ability [$F(1,58) = 1.63$, $p = .21$]. The main effect of instruction and problem type was consistent with previous experiments, suggesting that participants had an understanding of the differences between both instruction and problem types.

There were two significant interactions. The first was between instruction and problem type [$F(1,58) = 18.58$, $p < .001$, $\eta_p^2 = .24$], which supports the search for alternatives on PW problems under possibility instructions (see figure 4.3); and is consistent with experiments 2 and 3.
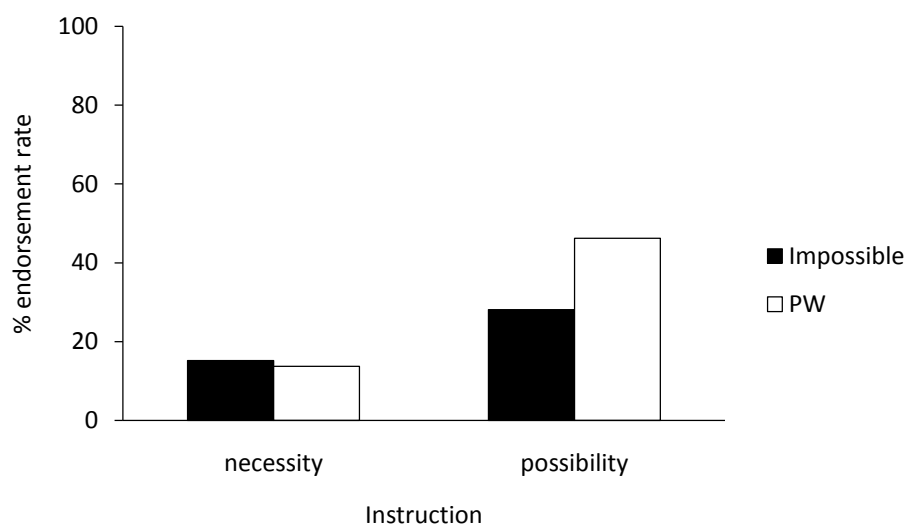
Figure 4.3. Mean percentage endorsement rates for Necessary and PW problems under necessity and possibility instructions

This effect was confirmed by follow up t-tests, which revealed significant differences between Impossible and PW problems under possibility instructions [$t(59)$ = 4.77, $p <$ .001], but not when the problems were presented under necessity instructions [$t(59)$ = .62, p = .54].

The second interaction was between instruction and ability [$F(1,58)$ = 7.23, $p < .01$, η_p^2 = .11], whereby the higher ability participants endorsed more problems under possibility instructions, where a search for alternative models was required(see figure 4.4). This was confirmed by repeated measures t-tests where high ability participants endorsed significantly more problems under possibility instructions [$t(1,58)$ = 2.40, p <.05], and the difference under necessity instructions was not significant [$t(1,58)$ 1.48, p = .15]. This finding was consistent with experiment 1 (syllogistic reasoning).
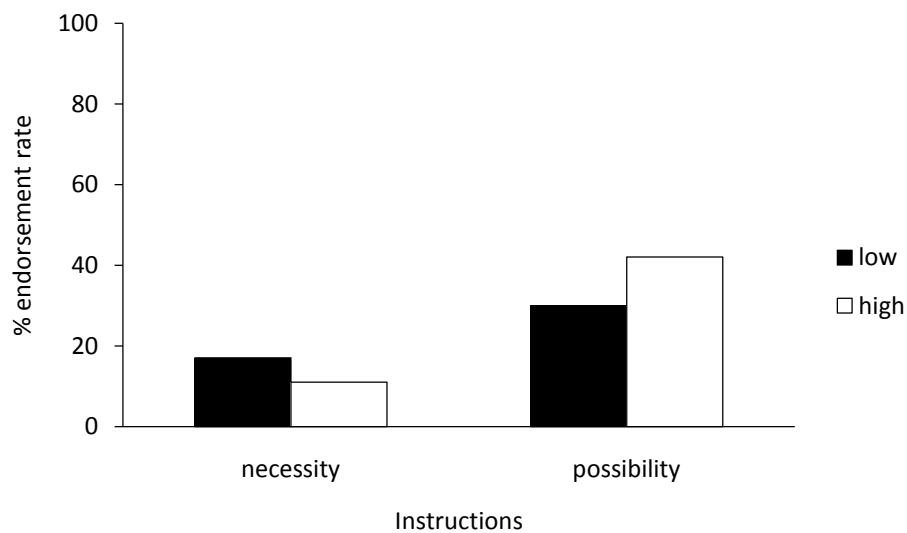
Figure 4.4. Mean percentage endorsement rates for the high and low ability group under necessity and possibility instructions

4.3.2 Reasoning times

The mean reasoning times for all problem types are shown in table 4.7, and are again broken down by instruction, problem type and ability. The cells for low ability group and the high ability group represent the mean reasoning times for the responses from 30 participants, and are shown in milliseconds.

Table 4.7
Mean reasoning times (milliseconds) for all problem types (N = 60, SD in brackets)

	Necessary	PS	Impossible	PW
Necessary				
Low	11333 *(4437)*	10857 *(3135)*	11173 *(3753)*	10787 *(3351)*
High	10170 *(4505)*	11020 *(5593)*	10277 *(4795)*	10999 *(4685)*
M	10752 *(4472)*	10938 *(4496)*	10725 *(4293)*	10893 *(4040)*
Possibility				
Low	11102 *(3111)*	11427 *(3342)*	11791 *(3072)*	11811 *(3393)*
High	9436 *(3243)*	9883 *(3305)*	10339 *(4402)*	10414 *(3966)*
M	10268 *(3261)*	10655 *(3386)*	10915 *(3866)*	11113 *(3726)*

Necessary and PS problems

A 2 (instruction) x 2 (problem type) x 2 (ability) mixed ANOVA revealed no main effects of instruction [$F(1,58) = .96$, $p = .33$], problem type [$F(1,58) = 2.19$, $p = .14$], or ability [$F(1,58) = 1.87$, $p = .18$]. There were no significant interactions.

Impossible and PW problems

A 2 (instruction) x 2 (problem type) 2 x (ability) mixed ANOVA test was carried out, but there were no main effect of instruction [$F(1,58) = .21$, $p = .65$], problem type [$F(1,58) = 1.16$, $p = .29$], or ability [$F(1,58) = 1.20$, $p = .28$]. There was a significant interaction

between problem type and ability [$F(1,58) = 4.65$, $p < .05$, $\eta_p^2 = .07$]. Follow up within subject *t*-tests were carried out, on the high ability group (see figure 4.5), when there was found to be a significant difference in reasoning times between Impossible and PW problems [$t(1, 29) = 2.23$, $p < .05$], but there was no evidence of this for the low ability group [$t(1, 29) = .78$, $p = .44$]. The processing speeds for the low ability were longer, which was as predicted.
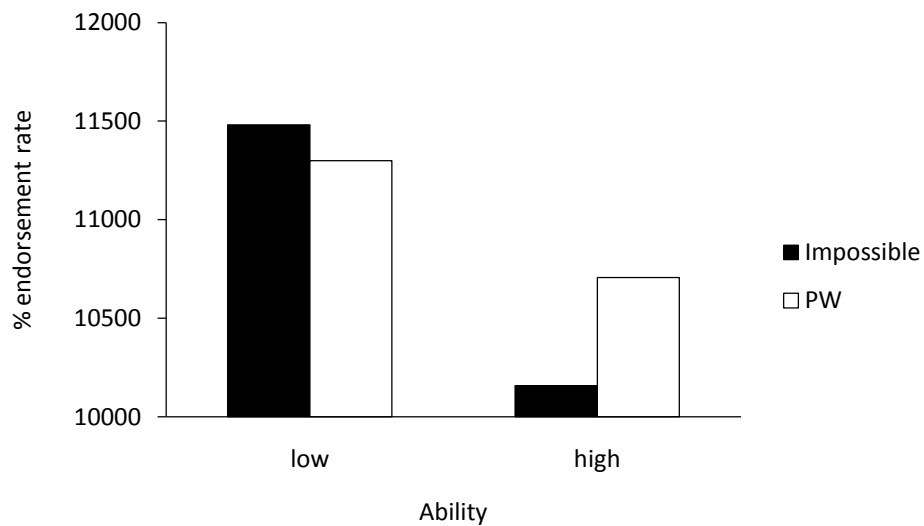
Figure 4.5. Mean reasoning times (in milliseconds) for Impossible and PW problems under necessity and possibility instructions

4.4 Discussion

The purpose of this study was to further investigate whether reasoners searched for counterexamples or alternative models when making conditional inferences with abstract materials; and whether cognitive ability is a good predictor of whether this search takes place. Consistent with the previous experiments reported in this thesis, there was evidence from the endorsement rate data to suggest that participants

understood the differences between necessary and possibility instructions, and also the differences between Necessary and PS problem structures and the differences between Impossible and PW problem structures.

There was clear supporting evidence from the endorsement rate data that the high ability participants were searching for counterexamples on PS problems under necessity instructions; and that both ability groups were searching for alternative models on PW problems under possibility instructions. These effects are consistent with experiment 3 (the second transitive inference experiment), although the ability effects were different, in that experiment 3 found ability effects relating to PW problems under possibility instructions, rather than PS problems under necessity instructions found in experiment 4.

When considering the latency data from experiment 4, previous research (Schroyens & Schaeken, 2008; Schroyens et al., 2003) has found that reasoners display a tendency to search for counterexamples when there are no time constraints, leading to an increased number of correct evaluations of indeterminate problems. However, even though there were no time constraints in place for experiment 4, there was only limited evidence from the latencies to support the search for counterexamples or alternatives. Analysis suggests that high ability reasoners took significantly longer on PW problems across both problem types; but we can only speculate that this was because of the time taken to carry out a search for alternative model under possibility instructions. One possible explanation that has been offered for consideration (Evans et al., 2007) was that in past research the processing time for MT (*if p then q, not q*) and DA arguments (*if p then q,*

not p) is longer because of the need to process negations. However in experiment 4, it is unlikely that this is the case, because the reasoning problems were made up of a balanced mix of problem types; although it may be that the effect of processing negations overrides, or at least influences, the time taken to access additional alternatives.

In looking at the results from a theoretical viewpoint, it would seem that while the mental model theory proposed by Johnson Laird (1983; Johnson-Laird & Byrne, 1991), claims that a search for counterexamples is the default in deductive reasoning; it may be that this claim needs to be modified. Clearly, as has been shown in experiment 2, 3 and 4, people can and sometimes do carry out a search for alternative models, when given instructions to evaluate reasoning problems for logical necessity or possibility.

Schroyens et al. (2003) found that temporal and motivational constraints affect the likelihood with which people engage in a search for counterexamples; with an increased tendency for individuals to look for falsification when given time to test the inferences made, and when the emphasis on logical necessity was increased. This in turn led to more logically correct evaluations, and to the identification of three main factors which contribute to whether reasoners search for counterexamples (Schroyens & Schaeken, 2008; Schroyens et al., 2003), which were motivation, ability to construct models, and ability to evaluate the model in order to accept or reject the initial model constructed. These factors are linked to the model based theory SSCEPTRE (Schroyens et al., 2001), which takes into account validation by falsification, and rejection by confirmation.

Although the response time data does not support the search for counterexamples, and does not *directly* support the search for alternative models, the high ability group took longer on PW problems in relation to Impossible problems, and were generally quicker reasoners. This would support the notion that the high ability participants can recognise that PW problems are different to Impossible problems, which is consistent with the general findings reported across the experimental studies, that higher ability leads to a greater capacity to distinguish between different instructions and different logical structures.

To summarise, experiment 4 has produced evidence to support the prediction that individuals with higher cognitive ability are able to and do carry out a search for counterexamples on PS problems under necessity instructions; and also support for the search for alternatives on PW problems under possibility instructions. The two experiments reported in the next chapter will extend the range of conditional inference experimental materials, to investigate the inferences people are prepared to make on conditional inference problems with realistic everyday content, where one event causes the other.

The experiments that have been reported so far have adopted abstract or non-thematic content, alongside clear logical structures, and the introduction of everyday materials enabled an investigation into the extent to which reasoning behaviours are influenced by the availability of other causes, and the relationship between these and the logical properties of the conditional argument forms. A second experiment will look at the influence of content on the inferences that participants are prepared to make, when

there are no logical properties to the inferences. In each of these experiments, both necessity and possibility instructions will be used, and a measure of cognitive ability will also be taken look at how these interact with endorsement rates and reasoning times.

Chapter 5

The search for counterexamples & alternative models on causal conditional inference tasks

The experiments that have been reported up to this point in the thesis have adopted reasoning tasks with clear logical structures, and either abstract or spatially related content, to look at whether reasoners search for counterexamples or alternatives when evaluating a conclusion. The results from analysis of the conditional inference data, and the second transitive inference experiment, revealed clear evidence to support the search for both counterexamples and alternatives; although there was limited support from the first transitive inference experiment, and no evidence from the abstract syllogistic reasoning experiment reported in chapter 2. However, it was also clear that individuals made a substantial amount of errors, which is the motivation for the choice of everyday materials in the two experiments reported in this chapter.

Research has consistently shown that human reasoning performance is heavily influenced by the content or subject matter, and reasoning with identical formal properties but different subjective content, frequently produces different levels of

performance. The two experiments use *'if p then q'* conditional inferences of the type used in experiment 4, but this time they are constructed with everyday realistic content rather than abstract terms, to form *causal* conditional statements with few *and* many alternative causes. This enables the investigation not only of whether a search for counterexamples or alternative models takes place, and the degree to which this is mediated by ability, but also the extent to which the number of alternative causes to the rule under consideration influences this search. In order to clarify the terms, *alternative models* is used to describe the search for other models under instructions of possibility, and *alternative causes* is used to indicate the number of other causes to the rule.

5.1 Introduction to experiments 5 and 6

As discussed in chapter 4, conditional reasoning involves drawing inferences about situations in which the occurrence of one event is contingent on the occurrence of another. The standard conditional inference tasks which have been frequently used, particularly in earlier reasoning research, generally looked at the factors which influence performance on the logical validity of 'if p then q' arguments for valid MP and MT inferences, and AC and DA fallacies (see table 5.1). Typically, endorsement rates are close to 100% for MP arguments, for MT arguments they are around 74% or less, but often as many as 80% of people are prepared to erroneously endorse AC and DA arguments. There is however much variation between studies, particularly with AC and DA arguments, although this is widely believed to be because of the different methodologies that have been used.

Table 5.1
The four conditional arguments

MP	if p then q, p, therefore q
MT	if p then q, not q, therefore p
AC	if p then q, q, therefore p
DA	if p then q, not p, therefore q

Although early conditional reasoning research with abstract materials has made a significant contribution to our understanding of logical competence, in everyday life we are required to reason with meaningful and content-rich materials, which draw on our prior knowledge. For instance take the statement, *if butter is heated it will melt*, followed by the statement *the butter did not melt*. Most adult individuals would be inclined to conclude that the butter had not been heated, by accessing their prior knowledge that heating butter causes it to melt. This conclusion is both logically correct, and congruent with our understanding of the relationship between heat and butter, although, less than 74% of people are prepared to endorse this MT argument when it is presented with abstract content.

Seminal work by Byrne (1989) found that it was easy to manipulate the willingness of participants to endorse conclusions to conditional inference reasoning problems, by increasing or decreasing the availability of additional information within each conditional statement. The effect of manipulating a response became known as the suppression effect, when the number of inferences that that reasoners are prepared to make, decreases with the availability or number of counterexamples. For example, when the additional information shown in italics was added to the causal conditional

statement shown below, reasoners were less inclined to accept the inference '*the light will go on when the fridge was opened*':

> If you open the fridge, then the light inside will go on
> *If the light bulb is working, then the light inside will go on*
> Somebody opens the fridge

Early methodology which was used to produce the *suppression effect,* presented the additional information after the conditional statement, but prior to the conclusion under evaluation. However, later research showed that the effect could also be achieved if alternative causes were not explicitly mentioned, but could merely be retrieved from the knowledge that people had, based on their everyday experiences of the world. The factors affecting the retrieval of alternative causes were explored by a frequently cited programme of research carried out by Cummins et al. (1995; 1991). When a pre-test was carried out in which participants were asked to generate counterexamples to a variety of causal conditional statements, Cummins et al. (1995; 1991) found that the retrieval process was sensitive to two factors: alternative causes and disabling conditions. Take for example the following causal conditional statement:

> If the brake is depressed, the car will slow down

An *alternative cause* suggests that there may be a reason for the car slowing down, other than the one cited in the rule under consideration, which produces the effect mentioned; for example *running out of fuel*, or *climbing a steep hill*. On the other hand a *disabling condition* prevents the effect from occurring despite the presence of the cause; which may be *ice on the road*, or *fractured brake lines*. When conditional inference problems were constructed to test for the effects of *alternative causes* or *disabling*

conditions, Cummins et al. (1995; 1991) found that the valid inferences MP and MT were more likely to be made for conditionals with few *disablers*, and the invalid AC and DA inferences were more likely to be made for conditionals with few *alternative causes*. Consider for instance the following example and narrative, taken from recent research (Evans, Handley, & Neilens, 2010):

If global warming continues (p) then London will be flooded (q)

A participant might think that global warming will cause a rise in sea levels and that London being quite low lying, will be at risk of flood. But then they may think of a disabling condition, such as expensive flood barriers that a major city would invest in the time to consider the problem. On this basis they may decline the MP inference that London will necessarily be flooded. When offered the AC inference however, they might well accept it. It seems unlikely that London would be flooded except by such a major environmental disaster. So for this person, the statement would be high in disablers, but low in alternatives (Evans et al., 2010, p. 894).

Other related work on causal conditionals carried out by Thompson (1994, 2000) used the terms *necessity* and *sufficiency* instead of alternative causes or disablers, and asked participants to rate sentences such as the above example for (a) is it necessary for *p* to happen in order for *q* to happen, or (b) *p* happening is enough to ensure that *q* will happen. There were strong effects of perceived *sufficiency* on the acceptance of MP and MT conditionals, and strong effects of perceived *necessity* on AC and DA conditionals.

So, when considering the causal statement 'if global warming continues then London will be flooded', the statement is rated as *low sufficiency* and *high necessity*, producing similar outcomes to those reported by Cummins et al. (1995; 1991), in that there would be low rates of MP and MT endorsements, but high rates of AC and DA endorsements.

An alternative view which has been used to explain the processes which underlie the conditional inferences that people make is that people treat them as probabilistic statements (Liu, Lo, & Wu, 1996; Oaksford & Chater, 1998, 2001). When making judgements of probability, the theory proposes that participants calculate the probability that exceptions to the event will occur; which will in turn affect the likelihood of individuals endorsing a given conclusion. Liu et al. (1996) rephrased conditionals of the nature used by Cummins et al. (1995; 1991), and asked participants to rate perceived probabilities; this resulted in an increase in correct responses with the perceived probability of q, given p, for each of the four forms of conditional arguments: MP, MT, AC and DA.

Comparisons were made between inferential reasoning and probabilistic reasoning (Markovits & Handley, 2005) using identical *if then* statements with everyday content. The participants were either asked to rate the probability on a likert scale, or respond either *yes* or *no* to question, for instance *Michael's dog has fleas, is it certain that Michael's dog will scratch constantly*? In a second experiment problems were presented either under instructions of logical necessity, or instructions asking 'what is the probability of' on a scale of between 0 and 100%. The results suggested that deductive and probabilistic inferences are not structurally similar, and highlighted the difficulties

in try to consider these two systems in terms of them both having a single underlying process.

5.1.1 Mental Model Theory

The mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991) predicts that when reasoners are presented with a causal conditional inference problem, an initial internal model of the information is constructed. In order to reject invalid AC and DA inferences under instructions of logical necessity, participants must flesh out the full set of models to indicate that *p* may *not* occur in the presence of *q,* and *q* may *not* occur in the presence of *not* p (disconfirming DA). Therefore when there are *few* alternative causes, reasoners are less likely to find them, and are more likely to produce a logically incorrect response under instructions of necessity, for instance take the following example where there are few causes for the consequent other that the one mentioned:

If butter is heated then it will melt

The butter was heated (p)   The butter melted (q)

The conclusion the *butter was heated* is frequently endorsed, because there are few (if any) other events that would cause the butter to melt. Consider in contrast the following example, where there are many alternative causes scenario. The conclusion that *the stone was kicked* is less frequently endorsed, as there are many other causes that might be responsible for the movement of the stone (see next page)

If the stone is kicked it will move

The stone was kicked (p)	The stone moved (q)
The stone was thrown (p)	The stone moved (q)
A dog picked up the stone (p)	The stone moved (q)
There was an earth tremor (p)	The stone moved (q)

5.1.2 Dual process theories

There are also theoretical implications within the framework of dual process theories, in terms of the influence of the number of alternative causes to the conditional rule under consideration, which have more recently been investigated (Evans, 2010). In conditional reasoning research, interest in dual process theories has been increasing in recent years to reflect research about the effect of the number of causes that can be accessed from our everyday knowledge of the world. Within a dual process theory framework, the intuitive Type 1 response is to endorse a conclusion because it is consistent with underlying beliefs, unless the less intuitive and more deliberate Type 2 process takes over to decouple the logical properties of a conditional from the reasoners stored knowledge of the world[19].

In a recently published study, which was carried out after experiment 5 in this thesis was planned and executed, Evans et al. (2010) used both *necessity* instructions (similar to those used in this thesis) and *pragmatic* instructions; together with a measure of cognitive ability. The task under *pragmatic* instructions concerned the ability of people

---

[19] Type 1 (system 1) is fast, automatic and intuitive, whereas Type 2 (system 2) is slow and more considered - see Chapter 1, and Evans (2010).

ability to reason about real-life situations as opposed to a test of logical reasoning; therefore no reference was made to assuming the premises to be true, or to logical necessity. Participants were asked to rate their degree of belief in the conclusion given, thus encouraging reasoners to focus on the believability rather than the validity of an event. The responses to the *pragmatic* instructions produced responses on a scale, rather than yes/no binary responses, which was in a similar form to the data collected by (Liu et al., 1996). The content of the conditional reasoning problems introduced a conflict between logic and belief under necessity instructions, enabling comparisons to be made between logical instructions and pragmatic instructions.

The results clearly show that higher ability participants were significantly less influenced by the everyday content of the reasoning problems when presented under strict logical instructions, but under pragmatic instructions (supposing the following …….. to what extent would you believe that) this effect disappeared. This led to the conclusion that that ability and specific instructions are required in order for people to reason in an abstract and decontextualized manner, despite the content being linked to our everyday experiences.

5.1.3 Reasoning times

Although very few studies have looked at whether the retrieval of few or many alternative causes has an effect on the time course of conditional reasoning, it would seem a reasonable assumption that due to a more extended search process, within the framework of the mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991), reasoning times would be longer when there are *many* alternative causes than

when there are *few* alternative causes. The implications of this are that the reasoning times would be longer for inferences such as *if the stone is kicked it will move* than for inferences with few alternative causes such as *if the butter is heated it will melt*. One study which did successfully collect latency data (De Neys et al., 2002) hypothesised that reasoning with standard argument form conditionals would produce longer reasoning times when there were many alternative causes, than when there were few alternative causes. Consistent with this hypothesis, participants took significantly longer on MP and AC argument forms, but not for MT and DA argument forms. The lack of support for MT and DA arguments forms was thought to be because of a trade off between the search process, and the need to process negations; MT (if p then q, not q) and DA (if p then q, not p) arguments.

5.1.4 Aims and rationale for experiment 5

The approach to Experiment 5 was novel, in that it combines the presentation of causal conditional problems with few and many alternatives causes, under instructions of logical *necessity* and instructions of *possibility* (as previously used in this thesis), to collect endorsement rates *and* reasoning times. A measure of cognitive ability was also taken, and recent work by Verschueren et al. (2005) suggests that participants with greater cognitive capacity are more likely to retrieve, and selectively use counterexamples to reject invalid (AC and DA) conditional inferences, which is consistent with the literature. Furthermore, the effect of ability has been shown to increase with development throughout childhood (Janveau-Brennan & Markovits, 1999).

The aim was to investigate the effect that the number of alternative causes to the rule under consideration had on judgements of *necessity* and *possibility*, when evaluating conclusions to the four different problems types used in experiments 1 – 4. The development of Necessary, PS, Impossible and PW problems was facilitated by the use of argument forms with standard direction conclusions, and argument forms with opposite direction conclusions. This was a similar method to that used for experiment 4 (conditionals with abstract content), but with the addition of an extra level in terms of *few* and *many* alternative causes. The standard conclusion inferences lead to the composition of Necessary and PS problem types; under necessity instructions the conclusion on MP and MT arguments forms is logically valid, but not under possibility instructions, and are referred to as Necessary problems. The standard conclusion AC and DA arguments are not logically valid under necessity instructions, but they are possible, and are referred to as PS problems.

The opposite conclusions on MP and MT inferences are impossible under both types of instruction, as logically the conclusion is inconsistent with all models, and these are referred to as Impossible problems. The AC and DA problems with opposite conclusions are not *necessary*, but they are *possible,* because although logically they are not consistent with an initial model of representations of the premises, there is a model or models which does not support the premises, and these are referred to as PW problems. An example of a PW problem with many alternative causes is; *if Simon cuts his finger it will bleed*, given that *Simon's finger is bleeding*, the first model that comes to mind is that *Simon cut his finger*, when in fact his finger may be bleeding because *he caught it on a bramble*, *grazed it on gravel*, or *was bitten by a hamster.*

5.1.5 Predictions for experiment 5

The following predictions are made on the basis of previous research (Cummins, 1992, 1995; Evans et al., 1999; Thompson, 1994, 2000), and the general assumptions of the mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991).

1. There will be more endorsements under possibility instructions than necessity instructions, because lots of everyday events are possible, but not necessary.

2. If participants search for counterexamples under necessity instructions: when there are many alternative causes there will be fewer endorsements of PS problems than Necessary problems, because other causes are more readily available because the first model under consideration supports the conclusion.

3. If participants search for alternative models under possibility instructions: when there are many alternative causes there will be more endorsements of PW problems than Impossible problems, because other causes are more readily available, given that the first model under consideration does not support the conclusion.

4. If participants search for counterexamples or alternative models: under necessary instructions the high ability group will produce more correct responses than the low ability group for problems with few alternative causes, and the high ability group will also produce more correct responses to PS problems under necessity instructions and PW problems under possibility instructions; this is because they are better at searching for counterexamples or alternative models.

5. High ability participants will be quicker reasoners, but may take proportionately longer on problems requiring a search for counterexamples or alternative models; more particularly PS inferences under necessity instructions, and PW inferences under possibility instructions.

5.2 Method for experiment 5

Design

This experimental study was carried out using a within-subjects design, when initially participants completed an AH4 cognitive ability test. This was followed by the causal conditional problem solving task, where participants were presented with one block of 32 problems with everyday content under necessity instructions, and one block of 32 of the same problems under possibility instructions, the order of which was counterbalanced to minimise order effects.

Participants

A total of 60 undergraduate students from the University of Plymouth took part in the study, in return for either payment or course credit. The sample consisted of 10 males and 50 females with a mean age of 21 years, and they were all native English speakers. No participants were dyslexic, or had received formal training in logic.

Materials and procedure

The procedure adopted was similar to that used in the previous 5 experiments, when participants were run in groups of between 2 and 6 in a laboratory containing several computers. Each participant was seated at their own workstation, to avoid distraction.

Cognitive Ability Test

Participants completed Parts I and II of the AH4 Test of Cognitive Ability (Heim, 1968), which was administered in accordance with the test instructions and followed the procedure used in experiment 1. Question booklets and answers sheets were collected by the experimenter before moving on to the inference task.

Causal conditional inference task

The 32 inference problems consisted of 8 problems in each of the four categories used in previous experiments: Necessary, PS, Impossible and PW, and within each category there were two statements with few alternative causes, and two statements with many alternative causes. In order to provide a balanced mix of argument structures (MP, MT, AC and DA) there were two different argument forms within each category, which were equally balanced for few and many alternative causes.

Initially, two sets of premise pairs were identified (see table 5.2) each of which contained two statements with many potential alternative causes, and two statements with few potential alternative causes. Presentation of the two sets was counterbalanced so that half of the participants received the content from set A for Necessary and PS problems and set B content for Impossible and PW problems, and the other half had the presentation order reversed. The two sets of problem content and structure are presented in appendix 5A and appendix 5B.

Table 5.2
The two sets of premise content for experiment 5

	Set A
Few	if butter is heated, it will melt
	if Simon cuts his finger, it will bleed
Many	if the stone is kicked, it will move
	if the brake is pressed, the car will slow down

	Set B
Few	if the paper clip touches the magnet, it will stick to it
	if water is frozen, it will become ice
Many	if the window is open, the room will be cold
	if the mug is dropped, it will break

A computer with a 15" monitor screen was used to present the problems with the computer program. The keyboard was adapted to include *yes* and *no* keys, which were systematically counterbalanced, so that half the participants had the *yes* key on the left of the keyboard and the *no* key on the right, while the other half had these positions reversed. The two sets of written task instructions which included examples of the screen layout, were printed on A4 paper, and were similar to those used in the previous experiments. These related to whether problems were being evaluated for either necessity correctness or possibility correctness. Examples of the screen layouts are shown in table 5.3, and a complete set of instructions is presented in appendix 5C and appendix 5D.

The instructions were distributed (necessity or possibility) for the first block of problems and after a short reading period, participants were given the opportunity to ask questions on any points that they were less clear about.

Table 5.3
Screen layouts included in task instructions

Screen 1

> Given:
>> If the stone is kicked then it will move
>> The stone was kicked

Screen 2

> Given:
>> If the stone is kicked then it will move
>> The stone was kicked
> *Is it necessary that*
>> The stone moved

Screen 1

> Given:
>> If Simon cuts his finger then it will bleed
>> Simon cut his finger

Screen 2

> Given:
>> If Simon cuts his finger then it will bleed
>> Simon cut his finger
> *Is it possible that*
>> Simon's finger bled

The participants were also told that they should ask the experimenter for the second set of instructions (necessity or possibility) as soon as a message appeared on the screen, and reminded that at the start of each block there were two practice questions.

Participant responses, *yes* or *no*, were recorded by the program, together with the time taken to indicate understanding of the problem (screen 1) and the time taken to complete the reasoning process (screen 2). These were saved to disc.

5.3 Results for experiment 5

The AH4 test sheets were scored in accordance with the test instructions, when one mark was given for each correct answer. There was a significant positive correlation between Parts I and II ($r = .49$, $p < .001$), therefore scores from both parts were added together to give an overall general ability score for each participant. The observed mean for participants was 102.78 ($SD = 12.82$), which was higher than to the available norm of 96.36 ($SD = 15.01$) for university students (Heim, 1968), higher than experiments 2 and 3, lower than experiment 1, and substantially lower than experiment 4. The sample was divided into high and low cognitive ability groups, on the basis of a median split on the AH4 test scores; cases below the median of 105 were classified as low ability and those above the median were classified as high ability. None of the participants recorded a score of 105.

All participants evaluated conclusions under both necessity instructions and possibility instructions. The first dependent variable was the mean percentage endorsement rates, i.e. the number of *yes* responses. The second dependent variable was the time course of the reasoning process; that is to say premise processing and response times together.

These were totalled for each problem type and instruction group to produce a mean reasoning time (in milliseconds). The results from the endorsement rate data are reported first; followed by the results from the reasoning time data. The inference rates and reasoning times which are further broken down into argument forms can be found in appendix 5E. All ANOVA tables for experiment 5 are shown in appendix 5F.

5.3.1 Conclusion endorsement rates

Necessary and PS problem types

The mean percentage endorsement rates for Necessary and PS inference problems are shown in table 5.4, broken down by instruction, problem type, alternative causes, and ability. The cells for the low ability group and the high ability group represent the mean percentage endorsement rates for the responses from 30 participants.

A 2 (instruction) x 2 (problem type) x 2 (alternative causes) x 2 (ability) mixed ANOVA test revealed a main effect of instruction [$F(1,58) = 32.14$, $p < .001$, $\eta_p^2 = .36$], reflecting higher endorsement rates when problems were presented under possibility instructions, and a main effect of problem type [$F(1,58) = 16.79$, $p < .001$, $\eta_p^2 = .23$], whereby Necessary problems were endorsed more frequently than PS problems. These effects which are consistent with the previous experiments and the literature, suggest that reasoners understood the differences between making judgements of necessity and judgements of possibility, and also reflects a tendency to endorse more valid (MP/MT) conclusions than invalid (AC/DA) conclusions.

Table 5.4
Mean percentage endorsement rates for experiment 5, on Necessary and PS inferences (N = 60, SD in brackets)

	N (few)		N (many)		PS (few)		PS (many)	
Necessary								
Low	82	*(25.37)*	85	*(22.36)*	80	*(28.16)*	64	*(40.83)*
High	90	*(16.87)*	92	*(13.67)*	80	*(36.20)*	69	*(39.21)*
M	86	*(21.77)*	88	*(18.68)*	80	*(32.15)*	67	*(39.77)*
Possibility								
Low	98	*(7.63)*	99	*(4.56)*	98	*(7.62)*	95	*(10.17)*
High	99	*(4.56)*	98	*(7.63)*	98	*(7.62)*	98	*(6.34)*
M	98	*(6.29)*	98	*(6.29)*	98	*(7.56)*	97	*(8.57)*

There was also a significant difference between few and many available alternative causes [$F(1,58)$ = 5.89, p < .05, η_p^2 = .09], whereby problems with few available alternative causes were endorsed more frequently than those with many available alternative causes, which is consistent with the literature, and suggests that when there were fewer alternative causes available participants were less likely to find them. There was no main effect of ability [$F(1,58)$ = .92, p = .34].

There was a three-way interaction between instruction, problem type and available alternative causes [$F(1,58)$ = 18.32, p < .001, η_p^2 = .24], indicating a different pattern of responding under necessity instructions, (table 5.1) than under possibility instructions (table 5.2). Under necessity instructions participants were more likely to search for counterexamples in order to reject PS conclusions (to AC/DA) when there were many alternative causes.
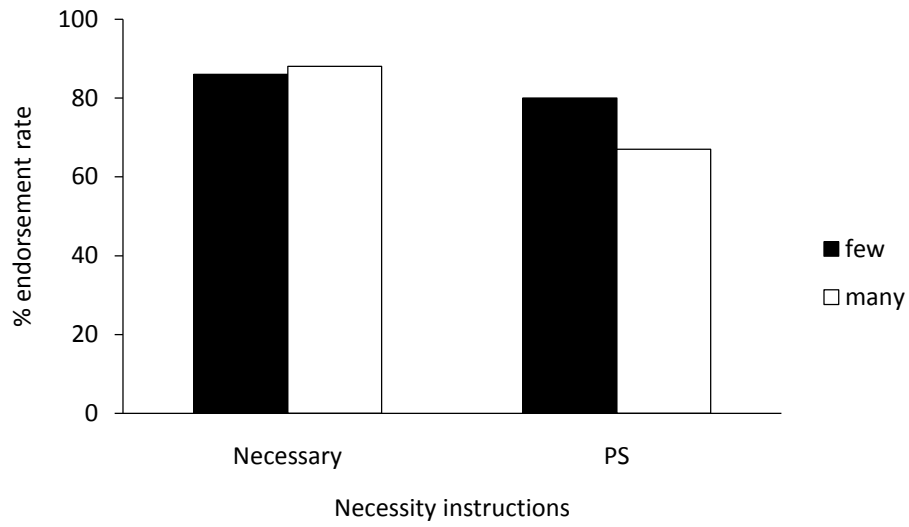
Figure 5.1. Mean percentage endorsement rates for experiment 5, on Necessary and PS inferences under necessity instructions
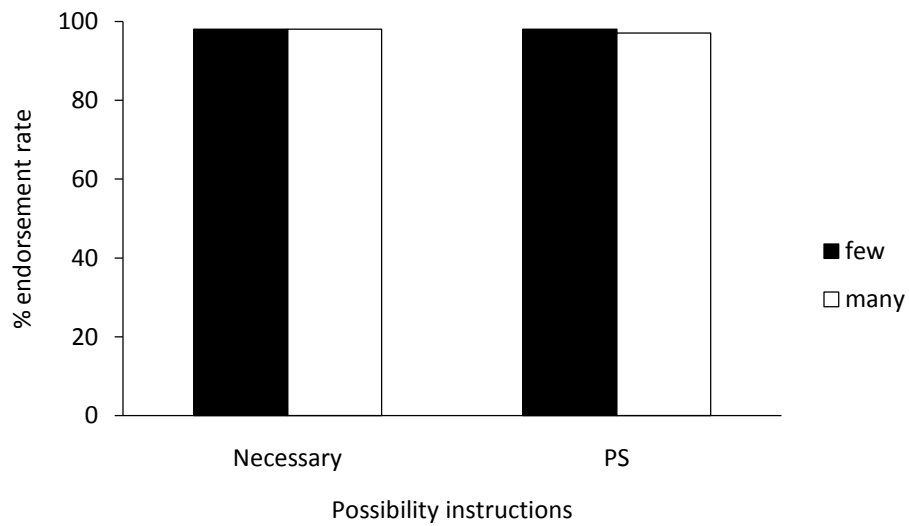
Figure 5.2. Mean percentage endorsement rates for experiment 5, on Necessary and PS inferences under possibiity instructions

Follow up within subjects *t*-tests indicate that inferences with many alternative causes show evidence of a search for counterexamples, in that there were significantly less endorsements of PS inferences than Necessary inferences, [$t(1,59) = 4.99, p < .001$]), but this effect was not present for inferences with few alternative causes [$t(1,59) = 1.57, p = .12$]. A within subjects *t*-test also indicated that there was a significant difference between PS inferences with few alternative causes and many alternative causes [$t(1,59) = 4.45, p < .001$]. The effects under instructions of necessity were not present under possibility instructions.

Impossible and PW problem types

The mean percentage endorsement rates for Impossible and PW problems are shown in table 5.5, broken down by instruction, problem type, alternative causes, and ability. The cells for the low ability group and the high ability group represent the mean percentage endorsement rates for the responses from 30 participants.

A 2 (instruction) x 2 (problem type) x 2 (alternative causes) x 2 (ability) mixed ANOVA test revealed a main effect of instruction [$F(1,58) = 25.49, p < .001, \eta_p^2 = .31$], reflecting higher endorsement rates when problems were presented under possibility instructions. This is consistent with both previous experiments and the literature, and reflects the fact that reasoners can differentiate between types of instruction. There was also a main effect of alternative causes [$F(1,58) = 44.02, p < .001, \eta_p^2 = .43$], when problems with few available alternative causes were endorsed less frequently than problems with many available alternative causes, which is consistent with past research, and confirms that the number of alternative causes increases the likelihood of

a given conclusion being endorsed by reasoners. There was no main effect of problem type [$F(1,58) = .73$, $p = .34$], and there was no main effect of ability [$F(1,58) = .84$, $p = .36$].

Table 5.5
Mean percentage endorsement rates for experiment 5, on Impossible and PW inferences (N = 60, SD in brackets)

	I (few)	I (many)	PW (few)	PW (many)
Necessary				
Low	15 *(22.61)*	27 *(32.12)*	7 *(13.02)*	8 *(18.74)*
High	2 *(9.13)*	7 *(25.37)*	11 *(21.46)*	6 *(20.43)*
M	8 *(18.22)*	17 *(30.42)*	9 *(17.23)*	7 *(19.46)*
Possibility				
Low	13 *(21.51)*	41 *(42.29)*	12 *(22.49)*	46 *(39.44)*
High	5 *(10.17)*	38 *(42.93)*	23 *(32.78)*	52 *(41.49)*
M	9 *(17.20)*	38 *(42.28)*	18 *(28.47)*	49 *(40.24)*

There were three significant interactions; the first of these was between instruction and problem type [$F(1,58) = 10.12$, $p < .05$, $\eta_p^2 = .15$]. Figure 5.3 suggests that this is because more conclusions were correctly accepted following a search for alternative models on PW problems under possibility instructions than under necessity instructions, and also more conclusions were endorsed on PW problems than Impossible problems under possibility instructions. These effects were confirmed by within subjects *t*-tests when significantly more PW problems were endorsed than Impossible under possibility instructions [$t(1,59) = 5.48$, $p < .05$]; and significantly more PW problems than Impossible problems were endorsed under possibility instructions [$t(1,59) = 2.14$, $p < .05$]. This replicated the results found in experiments 2, 3 and 4.
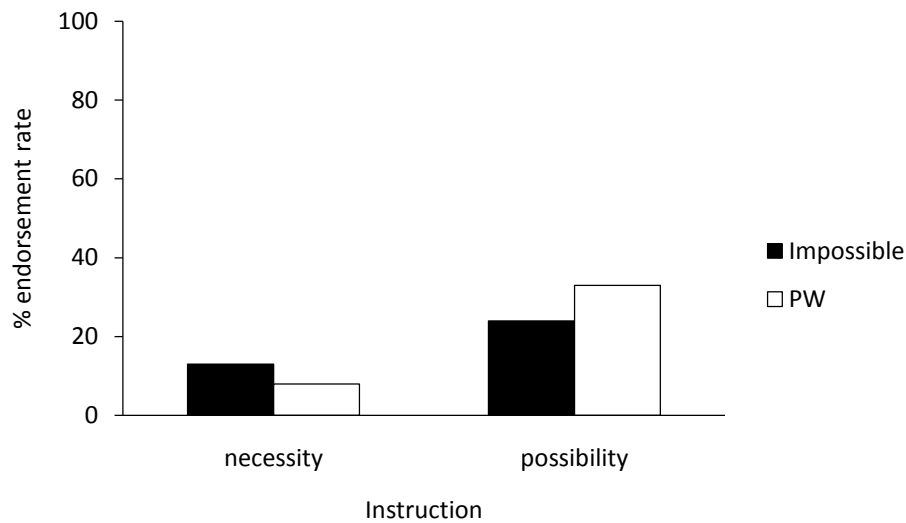
Figure 5.3. Mean percentage endorsement rates for experiment 5, on Impossible and PW inferences under necessity and possibility instructions

The second interaction was between instruction and the alternative causes [$F(1,58)$ = 35.36, p < .001, η_p^2 = .38], suggesting that under necessity instructions the number of available alternative causes did not affect the willingness of participants to endorse or reject conclusions, but under possibility instructions participants were more prepared to endorse problems with many available alternative causes than with few available alternative causes (see figure 5.4). The difference between few and many alternative causes under possibility instructions was confirmed by a follow up within subjects *t*-test, when problems with many alternative causes were endorsed significantly more often than problems with few alternative causes [$t(1,59)$ = 7.53, p < .001]. This pattern was predicted for PW problems; however there was no three-way interaction between instruction, alternative causes and problem type, suggesting that high rates of endorsement under many alternatives is present for both PW problems and Impossible problems.
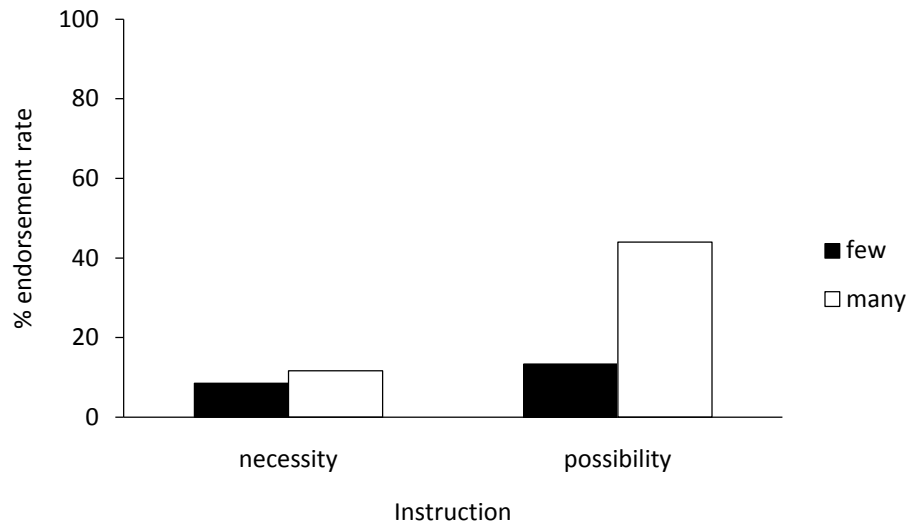
Figure 5.4. Mean percentage endorsement rates for experiment 5, on inferences with few and many alternative causes under necessity and possibility instructions
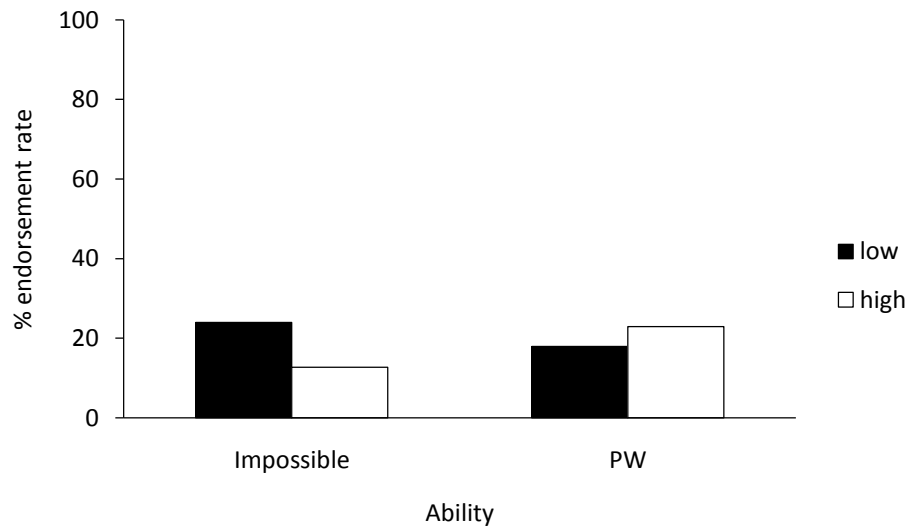
Figure 5.5. Mean percentage endorsement rates for experiment 5, on Impossible and PW inferences for low and high ability groups

The third interaction was between problem type and ability [$F(1,58) = 11.10$, $p < .05$, $\eta_p^2 = .16$] whereby participants in the high ability group performed better on Impossible problems, by rejecting them more frequently than the low ability group.

This was confirmed by a between subjects *t*-test when the high ability group endorsed significant less Impossible problems than the low ability group [$t(1,58) = 2.51$, $p < .01$]. Also, although the high ability group did not perform significantly better than the low ability group on PW problems, the effect was in the direction predicted [$t(1,58) = 1.29$, $p = .20$], and the high ability endorsed significant more PW problems than Impossible problems [$t(1,29) = 2.93$, $p < .005$] whereas the low ability did not [$t(1,29) = 1.77$, $p = .09$]. These effects are consistent with the literature.

5.3.2 Reasoning times

Necessary and PS problem types

The mean reasoning times for Necessary and PS problems are shown in table 5.6 which are broken down by instruction, problem type, alternative causes and ability. The cells for the low ability group and the high ability group represent the mean reasoning times for the responses from 30 participants, and are shown in milliseconds.

Table 5.6
Mean reasoning times in milliseconds for experiment 5, for Necessary and PS inferences (N = 60, SD in brackets)

	N (few)		N (many)		PS (few)		PS (many)	
Necessary								
Low	6742	*(2207)*	6675	*(1834)*	6798	*(2879)*	7032	*(2433)*
High	6811	*(3703)*	6573	*(2393)*	6554	*(2161)*	6661	*(2287)*
M	6776	*(3022)*	6624	*(2114)*	6676	*(2527)*	6846	*(2348)*
Possibility								
Low	6563	*(2651)*	6352	*(1703)*	6442	*(2192)*	7175	*(2519)*
High	6310	*(1966)*	6013	*(1894)*	5783	*(1432)*	5898	*(1383)*
M	6437	*(2318)*	6182	*(1794)*	6092	*(1869)*	6536	*(2115)*

A 2 (instruction) x 2 (problem type) x 2 (alternative causes) x 2 (ability) mixed ANOVA test was carried out, and there were no main effects of instruction [$F(1,58)$ = 3.36, p = .07], problem type [$F(1,58)$ = .06, p = .81], alternative causes [$F(1,58)$ = .10, p = .75], or ability [$F(1,58)$ = .91, p = .35]. There were no interactions.

Impossible and PW problem types

The mean reasoning times for Impossible and PW problems are shown in table 5.7 , broken down by instruction, problem type, alternative causes and ability. The cells for the low ability group and the high ability group represent the mean reasoning times for the responses from 30 participants, and are shown in milliseconds.

Table 5.7 *for experiment 5*
Mean reasoning times in milliseconds for experiment 5, on Impossible and PW inferences (N = 60, SD in brackets)

	I (few)		I (many)		PW (few)		PW (many)	
Necessary								
Low	9375	*(3773)*	8259	*(3286)*	8269	*(2825)*	8900	*(4770)*
High	7670	*(2210)*	7346	*(2678)*	8107	*(3906)*	7132	*(2417)*
M	8523	*(3184)*	7803	*(3007)*	8188	*(3213)*	8016	*(3950)*
Possibility								
Low	8358	*(2671)*	8148	*(3125)*	9171	*(2695)*	7515	*(3031)*
High	8156	*(3471)*	6773	*(1657)*	7477	*(2496)*	6925	*(2375)*
M	8258	*(3072)*	7461	*(2575)*	8324	*(2714)*	7220	*(2878)*

A 2 (instruction) x 2 (problem type) x 2 (alternative causes) x 2 (ability) mixed ANOVA test revealed a main effect of alternative causes [$F(1,58)$ = 13.15, p < .001, η_p^2 = .19], whereby problems with few alternative causes took longer than those with many alternative causes, and a main effect of ability [$F(1,58)$ = 4.07 p < .05, η_p^2 = .07] when

participants in the low ability group took longer than participants in the high ability group. There were no main effects of instruction [$F(1,58) = .88$, $p = .35$], or problem type [$F(1,58) = .11$, $p = .75$].

The main effect of ability is consistent with the literature, where it has been shown that high ability reasoners are generally quicker. However, the main effect of alternative causes suggests that participants were spending time looking for alternative causes when there were few available ones, which is inconsistent with findings reported by de Neys et al. (2002). The lack of a main effect for problem types may have been affected by the time taken to process negations. This is because Impossible problems consisted of MP and MT arguments with opposite conclusions, and PW problems consisted of AC/DA argument with opposite conclusions, which in this case would affect MP (if p then q, p, not q) and AC (if p then q, q, not p); although it would be expected that this would have affected each problem type equally.

There was a four-way interaction between instruction, problem type, alternative causes and ability [$F(59) = 9.58$, $p < .005$, $\eta_p^2 = .14$]. This was difficult to understand, and given that there was no clear interpretation, and that the latency effects in this thesis so far have been small and inconsistent; no attempt is made to draw any firm conclusions regarding this complex interaction.

5.4 Discussion for experiment 5

Experiment 5 used causal conditionals with everyday content, to explore the influence of the number of alternative causes that could be retrieved from our everyday knowledge of the world, on endorsement rates and latencies under necessity and

possibility instructions. A measure of cognitive ability also allowed comparisons to be made for performance and latencies between low and high ability groups. Consistent with experiment 1 - 4, there was clear evidence that participants understood the difference between instructions of necessity, and instructions of possibility, as there were more endorsements of possibility than of necessity. This supports previous research (Evans et al., 1999).

There was strong evidence to suggest that participants were accessing their knowledge of the world before responding, in terms of endorsement rates. Congruent with the literature (Cummins et al., 1991; Evans et al., 2010), there were more endorsements of inferences with few alternative causes than many alternative causes, on logically valid inferences (Necessary) and PS inferences where the first model supports the conclusion. There was also a novel finding that has not been reported previously in the literature, in that the number of possible causes led to inferences with few alternative causes being endorsed less frequently than inferences with many alternative causes, inferences with a logically Impossible conclusion, and indeterminate inferences where the first model under consideration does not support the conclusion (PW). These findings were not affected by ability. An example of inference content with few alternatives causes is *butter melting*, in that there are few causes other than it being heated which would cause butter to melt; on the other hand if a *room is cold*, in addition to *the window being open,* there may be many reasons why the room is cold, such as *an old heating system*, *extreme weather*, or *lack of insulation*.

The main analysis and comparisons were carried out between Necessary and PS problem types to explore whether reasoners searched for counterexamples on PS problems under necessity instructions, and whether reasoners searched for alternative models on Impossible and PW problems under possibility instructions. In each case a search for counterexamples or alternative models was necessary in order to provide the correct response to the conclusion under evaluation.

The patterns of behaviour were different, in terms of finding evidence to support the search for counterexamples under necessity instructions, and alternative causes under possibility instructions. There was a clear difference between endorsement rates for Necessary inferences and PS inferences, when there were more rejections of conclusions to PS inferences than to Necessary inferences. Consistent with experiments 1 - 4, inferences with necessarily correct conclusions (Necessary) were endorsed more frequently than inferences with indeterminate conclusions (PS), which suggests that reasoners were aware of the logical framework of the inferences. However the difference between Impossible problems and PW problems was not significant, indicating a lack of discrimination between inferences types.

Looking first at Necessary (MP/MT) inferences with logically valid structures and invalid PS (AC/DA) inferences. There was a strong indication that more reasoners were carrying out a search for counterexamples when there were *many* alternative causes under *necessity* instructions, than they were when there were *few* alternative causes. The theoretical implications of this are that when there are *many* alternative causes to an event, under *necessity* instructions, reasoners are more likely to

successfully search for counterexamples as proposed by the mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991). For instance take the invalid AC conditional 'if the brake is pressed the car will slow down', and the knowledge that the car slowed down, rather than accepting that the brake was pressed, participants appeared to think of other causes, for example *running out of fuel*, *climbing a hill*, *leaving the handbrake on*, *overheating, having a broken fan belt*. However, the likelihood of a search taking place was not affected by ability.

The analysis for Impossible and PW problems was less clear in terms of the relationship between the number of alternative causes and the search for alternative models; but there was support for the prediction that that reasoners searched for alternative causes on PW problems under possibility instructions, although this was not mediated by the number of alternative causes or by ability. Therefore, participants were equally likely to search for alternative causes for the *water becoming ice* other than it being *frozen*; and for the *mug breaking* because it was *dropped*.

The high ability group endorsed significantly more PW problems than Impossible problems, which was expected since Impossible problems are not *necessary* or *possible*, but PW problems are *possible* but not *necessary*. However surprisingly, the low ability group endorsed less PW problems than Impossible problems, which again was not affected by the number of causes.

There were particularly high endorsement rates for Impossible problems when there were *many* alternative causes, under possibility instructions. For instance, given the statement 'if the window is open, the room will be cold', and the information that the

window was open, many participants were willing to access their knowledge of the world to find reasons why the room was not cold; perhaps because *the heating was on*, *the curtain were drawn*, or *the weather was hot*. On the other hand given the statement 'if the paperclip touches the magnet, it will stick to it', and the information that the paper clip and magnet were stuck together, this was not the case. Interestingly, the high ability group did not override this effect, and the high number of erroneous responses to PW problems by both ability groups suggests that possibility instructions led participants to question the causal link more than when problems are presented under necessity instructions.

When looking at the results in relation to past research, Cummins et al. (1995; 1991) found that invalid AC and DA inferences were more likely to be made for conditionals with few available causes, which in this case are PS problems. There was clear evidence of this when 80% of inferences were erroneously endorsed when there were *few* alternative causes, as opposed to 67% when there were *many* alternative causes, and the results suggest that a successful search for counterexamples was carried out on inferences with *many* alternative causes, but was either not carried out or was unsuccessful on those with *few* alternative causes.

Clearly, our knowledge of the world plays an important role in the inferences we are prepared to make, although this was not as affected by ability as might be expected. In contrast to previous research by Evans et al. (2010), there was no evidence to suggest that cognitive ability played a role in performance under either necessity instructions or possibility instructions, on inferences that are logically valid or supported by an

initial model. However under possibility instructions the high ability group were better performers on Impossible inferences and PW inferences where the conclusions was not supported by the initial model

The next and final study of this thesis, simplifies experiment 5, to look at the consequents from specific antecedents under necessity and possibility instructions, where the emphasis was on content rather than the relationship between content and the logical structure.

5.5 Introduction to experiment 6

The previous study illustrated how different contents, in terms of the number of alternative causes that can be accessed from our everyday knowledge of the world, lead to the construction of certain models in order to evaluate a given conclusion relating to the rule under consideration. This final study adopts a simplified inference task, to look at the consequents from specific antecedents under different conditions, again using *necessity* and *possibility* instructions, a measure of cognitive ability, and the collection of inference latencies. The focus here is on possible disabling conditions (Cummins, 1995; Cummins et al., 1991; De Neys, Schaeken, & d'Ydewalle, 2003), so given the conditional '*if the brake is depressed, then the car will slow down*', possible disabling conditions are: a fractured brake lines, icy road conditions, or accelerating at the same time. Therefore 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 to slow the car down, and there are other conditions that need to be satisfied.

Unlike the preceding experiments reported in this thesis, and the practice most commonly adopted in studies of conditional inference, syllogistic reasoning, and transitive inference; responses to the tasks in experiment 6 do not require deductive reasoning skills, but merely a response that reflects the general perceptions people have of a given scenarios, by accessing their knowledge about the world.

The inference structure of the conditionals are shown in table 5.8, using abstract examples; but in order to ensure that the associative strength was in line with the inference category for all of the inferences i.e. PS inferences were highly likely, a pilot study was carried out to identify the specific antecedents and consequents.

Table 5.8
Inference structures for experiment 6

Type	Inference	Rationale
Necessary	If p, then q	The antecedent will almost certainly lead to the consequent
PS	If p, then q	The antecedent is highly likely to lead to the consequent, so individuals need to find a disabling condition in order to reject inferences under necessity instructions
Impossible	If p, then q	The antecedent will not lead to the consequent
PW	If p, then q	The antecedent is highly unlikely to lead to the consequent, so individuals need to find an enabling condition in order to endorse PW inferences under possibility instructions

5.6 Pilot study for experiment 6

The purpose of the pilot study was to select a reliable set of 32 simple inferences with an equal number in each of the four categories (Necessary, Impossible, PS and PW). A bank of 64 inferences was generated, where the causal inference broadly fell into one of four inference type categories (Necessary, PS, Impossible, PW). These were based on the relationship between p (the antecedent) and q (the consequent).

Participants

The participants were run in groups of 5, and were 20 undergraduate students from the University of Plymouth, who took part in the pilot study in return for either payment or course credit. The sample consisted of 6 males and 14 females with a mean age of 22 years, and they were all native English speakers.

Task materials and procedure

Table 5.9
Task instructions for pilot study on experiment for experiment 6

Read the sentence at the top of the page. Assuming this statement is true, please indicate the probability that the conclusion given in the second sentence is also true. This should be done by placing a mark on the scale in the appropriate place, where <u>you</u> think it should be, for instance:

If James jumps into the river
he will get wet

impossible	20%	40%	60%	80%	true
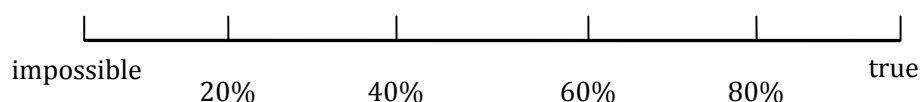
The 64 inferences were presented to participants in a randomised list with three inferences on each page; a set of six point likert scales, numbered 1 – 64 which again had three on each page; and the written task instructions (see table 5.9).

The mean ratings were calculated for all 64 inferences, when proposed Necessary inferences such as '*if it rains heavily, the streets will get wet*' with a mean score of 90% were retained for the main study, while inferences such as '*if a ruler is used, the line will be straight*' (mean score of 85%) were discarded. A full set of inferences used in the selection process are presented in appendix 5G. The selection criteria for identifying the inferences for use in experiment 5 are shown in table 5.10.

Table 5.10
Selection criteria for inferences used in experiment 6

Necessary	At least 50% of responses 100%	Mean response ≥ 90%
PS	At least 50% of responses between 50% and 80%	Mean response was between 60% and 85%
Impossible	At least 50% of responses 0%	Mean response ≤ 10%
PW	At least 50% of responses between 20% and 50%	Mean response was between 15% and 40%

5.6.1 Predictions for experiment 6

There are a number of predictions that can be made about the willingness of participants to endorsement or reject conclusions, which are based on the assumption that prior knowledge concerning disabling conditions for a given scenario, will affect whether or not a search is carried out for counterexamples or alternative models.

1. There will be more endorsements of possibility than of necessity, as although on PS and PW problems it cannot be concluded that the conclusion under consideration is necessary, it can be concluded that it is possible.

2. If reasoners consider disabling conditions for the scenario: under necessity instructions there will be higher endorsement rates for Necessary inferences than for PS inferences, because whilst PS inferences are not necessary, they are possible.

3. If reasoners consider disabling conditions for the scenario: under possibility instructions there will be higher endorsement rates for PW inferences than for Impossible inferences, because Impossible inferences are not possible under both types of instruction, but PW inferences are possible (but not necessary).

4. As the experiments reported so far have failed to find supporting evidence from the latencies for the search for counterexamples or alternative models, the only prediction that is made is that high ability participants will respond more quickly.

5.7 Method for experiment 6

Design

This experimental study was carried out using a within-subjects design, when initially participants completed an AH4 cognitive ability test. This was followed by the inference task, where participants were presented with one block of 32 randomised inferences under necessity instructions, and one block of 32 randomised inferences under possibility instructions, the order of which was counterbalanced to minimize order effects.

Participants

A total of 60 undergraduate students from the University of Plymouth took part in the study, in return for either payment or course credit. The sample consisted of 15 males and 45 females with a mean age of 24 years, and they were all native English speakers. No participants were dyslexic.

Materials and procedure

Using a similar procedure to the experiments reported in chapters 2, 3 and 4, participants were run in groups of between 4 and 7 in a laboratory containing several computers. Each participant was seated at their own workstation, to avoid distraction.

Cognitive Ability Test

Participants completed Parts I and II of the AH4 Test of Cognitive Ability (Heim, 1968), which was administered in accordance with the test instructions and followed the procedure used in experiment 1. Question booklets and answers sheets were collected by the experimenter before moving on to the inference task.

Causal inference task

The set of 32 one statement inferences identified in the pilot study, consisted of 8 simple inferences in each of four categories: Necessary, PS, Impossible, and PW; these were evaluated first under necessity instructions and then under possibility instructions, or vice versa. Examples of the inferences used are shown in table 5.11, and the full set off inferences can be found in appendix 5H.

Table 5.11
Examples of inferences for experiment 6

Necessary	If it rains heavily, the streets will get wet
PS	If a baby is hungry, he will cry
Impossible	If oil is added to water, they will mix
PW	If the dog falls into the canal, she will drown

A computer with a 15" monitor screen was used to present the inferences, with the computer program. The keyboard was adapted to include *yes* and *no* keys, which were systematically counterbalanced, so that half the participants had the *yes* key on the left of the keyboard and the *no* key on the right, while the other half had these positions reversed.

The two sets of written task instructions which included examples of the screen layout, were printed on A4 paper, and were similar to those used in the previous experiments. These related to whether problems were being evaluated for either necessity correctness or possibility correctness. Examples of the screen layouts are shown in table 5.14, and a complete set of instructions is presented in appendix 5I and appendix 5J.

The instructions were distributed (necessity or possibility) for the first block of inferences and after a short reading period, participants were given the opportunity to ask questions on any points that they were less clear about. The screen layouts are shown in table 5.15. The participants were also told that they should ask the experimenter for the second set of instructions (necessity or possibility) as soon as a

message appeared on the screen, and reminded that at the start of each block there were two practice questions.

Table 5.12
Screen layouts included in task instructions for experiment 6

Screen 1

Given that:
It is a lemon

Screen 2

Given that:
It is a lemon
Is it necessary that
It will taste sweet

Screen 1

Given that:
He cuts his finger

Screen 2

Given that:
He cuts his finger
Is it possible that
It will bleed

Participant responses, *yes* or *no*, were recorded by the program, together with the time taken to indicate understanding of the inference (screen 1) and the time taken to complete the reasoning process (screen 2). These were saved to disc.

5.8 Results for experiment 6

The AH4 test sheets were scored in accordance with the test instructions, when one mark was given for each correct answer. In line with the previous studies reported in this thesis, and the procedure generally adopted in the literature, the scores from the AH4 test parts I and II were totalled to give an overall general ability score for each

participant. The observed mean for participants was 96.95 (*SD* = 10.78), which was similar to the available norm of 96.36 (*SD* = 15.01) for university students (Heim, 1968), and to experiment 3 (first transitive inference experiment) and experiment 5 (everyday conditionals), but substantially higher than for the syllogistic reasoning experiment and the first transitive inference experiment, and lower than the abstract conditional reasoning experiment.

The sample was divided into high and low cognitive ability groups, on the basis of a median split on the total AH4 test scores; cases below the median of 96.5 were classified as low ability and those above the median were classified as high ability. All participants evaluated the causal inferences under both necessity instructions and possibility instructions. The first dependent variable was the mean percentage endorsement rates for each inference type, i.e. the number of *yes* responses. The second dependent variable was the time course of the evaluation process; that is to say both premise processing and response times together. These were totalled for each inference type and instruction group to produce a mean evaluation time (in milliseconds). The results from the endorsement rate data are reported first; followed by the results from the evaluation time data. All ANOVA tables for experiment 4 are shown in appendix 5K.

5.8.1 Inference endorsement rates

The mean percentage endorsement rates for all four types of inferences (Necessary, PS, Impossible and PW) are shown in table 5.13, broken down by instruction, inference

type and ability. The cells for the low ability group and the high ability group represent the mean percentage endorsement rates for the responses from 30 participants.

Table 5.13
Mean percentage endorsement rates for experiment 6, on all inference types (N = 60, SD in brackets)

	Necessary		PS		Impossible		PW	
Necessary								
Low	86	*(18.49)*	53	*(33.05)*	4	*(17.58)*	4	*(14.05)*
High	86	*(16.20)*	47	*(31.94)*	4	*(7.44)*	3	*(8.00)*
M	86	*(17.24)*	50	*(32.34)*	4	*(7.45)*	4	*(11.34)*
Possibility								
Low	100	*(2.28)*	100	*(0.00)*	12	*(13.10)*	78	*(26.24)*
High	97	*(5.37)*	98	*(7.14)*	15	*(13.57)*	84	*(19.12)*
M	98	*(4.28)*	99	*(5.08)*	13	*(4.29)*	82	*(22.96)*

Necessary and PS inferences

A 2 (instruction) x 2 (inference type) x 2 (ability) mixed ANOVA test revealed a main effect of instruction [$F(1,58) = 103.08$, $p < .001$, $\eta_p^2 = .64$], reflecting higher endorsement rates when inferences were presented under possibility instructions; and a main effect of inference type [$F(1,58) = 125.88$, $p < .001$, $\eta_p^2 = .69$], whereby Necessary inferences were endorsed more frequently than PS inferences. There was no main effect of ability [$F(1,58) = .58$, $p = .45$]. This indicated that participants had an understanding of the difference between necessity and possibility instructions, which is consistent with the previous experiments reported in this thesis, and also that participants were discriminating between inferences where there were no disabling conditions such as *if butter is heated, it will melt*; and inferences where there were few disabling condition such as *if a baby is hungry, he will cry*.

There was a highly significant interaction between instruction and inference type [F(1,58) = 133.36, p < .001, η_p^2 = .70], suggesting that participants were successfully finding disabling conditions in order to reject the conclusion on PS inferences under necessity instructions, as illustrated in figure 5.6
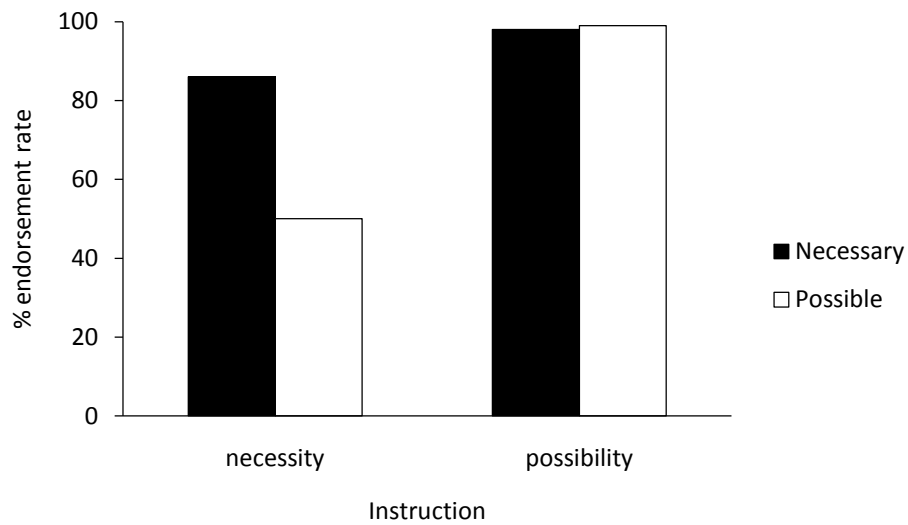
Figure 5.6. Mean percentage endorsement rates for experiment 6, on Necessary and PS inferences under necessity and possibility instructions

For instance given the statement *if the baby is hungry he will cry*, participants were finding at least one disabler for a *baby not crying when it is hungry*. Follow up within subjects *t*-tests were carried out, which confirmed that there was a significant difference between endorsements on inferences under instructions of necessity [t(59) = 11.65, p < .001]. On the other hand, under possibility instructions, where both Necessary and PS inferences are possible, the difference was not significant [t(59) = 1.27, p = .21].

Impossible and PW inferences

A 2 (instruction) x 2 (inference type) x 2 (ability) mixed ANOVA test revealed a main effect of instruction [$F(1,58)$ = 352.83, p < .001, $\eta_p{}^2$ = .85], reflecting higher endorsement rates under possibility instructions; and of inference type [$F(1,58)$ = 342.29, p < .001, $\eta_p{}^2$ = .85], whereby PW inferences were more frequently endorsed than Impossible inferences. There was no main effect of ability [$F(1,58)$ = .92, p = .34]. Again, the main effect of instruction suggests that participants had an understanding of the differences, which is consistent with previous experiments; and the inference type effects suggest that participants were able to discriminate between Impossible statements such as *if it is night time it will be sunny*, and PW statements such as *if he has a cold, he will cough or sneeze.*
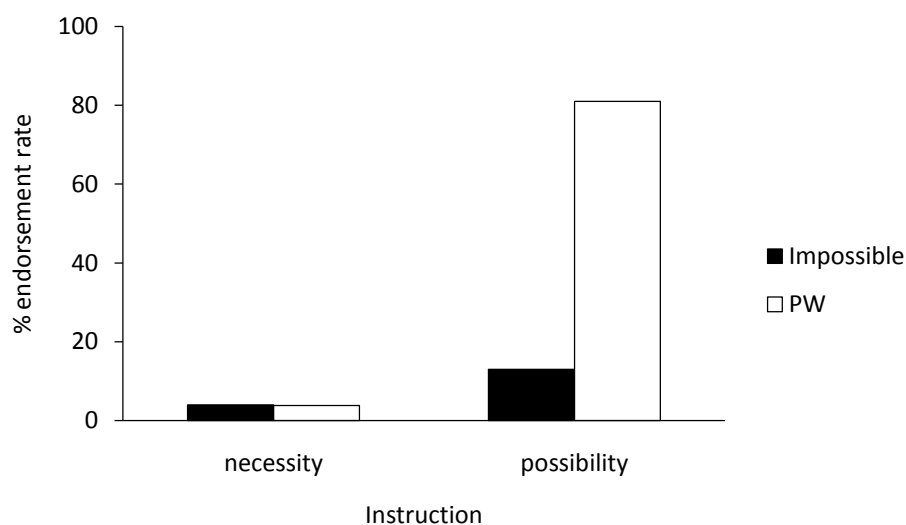
Figure 5.7. Mean percentage endorsement rates for experiment 6, on Impossible and PW inferences under necessity and possibility instructions

There was also a significant interaction between instruction and inference type [$F(1,58)$ = 495.54, p < .001, $\eta_p{}^2$ = .89], which is illustrated in figure 5.7. This interaction

suggests that under possibility instructions, participants looked past their first thought, to find an enabling condition where the statement was true. Follow up between subjects *t*-tests were carried out, which confirmed that there was a significant difference between inference types under possibility instructions [$t(59) = -23.23$, $p < .001$], but not under necessity instructions [$t(59) = .13$, $p = .90$].

5.8.2 Inference evaluation times

Table 5.14
Mean evaluation times in milliseconds for experiment 6, on all inference types
(N = 60, SD in brackets)

	N		PS		I		PW	
Necessary								
Low	4154	*(1846)*	4368	*(1551)*	4672	*(1373)*	4353	*(1188)*
High	3223	*(904)*	3284	*(986)*	3501	*(992)*	3344	*(945)*
M	3688	*(1400)*	3826	*(1400)*	4086	*(1326)*	3849	*(1180)*
Possibility								
Low	3420	*(1070)*	3329	*(1241)*	4492	*(1241)*	4436	*(1344)*
High	2911	*(811)*	2717	*(960)*	3786	*(860)*	3618	*(1283)*
M	3166	*(976)*	3023	*(1205)*	4139	*(1205)*	4027	*(1367)*

The mean inference evaluation times for all inference types are shown in table 5.14, which are again broken down by instruction, inference type and ability. The cells for low ability group and the high ability group represent the mean inference evaluation times for the responses from 30 participants, and are shown in milliseconds.

Necessary and PS inferences

A 2 (instruction) x 2 (inference type) x 2 (ability) mixed ANOVA test revealed a main effect of instruction [$F(1,58) = 25.61$, $p = .001$, $\eta_p^2 = .31$], reflecting longer evaluation

times under necessity instructions; and also of ability [$F(1,58) = 9.28$, $p < .005$, $\eta_p^2 = .14$], when the high ability group were quicker than the low ability group, as predicted. There was no main effect of inference type [$F(1,58) = .00$, $p = .98$], and there were no significant interactions.

Impossible and PW inferences

A 2 (instruction) x 2 (inference type) 2 x (ability) mixed ANOVA test revealed a main effect of ability [$F(1,58) = 20.98$, $p < .001$, $\eta_p^2 = .27$], whereby the high ability group were quicker than the low ability group, again as predicted. There was no main effect of instruction [$F(1,58) = .88$, $p = .35$]; or inference type [$F(1,58) = .23$, $p = .63$], and there were no significant interactions.

5.9 Discussion for experiment 6

The aim of this final experiment was to show how cognitive ability and our knowledge of the world influences the willingness of people to accept or reject a conclusion, when inferences were presented either under necessity instructions, or the more relaxed possibility instructions. There was no logical structure involved, which enabled confirmation of the effect of everyday content in relation to specific antecedents, for each of the problem types and structures which have featured throughout this thesis.

As predicted, the high ability participants were generally quicker for all inference types; and on Necessary and PS inferences the response times were longer for all participants under instructions of necessity, suggesting that reasoners were finding it more difficult to make judgements of necessity than judgements of possibility. There were also more endorsements of possibility than necessity, which was congruent line

with past research (Evans et al., 1999), and consistent with the predictions and findings from experiments 1 – 5.

When making comparisons between scenarios where the event was certain to occur (Necessary), and those where it was highly likely to occur (PS); there was a significant difference suggesting that participants were discriminating between the two, by accessing disabling conditions for the rule under consideration on highly possible scenarios. For instance, participants were prepared to concede that it is not necessary that the *baby will cry if he is hungry*, by perhaps considering that *he is sucking his dummy, or has temporarily lost his voice.* There was firm evidence from the interaction between instruction and inference type that this search for disabling conditions was strongest for PS inferences under instructions of necessity.

The endorsement rates for PW inferences were significantly higher than for Impossible inferences, suggesting that participants were discriminating between Impossible scenarios and ones that are unlikely, but may still occur. There was also evidence to support the prediction that participants were searching for enabling conditions on PW inference under possibility instructions. This suggests that when there were no logical constrictions on the structure of a task, the decision on whether or not to endorse the conclusion under consideration was heavily influenced by our knowledge of everyday events in the world. For example given an inference such as *if it is stormy weather;* participants were searching for enabling conditions for *the oil tanker sinking,* such as *it being heavily laden, or hitting a rock.*

In summary, experiment 6 has shown very clearly that people can discriminate between the meaning of necessity and possibility, when making judgements that rely primarily on their beliefs about the world, and also between instances where the scenario is almost certainly true, highly likely (PS), less likely (PW), and impossible.

5.10 Discussion for experiments 5 and 6

The two final experiments reported in this experimental program of research, fully support the pattern of results reported in the large body of literature on causal conditional reasoning with everyday statements (i.e. Cummings et al., 1991) which confirms that knowledge influences the extent to which people are prepared to accept or reject each of the four conditional inferences, and also inferences that are based solely on content.

This suggests that the nature and number of alternative causes and disabling conditionals is important in the decisions that people make on a daily basis. In experiment 5 more inferences with few alternative causes were endorsed, than inferences with many alternative causes, on Necessary and PS inference structures, and this was reversed for Impossible and PW inference structures. In experiment 6, the endorsement rates were also heavily influenced by content.

In terms of the theoretical implications, there was strong evidence in support of the mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991), for the search for counterexamples under necessity instructions. There was also support for the predictions that individuals would search for alternative models under possibility instructions, when the first model on an inference which was possible did not support

the conclusion. In experiment 5, where there was a distinction between conditional statements with *few* and *many* alternative causes, the effect of a search for counterexamples was present only when there were many alternative causes, but it was across both causal groups under possibility instructions. These are novel findings. There was also very strong evidence to support the search for counterexamples and alternative models from experiment 6, which further confirms the influence of content on the inferences which individuals are prepared to make.

When considering the endorsement rate results for experiment 5 and 6 from a dual process perspective, Evans et al. (2010) found that *high* ability participants were less influenced by content, than *low* ability participants, under instructions of logical necessity, although experiment 5 found no evidence to support this. The presence of this effect under necessity instructions was concluded to be because high ability people are more able to decontextualize the content of an inference, with the logical properties (Evans et al., 2010).

A further theory which was discussed in the introduction is that some researchers (i.e. Liu, Lo, & Wu, 1996; Oaksford & Chater, 1998, 2001) propose that people make inferences on *if then* conditionals with everyday content, according to the perceived probability of *q*, *given p*, for all four inferences. However Markovits and Handley (2005) found considerable evidence to suggest that these two systems are not isomorphic. Given the aims and nature of this programme of research, no direct comparisons are possible, although this may be an area of research that might be explored in the future.

The final chapter of this thesis will provide a discussion of the findings reported in each experimental chapter for the six studies which we carried out. The theoretical implications will also be discussed, together with directions for future research.

Chapter 6

General Discussion

The principal aims and objectives of this programme of experimental research were threefold. First, to extend the investigations on reasoning about necessity and possibility carried out by Evans et al. (1999), to include other paradigms; using a range of deductive arguments and types of inference. Second, to incorporate a measure of cognitive ability; and third to record the time course of the reasoning process, to evaluate whether this is a more sensitive measure in gauging reasoning behaviours. The predictions were derived from the principles of the mental model theory of human reasoning (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991), more specifically the third stage, which proposes that deductive competence is achieved by people searching for counterexamples, to identify other models in order to justify rejection of a putative conclusion.

The approach used in the preparation of this thesis is novel, in that not only do the experiments present reasoning problems with abstract and everyday content, under necessity instructions, across a range of reasoning paradigms; but the problems are

also presented under the more relaxed instructions of possibility. The addition of measures of cognitive ability and the time course of processing judgements, enabled a systematic investigation of whether people search for counterexamples under necessity instructions when the initial model fails to disconfirm the conclusion; and whether they search for alternative models under possibility instructions, when the initial model fails to support the conclusion. This work is theoretically important, because the literature suggests that people tend to be errorful in reasoning, and that they do not consider other possibilities; while mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991) suggests that people do search for other models but are sometimes unsuccessful in this search.

Although the mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991) has generated a large body of reasoning research over the past twenty five years, and has provided a framework for testing reasoning behaviours across a wide range of experimental studies, surprisingly little support has been found for the idea that people actually do search for counterexamples. There is a growing consensus in recent reasoning literature, that if people can find a model that supports the conclusion under consideration, they will satisfice on this model, rather than engaging in a more effortful approach to reasoning. This view is consistent with the satisficing principle, which is the third principle of hypothetical thinking theory, originally proposed by Evans, Over, and Handley (2003), and more fully developed by Evans (2007a).

To facilitate the aims of the reported in this thesis, four different types of reasoning problem were used, which were broadly similar for each paradigm. The structure and

composition of these was inspired by the work of Evans et al. (1999), who carried out a large syllogistic reasoning study which produced a useful database of endorsement rates for all 256 syllogistic combinations, under both necessity and possibility instructions. This enabled the selection of materials for experiment 1, in terms of valid syllogisms, where the conclusion for evaluation is necessarily true and therefore possibly true (Necessary problems), syllogisms with Impossible conclusions, and also two sets of indeterminate syllogisms. Evans et al. (1999) found that problems supporting possible conclusions fell into two categories, which were termed PS (possible strong) and PW (possible weak); when PS problems were regularly taken to imply necessary conclusions, because the first model supported the conclusion, but PW problems were rarely taken to imply necessary conclusions, because the conclusion was not supported by the first model. These two types of possible problems have been pivotal in the development of materials, and the general analysis of the experiments carried out.

We first review and discuss the results from each of the six experimental studies, followed by an evaluation of the reasoning time results and the findings in relation to individual differences in cognitive ability. Following this the theoretical implications will be considered, before moving on to look at areas for future research and our concluding comments.

Summary of key experimental findings from the endorsement rate data

We presented deductive reasoning problems and inferences to participants across three paradigms; syllogistic reasoning, transitive inference and conditional inference.

The syllogistic reasoning problems had abstract content; the transitive inference problems described non-ambiguous relationships between red, blue and green lines akin to 3-term series problems; and the conditional inferences had either abstract content or everyday content. We recorded reasoning behaviours under standard logical instructions, and under instructions asking participants to evaluate whether conclusions possibly followed. We also administered a measure of cognitive ability, and recorded the time course of the reasoning process.

The data was interpreted within the framework of the mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991), and comparisons were made between Necessary and PS problems/inferences, and Impossible and PW problems/inferences. This facilitated the exploration of whether a search for counterexamples and alternative models was carried out, because on PS problem under necessity instructions, a search for counterexamples is needed to find a model that falsifies the initial conclusion, and on PW problems under possibility instructions, a search for alternative models is necessary because the first model disconfirms the conclusion.

The search for counterexamples under necessity instructions

We made a number of predictions, which were systematically tested across the six experiments reported in this thesis. First, conclusions would be more frequently endorsed under possibility instructions than under necessity instructions, because only one model will suffice for a possible conclusion. Also, there would be more endorsements of Necessary problems than PS problems, because PS conclusions are only true under instructions of possibility.

Second, we predicted that a search for counterexamples would result in an interaction between instruction and problem type (Necessary and PS), since under possibility instructions no search for other models is required for either type of problem, but a search is required on PS problems under necessity instructions. As discussed in chapter 2, although there is no evidence to suggest that people know in advance whether they need to search for counterexamples on valid reasoning problems, we do know that they only need to confirm that a given conclusion is correct. However, with invalid problems where the first model supports the premises (PS problems), mental model theory predicts that people need to search for counterexamples in order to reject a given conclusion.

Third, people with higher cognitive ability would be more able to discriminate between problems which were necessarily true (Necessary), and those which were merely possible (PS). Also, people with higher cognitive ability would be more discriminating between the two types of instruction, and would generally be better performers. The ability predictions were based on previous work (i.e. Stanovich and West, 1998a; 1998b).

Our final prediction relates specifically to experiment 5 where the number of alternative causes was manipulated: on inferences with *many alternative causes*, there would be fewer endorsements of PS inferences than Necessary inferences because other causes are more readily availability; however we posited that the effect may not be present when there were *few alternative causes*. This prediction was informed by previous research (Cummins et al., 1995; 1991; De Neys et al., 2003).

Table 6.1

Table of effect sizes (η_p^2) on significant differences in endorsement rates for main effects and interactions - Necessary and PS problems/inferences

	Instruction*	Problem**	Ability	Causes***	Instruction problem	Instruction x ability	Problem x ability	Instruction x causes	Problem x causes
Syllogisms	.14	.11	-	x	-	-	-	x	x
Transitive inference (1st)	-	.17	-	x	-	-	-	x	x
Transitive inference (2nd)	.12	.29	-	x	.20	.18	.08	x	x
Abstract conditionals	.45	.13	-	x	.15	.08	-	-	x
Everyday conditionals	.36	.23	-	.09	.15	-	-	.08	.24
Everyday inferences	.64	.69	-	x	.70	-	-	x	x

*   less endorsements under necessity instructions than under possibility instructions
**   more rejections of PS problem conclusions than Necessary problem conclusions
*** few alternative causes more frequently endorsed than many alternative causes

A summary of effect sizes for significant main effects and interactions is presented in table 6.1. As predicted, we found across all six experiments that people were more likely to endorse conclusions about what they believed was possible, rather than what they believed to be necessary[20], which is in line with previous research using necessity and possibility instructions (i.e. Evans et al., 1999). The effect was strongest for experiment 6 (simple inferences), where no reasoning was required, and responses were based solely on the content of the inference.

---

[20] There was one notable exception; this was in the first transitive inference experiment, when there appeared to be a lack of clarity on the part of the participants, in terms of how to interpret the terms.

Also as predicted and shown in table 6.1, across all six experiments, there was a significant difference between endorsement rates for Necessary problems and PS problems; when Necessary problems were endorsed more frequently than PS problems, indicating that people were rejecting *some* invalid (PS under necessity instructions) inferences. Again, this was strongest for the simple everyday inferences presented to participants in experiment 6. It should also be noted that the frequency of Necessary and PS acceptance rates reported in each experimental chapter was high, particularly under possibility instructions, which was expected given our experimental manipulation. Turning now to look at the extent to which interactions between instruction and problem type provide evidence for the search for counterexamples, and also other interactions present; each chapter will be reviewed in order of presentation in this thesis, with reference to the literature where appropriate.

We first consider experiment 1, which was a replication and extension of previous work carried out by Evans et al (1999), using the syllogistic reasoning paradigm. There was no evidence to support the search for counterexamples (see table 6.1), which confirms comments from Evans at al. (1999) suggesting that little search for alternative models occurs when the first model identified confirms the conclusion. This may well be because individuals were satisficing as proposed by hypothetical thinking theory (Evans, 2007a); because of the structural complexity of syllogisms. The theoretical implications of this will be discussed later in the chapter. In addition, there was no evidence that ability affected reasoning behaviours (see table 6.1). The lack of ability effects is somewhat at odds with some of the earlier reasoning literature (i.e. Stanovich and West, 1998a; 1998b), when it was suggested that higher ability people were better

at syllogistic reasoning; although more recently research in syllogistic reasoning within the belief bias paradigm (i.e. Handley et al., 2010; Evans, et al., 2010) has found that higher ability people are merely less belief biased than lower ability people, when reasoning under instructions of logical necessity.

Chapter 3 introduced the transitive inference paradigm to the problem structures used in the first experiment, and benefitted from the use of a database of endorsement rates collected by Knauff et al. (1995) in a study investigating differences in model variation. The database provided preferred conclusions on indeterminate spatial relationships of transitivity, as well as valid relationships of transitivity, and was used to inform the choice of our transitive inference problems.

The first of our two experiments using transitive inference problems failed to provide evidence to support the search for counterexamples, in that there was no interaction between instruction and problem type, and there were no ability effects (see table 6.1). A number of reasons were considered, which may have contributed to this, aside from the fact that participants were merely satisficing on PS problems after meeting the criteria for adequacy rather than search for the optimum solution as proposed by Goodwin and Johnson-Laird (2005).

The decision to re-run the transitive inference experiments was motivated by three things; first, the observed mean for experiment 2 was considerably lower than the available norm, and the mean, and the lack of results may have been due to participants not having the cognitive ability to carry out this search. It is generally accepted that ability is a good predictor of logically correct performance on a number of reasoning

tasks, and this is discussed in more detail in chapter 1. Second, and in line with the argument put forward by Anderson (1978), participants may not have had a clear understanding of the content and format of the spatial compositions used for the problems, which in turn may have led to interpretational problems. Third, it has been found by Prowse et al. (2009), that immediate feedback is more effective in remedying some of the systematic misunderstandings found in reasoning tasks, which in turn leads to improved performance; although it is acknowledge that improved performance does not necessarily mean that people are searching for counterexamples.

These concerns led us to carry out a second transitive inference experiment, with a learning and practice phase which provided immediate feedback; more in line with the one used in the research carried out by Knauff et al. (1995), in terms of providing immediate feedback as to the correctness of responses during the practice phase. In sharp contrast to experiment 2, the findings from experiment 3 provided firm support for our predictions (see table 6.1), when not only was there was evidence of a search for counterexamples in the form of an interaction between instruction and inference type, but there were also ability effects in that the high ability people were more able to discriminate between types of instruction, and between problem structures.

The fourth experiment, which was reported in chapter 4, was the first of three experiments that we carried out using the conditional inference paradigm. Experiment 4 produced firm evidence in support of the search for counterexamples (see table 6.1), confirming previous research with similar aims (Schroyens, Schaeken and Handley, 2003; Schroyens & Schaeken, 2008). In addition we found that people with higher

cognitive ability were more able to discriminate between instructions of necessity, and instructions of possibility. In chapter 4, we also reported a three way interaction, confirming the predictions made, that it was people with higher cognitive ability who were more likely to search for counterexamples.

The first of the two experiments reported in chapter 5, introduced another factor, in that we varied the number of alternative causes to the rule under consideration that could be accessed from people's everyday knowledge of the world. Early research by Cummings et al. (1995; 1991; De Neys et al., 2003) found the inferences that people were prepared to make could be suppressed by manipulating the number of alternative causes retrievable from the everyday experiences that people have. We confirmed these predictions which were based on this previous research, and found there to be significantly more endorsements of inferences with few alternative causes, than inferences with many alternative causes (see table 6.1). In addition we found that higher ability people were more able to discriminate between types of instruction, and Necessary and PS inference types (see table 6.1). We also found firm evidence to support the prediction that a search would be carried out for counterexamples across few and many alternative causes (see table 6.1), and as we reported in chapter 5, this search was more likely to be successful when there were many alternative causes, than when there were few alternative causes.

Our final experiment used simple inferences, to look at the consequents from specific antecedents, where the emphasis was on content, rather than the interaction between content and logical structure (Cummins et al., 1995; 1991). The content of the

inferences used in this study, were selected by carrying out a pilot study to ensure that the terms reflected the general perceptions that people hold about the world. An example of a Necessary inference is *if butter is heated it will melt*, and it is generally agreed that there is no disabler that will enable butter to stay firm when melted. Alternatively, although it is highly likely that people will die *if the aeroplane crashes,* there are a small number of conditions under which this may not happen.

Although we found no effects of ability (main effects or interactions), there was very strong evidence to suggest that people carried out a search for counterexamples. This confirmed our predictions; and the interaction between inference and instruction suggests that people can discriminate between specific events that are almost certain to occur and those which are highly possible, which in turn interacted with instructions of necessity and possibility.

The search for alternative models under possibility instructions

In this section we will review the evidence for the search for alternative models under possibility instructions. Our predictions were methodically tested across the six experiments as discussed in the previous section. Our first prediction is that people would be more likely to endorse conclusions under possibility instructions than under necessity instructions, because although on PW problems it cannot be concluded that the conclusion under consideration is necessary, it is possible; under both types of instructions Impossible conclusions remain impossible. Also, we predicted that there would be more endorsements of PW problems than Impossible problems, because PW

problems are possible under possibility instructions, but Impossible problems remain impossible under both types of instruction.

The second prediction is that a search for alternative models would result in an interaction between instruction and problem type (Impossible and PW), because under necessity instructions no search for alternative models is required for either type of problem, but a search is required on PW problems under possibility instructions. Again, as discussed in chapter 2, we know that although people may not know that they do not need to search for alternative models on Impossible problems; we do know that if they search for and find alternative models on PW problems, the correct response will be facilitated.

The third prediction is based on previous work (i.e. Stanovich and West, 1998a; 1998b) that people with higher cognitive ability would be more able to discriminate between the problems which were impossible (Impossible), and those which were possible (PW), but not supported by the first model. Furthermore, people with higher cognitive ability would be more able to discriminate between the two types of instruction, and would generally be better performers.

The last prediction we made, which is specific to experiment 5, is that on inferences with *many alternative causes* there would be more endorsements of PW inferences than Impossible inferences, because other causes are more readily available from people's knowledge of the world (Cummins et al., 1995; 1991; De Neys et al., 2003).

The endorsement rate results, in terms of effect sizes for significant main effects and

interactions, are summarised in table 6.2. Consistent with previous research (i.e. Evans et al., 1999), and the findings from the analysis of Necessary and PS problems, we found across all six experiments that people were more likely to endorse conclusions about what they believed was possible, rather than what they believed was necessary. We found this to be strongest for experiment 6, where the nature of the task required people to make inferences based solely on the content of the inference (see table 6.2). We also found that, as expected, the frequency of acceptance rates under both types of instruction were low, which confirms the success of the experimental manipulation.

Furthermore, as predicted, there was a significant difference between endorsement rates for Impossible problems and PW problems on all experiments apart from experiment 5 (every day causal inferences), which is shown in table 6.2. The main effects of both instruction and problem type are strongest in the final experiment (simple inferences), where people merely made inferences based on their everyday knowledge (see table 6.2). We suggest the lack of main effect for experiment 5, may be because of an unexpectedly high number of *yes* responses to Impossible inferences with many alternative causes. We will now look at the extent to which the findings provide evidence to support the search for alternative models, resulting in an interaction between instruction and problem type; and also other interactions that were present. References will be made to the literature where appropriate.

Table 6.2
Table of effect sizes (η_p^2) on significant differences in endorsement rates for main effects and interactions - Impossible and PW problems/inferences

	Instruction*	Problem**	Ability	Causes***	Instruction x problem	Instruction x ability	Problem x ability	Instruction x causes
Syllogisms	.17	.46	-	x		.13	-	x
Transitive inference (1st)	.28	.37	-	x	.36	-	-	x
Transitive inference (2nd)	.51	.36	-	x	.48	.12	-	x
Abstract conditionals	.43	.20	-	x	.24	.11	-	x
Everyday conditionals	.31	-	-	.43	.15	.16	-	.38
Everyday inferences	.85	.85	-	x	.89	-	-	x

* less endorsement under necessity instructions than under possibility instructions
** more rejections of Impossible problem conclusions than PW problem conclusions
*** few alternative causes endorsed less frequently than many alternative causes

First we will consider the results from experiment 1, which replicated and extended previous work by Evans et al. (1999). There was no evidence to support the search for alternative models (see table 6.2). We did, however, find an effect of ability, in that the higher ability people were more able to discriminate between the two types of instruction. In the last section, we considered the possible explanations for the lack of support in the search for alternative models, in that the structural complexity of syllogisms may lead people to satisfice by accepting the first available model, rather than searching for other models to test their initial model.

In contrast to the results which were reported on the search for counterexamples, we found firm evidence from both of the transitive inference experiments to support the search for alternative models (see table 6.2). Our experiment was novel within the transitive inference paradigm, because to our knowledge this is the first time that transitive inference problems have been presented under instructions of necessity *and* possibility.

We suggested that the lack of support for the search for counterexamples under *necessity* instructions in experiment two, may have been because of overall low ability rates; participants having a poor understanding of the content and format of the spatial composition of the problems (Anderson, 1978); or the delay between providing a response and receiving feedback (Prowse et al., 2009). However, given that there was evidence to support the search for other models under possibility instructions in both experiments, and that reasoning with possibility instructions is easier; the low ability of the sample in experiment 2 is a plausible explanation for there being evidence to support the search for both counterexamples and alternative models in experiment 3. We are not saying however that the training given in experiment 3 was ineffective, and similarly that the provision of immediate feedback was not beneficial, but that all or some of these explanations may explain the support for the mental model theory in experiment 3, which was not present in experiment 2.

Our experiment using abstract conditional inferences, reported in chapter 4, produced firm evidence in support of the search for alternative models when the initial model does not support the conclusion. These findings were novel, as other research has not

used possibility instructions of this type, within this paradigm. Having said this, other research has used instructions of non-logical possibility within the belief bias paradigm (Evans et al., 2010), which included pragmatic instructions, but responses were recorded on a scale, rather than the yes/no binary responses used in the studies reported in this thesis. We also found that people with higher cognitive ability were more able to discriminate between the two types of instruction, which was in line with one of the general predictions made.

There was clear evidence to support our predictions relating to the effect of the number of alternatives causes for the scenario, when inferences with few alternative causes were endorsed less frequently than inferences with many alternative causes in experiment 5 (see table 6.2). This suggests that the number of alternative causes increases the likelihood of the identification of other causal models, and is a novel findings, because no previous research has used problems with this type of structure. In addition, we found that people were more prepared to endorse inferences with many alternative causes than those with few alternative causes, when the inferences were presented under instructions of possibility, which again is a novel finding.

There was also firm evidence of a search for alternative models across both causal groups, which is in contrast to when PS inferences were presented under necessity instructions, when the effect was only present when there were many alternative causes (see table 6.2). The only ability effect was that the higher ability people were more discriminating between the two types of instruction.

We found strong evidence from experiment 6 for the search for alternative models, as

shown in table 6.2, this is another novel finding, where our aim was to look at the consequents from specific antecedents, to confirm the effect of everyday content on the inferences that people are prepared to make. In line with the analysis on Necessary and PS inferences, there was no effect of ability, but the findings do confirm that people are able to discriminate between events that are rarely going to occur or impossible events under different types of instruction. For example it is unlikely that *students will be disappointed if a lecture if cancelled*, although it is slightly possible; on the other hand people are rarely inclined to endorse the inference *if it is night time, it will be sunny*.

Individual differences in cognitive ability

In the first chapter it was pointed out that some researchers have highlighted the need to develop a clearer understanding of individual differences in reasoning behaviours. The introduction of a measure of cognitive ability was used for all of the experiments reported in this thesis, but while there is some evidence to suggest that the search for alternative models was mediated by cognitive ability, there was no clear pattern, and our findings were inconsistent across the six experiments.

The predictions we made across all six of the experiments, were that people with higher cognitive ability would be more likely to carry out a search for other models, both under necessity instructions and under possibility instructions. These predictions were upheld in experiment 4 (abstract conditionals) under necessity instructions; and on the second transitive inference experiment under possibility instructions, where it was concluded that people with higher ability benefitted more from an improved learning and practice phase. However, the only other ability effects were that higher

ability people were more able to discriminate between instructions of necessity and possibility and between types of problems/inferences, although this was in terms of modifying endorsement rates not in terms of accuracy,

It may well be that we should think about the lack of individual differences results, in terms of considering whether there is a simple methodological explanation. The procedure which we used to categorise participants into low and high ability groups at the median value of their AH4 scores, is consistent with that used in a number of other studies (i.e. Evans et al., 2010).

However, it is observed that other studies have used larger sample sizes than the 60 participants used for the experiments reported in this thesis, which allowed the sample to be split in such a way as to maximise ability differences. For instance Newstead et al. (2004) split AH5 scores into top quartile ($n = 21$), middle two quartiles ($n = 54$) and bottom quartile ($n = 23$). Although with hindsight this procedure have may be more appropriate, it was not fitting for the experiments that we have reported in this thesis. This is because such a division into quartiles, would have resulted in quartiles consisting of too few participants to make the employed analyses valid, and would have reduced the power of the tests to a degree whereby the chance of a type II error was inflated; ultimately making it difficult to detect any differences which may exist in the population/a larger sample. In selecting the type of analysis, we decided to use a similar method to that of the study carried out by Evans et al. (1999), given that our first experiment was a replication and extension of that work. This method of analysis was systematically employed throughout the thesis. A possible next step for this work

may be to try and replicate the findings with a larger sample, which would allow the inclusion of ability/intelligence as a factor.

Although the results in terms of not finding a consistent relationship between cognitive ability and performance across the reasoning paradigms are disappointing; we did find evidence that higher ability people were slightly better performers on indeterminate problems/inferences. For instance on the syllogistic reasoning experiment, endorsement rates for PS problems under necessity instructions were 68% for the higher ability people, as opposed to 74% for the lower ability people (the logically correct response was to reject the conclusion). Similarly, under instructions of possibility, they were 35% for the lower ability people, and 46% for the higher ability people (the correct response was yes). This is consistent with early research studies which have administered a measure of cognitive ability prior to a reasoning task presented under necessity instructions (Evans, Handley, Neilens, & Over, 2007; Newstead et al., 2004; Evans et al., 1983; Newstead et al., 2004; Newstead et al., 1992; Stanovich & West, 1998b; Torrens et al., 1999). Evans et al., Evans et al., 2007; Newstead et al., 2004; Newstead et al., 1992; Stanovich & West, 1998a, 1998b; Torrens et al., 1999)

Reasoning times

The predictions which we made relating to the time course of processing judgements of both necessity and possibility were derived from the principles that underlie the mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991). Our intention in recording the latencies was to enable the exploration of the deliberation

process or reasoning time, to see whether people were spending extra time when the task required a search for counterexamples or alternative models, for the provision of the correct normative response. We also considered how the number of alternative causes affected the latencies which were collected

However, the evidence was limited and inconsistent in relation to the predictions made, which were that latencies would increase when a search for alternative models was required, in order to evaluate the conclusion under consideration. When considering the search for counterexamples, the only experiment which provided us with evidence to support the predictions, was the first transitive inference experiment, when people took longer on PS inferences under necessity instructions. We found no evidence of increased latencies under instructions of possibility, but there were some other latency effects in that the high ability people were generally quicker on both of the transitive inference experiments.

A further prediction was made on experiment 5, which used everyday conditional inferences with few and many alternative causes to explore the effect of the number of alternative causes that could be recalled from people's everyday experience of the world. We found a main effect of alternative causes on PW inferences under possibility instructions, whereby longer latencies were recorded when there were few alternative causes than when there were many alternative causes. This is an important finding which confirms the predictions made, in that even when there were few other causes, people were still trying to find an alternative cause in order to disconfirm the conclusion under consideration.

Although measuring response times is quite common in psychology, and indeed many studies have successfully used this methodology to their advantage (i.e. Evans & Curtis-Holmes, 2005; Handley et. al., 2010; Kosinski & Cummings, 1999; Luce, 1986; Thompson et al., 2003) there are a number of disadvantages. For instance, Rubinstein (2007) identified the differences in the speed in which participants read and think, reporting very noisy data blurred by the behaviour of participants who choose without serious deliberation. One way around this is by increasing the sample size, to give a clearer picture of the relative time responses. It may also be that although the materials were systematically varied across each of the experiments carried out, this may have been confounded by presenting the stimuli on two consecutive screens, rather than showing the complete problem/inference for evaluation on one screen. It is possible, therefore, that there is a methodological explanation similar to the one considered for the lack of latency effects.

One other factor which may have affected the latency results relates to Evans' (2009) suggestion that people may approach a reasoning task in qualitatively different ways, depending on the materials used, for example: with the conditional inference task, it is assumed that people start with the major premise (if p then q), from whence they move towards a conclusion. However, in syllogistic reasoning perhaps people start with the easier premise; in other words *All of the A's are B's* is an easier relationship to comprehend than *Some of the B's are not C's*, resulting in a different processing order. This would lead to some syllogisms being approached in different ways, depending on the simplicity of each premise. A further comment which has been made (Evans et al., 1999) is that psychology lacks a good theory of how response latencies map onto

cognitive processes, particularly with complex problems such as syllogisms; and therefore perhaps a satisfactory theoretical explanation to this question has not yet been found. With the aforementioned discussion and comments in mind, the remainder of this chapter will focus on the findings from the endorsement rate data.

Theoretical Implications

The opening chapter of this thesis discussed deductive reasoning in the light of general reasoning theories; the mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991), mental logic theories (Braine & O'Brien, 1998; Rips, 1994), the VRH (Polk & Newell, 1995), dual process theories (Evans, 2008; Evans & Over, 1996; Kahneman & Frederick, 2002; Sloman, 1996; Stanovich, 1999), and probabilistic reasoning (Chater & Oaksford, 1999, 2001). Also theories that have been developed specifically to explain reasoning paradigms were discussed, for instance in the case of syllogistic reasoning, one of the theories considered was the atmosphere effect (Woodworth & Sells, 1935). These theories tend to assume cognitive universality, which is the assumption that all individuals reason in a similar way.

The theory that has been used as a framework in the preparation of this thesis is the mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991), which has been the dominant theory in reasoning studies over the past two decades; although dual process theories (i.e. Evans, 2008; Evans & Over, 1996; Kahneman & Frederick, 2002; Sloman, 1996; Stanovich, 1999) now seem to be more widely used to explain research findings across a range of paradigms.

The mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991) suggests that people first construct an internal model from the state of affairs described, second they try to form a parsimonious conclusion by fleshing out the initial model, and finally the stage that is of interest to our research, people search for alternative models (counterexamples) of the premises in which their putative conclusion is false. Although the theory does not extend to judgements of possibility, it follows that if a statement is possible, but not necessarily true; when asked if a set of given premises is possible, the correct response is 'yes' because possibility calls for only a single model of the premises to support the conclusion, whereas necessity calls for all models of the premises to support the conclusion.

In seeking to reconcile the findings with the claims made by the third stage of the mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991), the predictions we made are upheld by five of the six experiments which we ran. Therefore, we found evidence to support our predictions from the transitive inference experiments, and strong evidence from the conditional inference experiments when inferences were presented with abstract contents, and when they were presented with everyday content.

However, the support for the mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991) did not extend to syllogistic reasoning, where we failed to find evidence that people carried out a search for counterexamples, or evidence to suggest that people searched for alternative models under instructions of possibility. However, when preparing materials for this experiment, the main reasoning behaviours that have

been observed in relation to syllogistic reasoning (discussed in chapter 1) were considered, and taken into account, so it is not thought they had a bearing on the results. Specifically, we used a range of problem structures and difficulty, and each category (N, PS, I and PW) had two problems from each of the four syllogistic figures (one a - c and one c - a direction), with one problem in each mood in a - c and c - a direction, thus controlling for figural and atmosphere effects, and conclusion direction.

It is of course conceivable that some people did not understand the quantifiers *all* and *some .... not* and thought that each implied their converse, for example they held the belief that *All of the A's are B's* means the same as *All of the B's are A's.* This is the main characteristic of conversion theory (Chapman & Chapman, 1959), and there is further discussion of this in the following section when directions for further research are considered.

The VRH (Polk & Newell, 1995) offers a possible explanation for the syllogistic reasoning data. The VRH (Polk and Newell, 1995) is a model theory where people evaluate a conclusion by repeatedly encoding the premises until a legal conclusion interrelating the premise terms is generated. According to the VRH (Polk & Newell, 1995), the default mechanism is that reasoners do not proceed past the first model, and emphasis on model formation is less important than in the model theory developed by Johnson-Laird. Therefore within the parameters of the VRH (Polk & Newell, 1995), because syllogisms are linguistically more complex reasoning problems, there may be an increased tendency to produce errors. However the VRH (Polk & Newell) fails to

provide an adequate explanation for poor performance on the other less complex reasoning task tasks that we have used in our research.

The rule based theories of Rips (1994) and Braine and O'Brien (1998) assume that reasoning is carried out by applying rules of inference stored in a mental logic, and problem difficulty is accounted for in terms of the number of rules that must be applied and the faulty application of these rules. Again, although rule based theories may offer an explanation for the syllogistic reasoning data, the nature of the tasks employed in transitive inference and conditional inference research are widely considered to be simpler, so we might expect the rules to be easier to apply, leading to a greater number of correct responses than we have reported throughout this thesis.

An alternative explanation for the data, and which allows for the failure to search for counterexamples in the syllogistic reasoning data that we collected, is hypothetical thinking theory (Evans 2007a). The theory has been referred to on a number of occasions throughout this thesis, and is currently gaining popularity in the reasoning literature (for a comprehensive review see Evans, 2007a). Hypothetical thinking theory consists of three principles; the *singularity principle*, *the relevance principle* and the *satisficing principle.* It is the third of these principles, *the satisficing principle,* which is a key component when reasoning with premises which require either a search for counterexamples in order to reject an initial conclusion, or a search to disconfirm an initial conclusion. The *satisficing principle* suggests that reasoners are prepared to settle for what is 'good enough; in other words, reasoners accept the first model under consideration, unless there is good reason to reject, modify, or replace it. Moving past

this initial model requires effortful active reasoning, which is motivated by external factors or specific instructions which encourage extra reasoning (Evans et al., 2010).

When considering whether hypothetical thinking theory (Evans, 2007a) which was developed from research on syllogistic reasoning, is a good overall explanation for the data reported in this in this thesis, there were a number of reasons why we rejected this account. While it is a plausible explanation for syllogistic reasoning with abstract content, the data from the transitive inference and conditional inference paradigms clearly suggest that satisficing is perhaps not as widespread as is claimed by hypothetical thinking theory (Evans, 2007a), and that people can and do go past the first possible model to find another possible model which disconfirms the initial preferred response. We therefore conclude that hypothetical thinking theory, while offering an explanation for the syllogistic reasoning data, fails to support the data which we collected from the other reasoning paradigms.

In thinking about the wider implications of the data, we make particular reference to experiment 5 where inference problems were presented to participants with both *few* and *many* alternative causes. There was strong support for the influence of the number of alternative causes on the search of counterexamples, which is consistent with the literature (Cummins, 1995; Cummins et al., 1991), and also the influence of the number of alternative causes on the search for alternative models under instructions of possibility. However there may be an alternative explanation for these findings in terms of the probability heuristics model (Oaksford & Chater, 2001), which proposes

that people do not employ logic at all, but rather make judgements with reference to a single probabilistic dimension.

The probability heuristics model (Oaksford & Chater, 2001) suggests that errors and biases arise because people draw incorrect probabilistic inferences based on their knowledge of the world. This rather complex model which is domain specific and composed of computational and algorithmic levels, has been found to account for variation in performance in a number of studies reported in the literature, and has led to an ongoing debate as to whether deductive reasoning and probabilistic reasoning are the same, or whether they are two distinct processes. However Markovits and Handley (2005) carried out a study which made direct comparisons between probabilistic and deductive reasoning, and found convincing evidence to suggest that the two are not interchangeable. On the other hand an earlier study by Liu et al. (1996) reported that when people were asked to rate perceived probabilities on a likert scale, most people treated a conditional as a probabilistic statement, concluding a probabilistic model is an appropriate one to adopt.

According to a probabilistic account of reasoning, judgements of possibility and necessity such as those used in the studies reported in this thesis, should differ only in terms of where participants place the threshold for response. Therefore under possibility instructions participants should generally be less sensitive in the placing of this threshold. However, the findings of this thesis show an interaction between instruction and problem/inference type, suggesting that participants approach judgements of possibility and necessity in qualitatively different ways. This is

consistent with recent work which suggests that deductive and inductive judgements are accomplished through distinct processes (Heit and Rotello, 2010, Rips, 2001).

Directions for future research

Although the evidence to support the search for counterexamples and alternative models was a fairly robust effect across the transitive inference and conditional inference paradigms (replicated five times across these experiments), there was no evidence of this in our first experiment, where people were presented with abstract syllogistic reasoning problems. This is somewhat surprising, since the search for counterexamples is a key component of mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991), which was originally developed using the syllogistic reasoning paradigm, and later projected as a general theory of reasoning. The suggestions for future research are based on the effect of training on the terminology used in the transitive inference experiments reported in this thesis, and how this might be transferred to further explore the relationship between syllogistic reasoning with abstract content, training, and cognitive ability.

The second and third experiments (transitive inference) presented in this thesis raised questions on which is whether people failed to search for counterexamples in experiment 2 because the sample had an ability rate that was 6 points below the norm and below the ability rate recorded in experiment 1. Alternatively it may be that the participants did they did not fully understand how to interpret the terms used to construct the transitive inference problems; or whether they may have benefitted from immediate feedback during the learning and practice phase. These concerns led to the

design of a second transitive inference experiment which is also reported in chapter 3. The second transitive inference experiment subsequently produced strong evidence to support both the search for counterexamples under instructions of necessity, and alternative models under possibility instructions.

Evans (1972) refers to interpretational problems within reasoning studies, because of a misunderstanding relating to the meaning of the terms or relationships used. Although it was only one of the reasons considered for the lack of evidence to support the search for counterexamples found in experiment 2, it may well be that by introducing a training phase to syllogistic reasoning experiments, based on the one used in the second transitive inference experiment, evidence will emerge to support the search for counterexamples. This will then give us a clearer picture of the reasoning behaviours produced by syllogistic reasoning with abstract content. Indeed, an early theory specific to syllogistic reasoning, conversion theory (Chapman & Chapman, 1959) suggests that some people do not understand the quantifiers *all* and *some …. not* and believe that each implies its converse, for example *All A's are B's* means the same as *All B's are A's.* Also, early research (Henle, 1962) argues that people do not commit logical errors, but they merely misinterpret the premises presented to them.

To our knowledge, no other work has focussed on designing a training phase to explore whether this facilitates people to search for counterexamples or alternative models in syllogistic reasoning. Studies looking at the impact of training and instruction on both logical and statistical reasoning tasks (Fong, Krantz, & Nisbett, 1986; Lehman, Lempert, & Nisbett, 1988; Nisbett & Ross, 1980); report significant improvements in the

reasoning skills of individuals after training and practice, and although the aim is not to improve reasoning skills per se, it is anticipated that this may also be the case.

One key point to bear in mind when considering how best to introduce a training and practice phase into syllogistic reasoning studies, is that research has shown individuals tackle syllogistic reasoning problems in qualitatively different ways. Ford (1995) made a basic distinction between verbal and spatial reasoners, which has been confirmed by a number of studies (i.e. Bacon 2003; Serpell, 2004), and this research has shown that *some* people adopt predominantly spatial strategies in conclusion evaluation tasks, and *some* use verbal strategies. Notwithstanding this, there have been found to be few performance differences between strategy groups, despite each strategy group finding the other group's strategy quite alien. A syllogistic reasoning strategy questionnaire was later developed and used by Bacon (2003) and Serpell (2004), which was found to reliably identify strategic preferences independently of reasoning tasks. This might usefully be employed when developing a training phase, for presentation to participants prior to a series of syllogistic reasoning problems, thus avoiding cueing people into adopting an alien strategy. This is a novel methodology and approach to individual difference in syllogistic reasoning, and to our knowledge has not been reported previously in the literature.

The role of individual differences in cognitive ability, as a mediating factor in the effectiveness of training in deductive reasoning was extensively investigated by Neilens (2004); who reported that participants of higher ability were more able to understand and apply the principles they had been taught in order to transfer these skills and

knowledge to a number of reasoning and problem solving tasks. The present experiments suggest that this is also the case within the transitive inference paradigm, since with the benefit of an improved training phase, there was evidence that the higher ability group were more likely to search for alternative models. As mentioned previously the procedure which we used to categorise participants into low and high ability groups at the median value of their AH4 scores, is consistent with that used in a number of other studies (i.e. Evans et al., 2010). However, our sample size was relatively small, and a larger sample would allow the sample to be split into quartiles, and to include ability/intelligence as a factor.

Overall, the above suggestions would allow the research presented in this thesis to be extended to investigate whether training in quantifier interpretation leads to reasoning behaviours which are consistent with the third stage of the mental model theory (Johnson-Laird, 1983; Johnson Laird & Byrne, 1991), and whether these behaviours are linked to individual differences in cognitive ability. This would also afford a better understanding of syllogistic reasoning, and add something novel to the extensive literature that currently exists.

Concluding comments

This programme of study has provided a number of novel findings to advance our understanding of the extent to which reasoners think about possibilities, when reasoning deductively under instructions of logical necessity, and under the more relaxed instructions of possibility. First there was firm evidence to suggest that individuals can, and do search for other possibilities to the first available model, when

making transitive inferences, and reasoning with conditional inferences containing both abstract and everyday content. There was also strong evidence confirming past research (Cummins, 1995), that the likelihood of a successful search being carried out is mediated by the number of alternative causes; as the search more frequently took place when an individual could access many alternative causes from their everyday experience of the world.

Second, for the first time across reasoning paradigms, two other measures were introduced (cognitive ability and the time course of the reasoning process) to evaluate whether they were more sensitive in gauging whether people can and do search for other models, before providing a correct evaluation to a given conclusion. While there was evidence that high ability people were more likely to search for counterexamples on abstract conditionals, and for alternative models when making transitive inferences; there was limited and inconsistent evidence for the predictions that if people searched for models they would take longer.

The thesis also highlights the fact that syllogisms are, as many researchers believe, unique in terms of the reasoning behaviours that they produce. This view is supported by the experimental studies which we have reported in this thesis, where we found evidence to support the search for counterexamples from all of our studies, except the first experiment which used abstract syllogistic reasoning problems.

In conclusion, the experiments that we have carried out and reported in this thesis offer support for the mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991), and provided some fresh insight into how individuals consider

possibilities, based upon the knowledge and information available at that time. We believe it also leads to other avenues of research in the field of human reasoning, in order to facilitate and further our knowledge of what is an inherently human characteristic.

Reference list

Alexopoulos, D. S. (1997). Reliability and validity of Heim's AH4 in Greece. *Personality and Individual Differences, Volume 22*(3), 429-432.

Allen, J. F. (1983). Maintaining knowledge about temporal intervals. *Communications of the ACM, 26*, 832-843.

Anderson, J. R. (1978). Arguments Concerning Representations for Mental Imagery. *Psychological Review, 85*(4).

Anderson, J. R. (1982). Acquisition of Cognitive Skill. *Psychological Review, 89*(4), 369-406.

Andrews, G., & Halford, G. S. (1994). Relational complexity and sentence processing. *Australian Journal of Psychology, 46*(1).

Andrews, G., & Halford, G. S. (2002). A cognitive complexity metric applied to cognitive development. *Cognitive Psychology, 45*, 153-219.

Bacon, A. M. (2003). Individual Differences and Strategies for Human Reasoning. University of Plymouth, Plymouth.

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

Barrouillet, P., & Lecas, J. F. (1998). How can mental models theory account for content effects in conditional reasoning? *Cognition, 67*(209-253).

Begg, I., & Denny, J. P. (1969). Empirical reconciliation of atmosphere and conversion interpretations of reasoning errors. *Journal of Experimental Psychology, 81*, 351-354.

Begg, I., & Denny, J. P. (1982). On the interpretation of syllogisms. *Journal of Verbal Learning and Verbal Behaviour, 21*, 595-620.

Bell, V., & Johnson-Laird, P. N. (1998). A Model Theory of Modal Reasoning. *Cognitive Science, 22*(1), 25-51.

Blanchette, I. (2006). The effect of emotion on interpretation and logic in a conditional reasoning task. *Memory and Cognition, 34*(5), 1112-1125.

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

Braine, M. D. S., & O'Brien, D. R. (1991). A theory of If: A lexical entry, reasoning program, and pragmatic principles. *Psychological Review, 98*, 182-203.

244

Bucciarelli, M., & Johnson-Laird, P. N. (1999). Strategies in Syllogistic Reasoning. *Cognitive Science, 23*(3), 247-303.

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

Byrne, R. M. J., & Johnson-Laird, P. N. (1990). Remembering conclusions we have inferred: What biases reveal. In J. P. Caverni, J. M. Fabre & M. Gonzalez (Eds.), *Cognitive Biases: Their Contribution for Understanding Human Cognitive Processes* (Vol. 68). Amsterdam: North-Holland.

Capon, A., Handley, S., & Dennis, I. (2003). Working memory and reasoning: An individual differences perspective. *Thinking and Reasoning, 9*(3), 203-244.

Chapman, L. J., & Chapman, J. P. (1959). Atmosphere effect re-examined. *Journal of Experimental Psychology, 58*, 220-226.

Chater, N., & Oaksford, M. (1999). The Probability Heuristics Model of Syllogistic Reasoning. *Cognitive Psychology, 38*, 191-258.

Chater, N., & Oaksford, M. (2001). Human rationality and the psychology of reasoning: Where do we go from here? *British Journal of Psychology, 92*, 193-216.

Clark, H. H. (1969). Influence of language on solving three term series problems. *Journal of Experimental Psychology, 82*, 205-215.

Clark, H. H., & Chase W.G. (1974). Perceptual coding strategies in the formation and verification of descriptions. *Memory and Cognition, 2,* 14, 101-111.

Clark, H. H., & Clark, E. (1976). *Psychology and language: An introduction to psycholinguistics.* New York: Harcourt Brace Jovanovich.

Clement, C. A., & Falmagne, R. J. (1986). Logical reasoning, world knowledge, and mental imagery: Interconnections in cognitive processes. *Memory and Cognition, 14*, 299-307.

Colom, R., Rebollo, I., Palacios, A., Juan-Espinosa, M., & Kyllonen, P. C. (2004). Working memory is (almost) perfectly predicted by g. *Intelligence, 32*, 277-296.

Cummins, D. D. (1992). Role of Analogical Reasoning in the Induction of Problem Categories. *Journal of Experimental Psychology: Learning, Memory and Cognition, 18*(5), 1103-1124.

Cummins, D. D. (1995). Naive theories and causal deduction. *Memory and Cognition, 23*(5), 646-658.

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

De Neys, W., Schaeken, W., & d'Ydewalle, G. (2002). Causal conditional reasoning and semantic memory retrieval: A test of the semantic memory framework. *Memory and Cognition, 30*(6), 908-920.

De Neys, W., Schaeken, W., & d'Ydewalle, G. (2003). Causal conditional reasoning and strength of association: The disabling condition case. *European Journal of Cognitive Psychology, 15*(2), 161-176.

De Soto, C. B., London, M., & Handel, S. (1965). Social reasoning and spatial paralogic. *Journal of Personality and Social Psychology, 2*, 513-521.

Dickstein, L. S. (1978). The effect of figure on syllogistic reasoning. *Memory and Cognition, 6*, 76-83.

Evans, J. St. B. T. (1972). On the problem of interpreting reasoning data: Logical and psychological approaches. *Cognition, 1*, 373-384.

Evans, J. St. B. T. (1977). Linguistic factors in reasoning. *Quarterly Journal of Experimental Psychology, 29,* 297-306.

Evans, J. St. B. T. (1989). *Bias in Human Reasoning: Causes and Consequences*. Hove: Lawrence Erlbaum Associates.

Evans, J. St. B. T. (2003). In two minds: dual-process accounts of reasoning. *Trends in Cognitive Sciences, 7*(10), 454-459.

Evans, J. St. B. T. (2004). Dual processes, evolution and rationality. *Thinking and Reasoning, 10*(4), 405-410.

Evans, J. St. B. T. (2007a). *Hypothetical Thinking*. Hove: Taylor and Francis Group.

Evans, J. St. B. T. (2007b). On the resolution of conflict in dual process theories of reasoning. *Thinking and Reasoning, 13*(4), 321-339.

Evans, J. St. B. T. (2008). Dual-processing Accounts of Reasoning. Judgement and Social Cognition. *Annual Review of Psychology, 59*(1), 255-278.

Evans, J. St. B. T. (2009). Dual-processing accounts of reasoning. judgement and social cognition. *Annual Review of Psychology, 59*(1), 255-278.

Evans, J. St. B. T. (in press). Dual-process theories of reasoning: Facts and fallacies. In K. J. Holyoak and R. G. Morrison (Eds). *The Oxford handbook of thinking and reasoning*. New York: Oxford University Press.

Evans, J. St. B. T., Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory and Cognition, 11*(3), 295-306.

Evans, J. St. B. T., Clibbens, J., & Rood, B. (1995). Bias in Conditional Inference: Implications for Mental Models and Mental Logic. *The Quarterly Journal of Experimental Psychology, 48A*(3), 644-670.

Evans, J. St. B. T., Handley, S., & Neilens, H. L. (2010). The influence of cognitive ability and instructional set on causal conditional inference. *The Quarterly Journal of Experimental Psychology, 63*(5), 892-909.

Evans, J. St. B. T., Handley, S. J., & Bacon, A. M. (2009). Reasoning Under Time Pressure. *Experimental Psychology., 56*(2), 77-83.

Evans, J. St. B. T., Handley, S. J., & Harper, C. N. J. (2001). Necessity, possibility and belief: A study of syllogistic reasoning. *The Quarterly Journal of Experimental Psychology, 54A*(3), 935-958.

Evans, J. St. B. T., Handley, S. J., Harper, C. N. J., & Johnson-Laird, P. N. (1999). Reasoning about necessity and possibility: A test of the mental model theory of deduction. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(6), 1495-1513.

Evans, J. St. B. T., Handley, S. J., Neilens, H., & Over, D. E. (2007). Thinking about conditionals: A study of individual differences. *Memory and Cognition, 35*(7), 1772-1784.

Evans, J. St. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human Reasoning: The Psychology of Deduction*. Hove: Lawrence Erlbaum Associates.

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

Evans, J. St. B. T., & Over, D. E. (2004). *If: supposition, pragmatics, and dual processes*. Oxford: Oxford University Press.

Evans, J. St. B. T., Over, D. E., & Handley S. J. (2003). A theory of hypothetical thinking. In D. Hardman & L. Maachi (Eds.), *Thinking: Psychological perspectives on reasoning, judgement and decision making* (pp. 3-22). Chichester, UK: Wiley.

Feeney, A. (2007). In A. Feeney & E. Heit (Eds.), Inductive reasoning: experimental, developmental and computational approaches. Cambridge: Cambridge University Press.

Fitts, P. M., & Posner, M. I. (1967). *Human Performance*. Belmont, California: Brooks Cole.

Fong, G. T., Krantz, D. H., & Nisbett, R. E. (1986). The effects of statistical training on thinking about everyday problems. *Cognitive Psychology, 18*, 253-292.

Ford, M. (1995). Two modes of mental representation and problem solution in syllogistic reasoning. *Cognition, 54*, 1-71.

Galotti, K. M., Baron, J., & Sabini, J. P. (1986). Individual differences in syllogistic reasoning: Deduction rules or mental models? *Journal of Experimental Psychology: General, 115*(1), 16-25.

Goel, V. (2007). Anatomy of deductive reasoning. *Trends in Cognitive Sciences, 11*(10).

Goodwin, G. P., & Johnson-Laird, P. N. (2005). Reasoning about Relations. *Psychological Review, 112*(2), 468-493.

Goodwin, G. P., & Johnson-Laird, P. N. (2006). Reasoning about the relations between relations. *The Quarterly Journal of Experimental Psychology, 59*(6), 1047-1069.

Gottfredson, L. S. (1997). Why g Matters: The Complexity of Everyday Life. *Intelligence, 241*(79-132).

Grice, P. (1975). Logic and Conversation. In P. Cole & J. L. Morgan (Eds.), *Studies in Syntax: Volume 3: Speech acts*. New York: Academic Press.

Handley, S. J., & Evans, J. St. B. T. (2000). Supposition and representation in human reasoning. *Thinking and Reasoning, 6*(4), 273-311.

Handley, S.J., Newstead, S.E., Trippas, D. (2010). Logic, Beliefs, and Instruction: A Test of the Default Interventionist Account of Belief Bias. *Journal of Experimental Psychology,* 139(4).

Heim, A. W. (1968). *AH4 group test of general intelligence manual*. Windsor, UK: N.F.E.R.

Heit, E., & Rotello, C. M. (2010). Relations Between Inductive Reasoning and Deductive Reasoning. *Journal of Experimental Psychology: Learning, Memory and Cognition, 36*(3).

Henle, M. (1962). On the relation between logic and thinking. *Psychological Review, 69*(4), 366-378.

Huttenlocher, J. (1968). Constructing spatial images: A strategy in reasoning. *Psychological Review, 75*, 550-560.

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*. Cambridge: University Press.

Johnson-Laird, P. N., & Bara, B. G. (1984). Syllogistic inference. *Cognition, 16*, 1-61.

248

Johnson-Laird, P. N., & Byrne, R. (1989). Only Reasoning. *Journal of Memory and Language, 28*.

Johnson-Laird, P. N., & Byrne, R. (1994). Models, necessity and the search for counter-examples: A reply to Martin-Cordero & Gonzalez-Labra and to Smith. *Behavioural and Brian Sciences, 17*, 775-777.

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

Johnson-Laird, P. N., & Steedman, M. (1978). The Psychology of Syllogisms. *Cognitive Psychology, 10*, 64-99.

Kahneman, D. (2003). A perspective on judgement and choice. *American Psychologist, 58*, 697-720.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited. In T. Gilovich, D. Griffin & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment.* Cambridge: Cambridge University Press.

Kern, L.H., Mirrels, H.L., & Hinshaw, V.G. (1983). Scientists' understanding of propositional logic: An experimental investigation. *Social Studies of Science, 13*, 131-146.

Klaczykski, P. A., & Daniel, D. B. (2005). Individual differences in conditional reasoning: A dual -process account. *Thinking and Reasoning, 11*(4), 305 - 325.

Klaczynski, P. A., Fauth, J., & Swanger, A. (1998). Adolescent identity: Rationality, critical thinking, and formal operations. *Journal of Youth and Adolescence, 27*, 185-207

Klauer, K. C., Musch, J., & Naumer, B. (2000). On belief bias in syllogistic reasoning. *Psychological Review, 107*(4), 853-884.

Knauff, M. (1999). The cognitive adequacy of Allen's interval calculus for qualitative spatial representation and reasoning. *Spatial Cognition and Computation, 1*, 261-290.

Knauff, M., & Johnson-Laird, P. N. (2002). Visual imagery can impede reasoning. *Memory and Cognition, 30*(3), 363-371.

Knauff, M., Rauh, R., & Schlieder, C. (1995). *Preferred Mental Models in Qualitative Spatial Reasoning: A Cognitive Assessment of Allen's Calculus.* Paper presented at the Proceedings of the Seventeenth Annual Conference of the Cognitive Society.

Knauff, M., Rauh, R., Schlieder, C., & Strube, G. (1998). *Continuity Effect and Figural Bias in Spatial Relational Inference.* Paper presented at the Twentieth Annual Conference of the Cognitive Society.

Kosinski, R. A., & Cummings, J. (1999). *The Scientific Method: An Introduction Using Reaction Time.* Paper presented at the 20th Workshop/Conference of the Association for Biology Laboratory Education (ABLE).

Kruger, J. & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own competence lead to inflated self-assessments. Journal of Personality and Social Psychology, 77, 1121-1134.

Kunda, Z. (1990). The Case for Motivated Reasoning. *Psychological Bulletin, 108*(3), 480-498.

Lehman, D. R., Lempert, R. O., & Nisbett, R. E. (1988). The effects of graduate training on reasoning: Formal discipline and thinking about everyday-life events. *American Psychologist, 43*, 431-443.

Liu, I., Lo, K., & Wu, J. (1996). A Probabilistic interpretation of 'If-Then'. *The Quarterly Journal of Experimental Psychology, 49A*(3), 828-844.

Luce, R. D. (1986). *Response Times: Their Role in Inferring Elementary Mental Organisation.* Oxford: Oxford University Press.

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

Markovits, H. (2000). A mental model of analysis of young children's conditional reasoning with meaningful premises. *Thinking and Reasoning, 6*(4), 91-123.

Markovits, H., & Barrouillet, P. (2004). Introduction: Why is understanding the development of reasoning important. *Thinking and Reasoning, 10*(2).

Markovits, H., & Handley, S. J. (2005). Is inferential reasoning just probabilistic reasoning in disguise? *Memory and Cognition, 33*(70), 1315-1323.

Morley, N. J., Evans, J. St. B. T., & Handley, S. J. (2004). Belief bias and figural bias in syllogistic reasoning. *The Quarterly Journal of Experimental Psychology, 57A* (4), 666-692.

Neilens, H. L. (2004). *Training and Dual Processes in Human Thinking.* University of Plymouth, Plymouth.

Newstead, S. E. (1989). Interpretational Errors in Syllogistic Reasoning. *Journal of Memory and Language, 28*, 78-91.

Newstead, S. E. (1995). Gricean implicatures and syllogistic reasoning. *Journal of Memory and Language, 34*, 644-664.

Newstead, S. E. (2003). Can natural language semantics explain syllogistic reasoning? *Cognition, 90*, 193-199.

Newstead, S. E., & Griggs, R. A. (1983). Drawing inferences from quantified statements: A study of the square of opposition. *Journal of Verbal Learning and Verbal Behaviour, 22*, 536-546.

Newstead, S. E., & Griggs, R. A. (1999). Premise Misinterpretation and Syllogistic Reasoning. *The Quarterly Journal of Experimental Psychology, 52A*(4), 1057-1075.

Newstead, S. E., Handley, S. J., & Buck, E. (1999). Falsifying mental models: Testing the predictions of theories of syllogistic reasoning. *Memory and Cognition, 27*(2), 344-354.

Newstead, S. E., Handley, S. J., Harley, C., Wright, H., & Farrelly, D. (2004). Individual differences in deductive reasoning. *The Quarterly Journal of Experimental Psychology, 57A*(1), 33-60.

Newstead, S. E., Pollard, P., Evans, J. St. B. T., & Allen, J. L. (1992). The source of belief bias effects on syllogistic reasoning. *Cognition, 45*, 257-284.

Newstead, S. E., Pollard, P., & Riezebos, D. (1987). The effect of set size on the interpretation of quantifiers used in rating scales. *Applied Ergonomics, 18*, 178-182.

Newstead, S. E., Thompson, V., & Handley, S. (2002). Generating alternatives: A key component in human reasoning? *Memory and Cognition, 30*(1), 129-137.

Nisbett, R., & Ross, L. (1980). *Human Inference: Strategies and shortcomings of social judgement.* Englewood Cliffs, NJ: Prentice-Hall.

Oaksford, M., & Chater, N. (2001). The probabilistic approach to human reasoning. *Trends in Cognitive Sciences, 5*(8).

Oaksford, M., & Chater, N. (2009). Précis of Bayesian Rationality: The Probabilistic Approach to Human Reasoning. *Behavioural and Brain Sciences, 32*, 69 - 120.

Osheron, D. N. (1976). *Logical ability in children: Vol. 4. Reasoning and concepts.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Polk, T. A., & Newell, A. (1995). Deduction as Verbal Reasoning. *Psychological Review, 102*(3), 533-566.

Prowse Turner, J. A., & Thompson, V. A. (2009). The role of training, alternative models, and logical necessity in determining confidence in syllogistic reasoning. *Thinking and Reasoning, 15*(1), 69-100.

Quayle, J. D., & Ball, L. J. (2001). *Cognitive Uncertainty in Syllogistic Reasoning: An Alternative Mental Models Theory.*

Rauh, R., Hagen, C., Schlieder, C., Strube, G., & Knauff, M. (2000). *Searching for Alternatives in Spatial Reasoning: Local Transformations and Beyond.* Paper presented at the Annual Conference of the Cognitive Society.

Revlis, R. (1975). Two models of syllogistic inference: Feature selection and conversion. *Journal of Verbal Learning and Verbal Behaviour, 14*, 180-195.

Rips, L. J. (1994). *The Psychology of Proof.* London: The MIT Press.

Rips, L. J. (2001). Two kinds of reasoning. *Psychological Science, 12*(2).

Roberts, M., Newstead, S., & Griggs, R. A. (2001). Quantifier interpretation and syllogistic reasoning. *Thinking and Reasoning, 7*(2), 173-204.

Rubinstein, A. (2007). Instinctive and Cognitive Reasoning: A Study of Response Times. *The Economic Journal, 117*, 1243-1259.

Rumain, B., Connell, J., & Braine, M. D. S. (1983). Conversational Comprehension Processes Are Responsible for Reasoning Fallacies in Children As Well as Adults: If Is Not the Biconditional. *Developmental Psychology, 19*(4), 471-481.

Sa, W., West, R. F., & Stanovich, K. E. (1999). The domain specificity and generality of belief bias. Searching for a generalisable critical thinking skill. *Journal of Educational Psychology, 91*, 497-510.

Schmidt, J., & Thompson, V. A. (2008). "At least one" problem with "some" formal reasoning paradigms. *Memory and Cognition, 36*(1), 217-229.

Schroyens, W., & Schaeken, W. (2008). Deductive Rationality in Validating and Testing Conditional Inferences. *Canadian Journal of Experimental Psychology, 62*, 163-173.

Schroyens, W., Schaeken, W., & d'Ywelle, G. (2001). The processing of negations in conditional reasoning: A meta-analytic case study in mental model and/or mental logic theory. *Thinking and Reasoning, 7*(2), 121-172.

Schroyens, W., Schaeken, W., & Handley, S. (2003). In search of counterexamples: Deductive rationality in human reasoning. *The Quarterly Journal of Experimental Psychology., 56A*, 1129-1145.

Serpell, S. M. P. (2004). *The effect of a cue-in task on strategy selection in syllogistic reasoning.* Unpublished Undergraduate Dissertation, Plymouth.

Shaver, P., Pieron, L., & Lang, S. (1975). Converting evidence for the functional significance of imagery in problem solving. *Cognition, 3*, 359-375.

Simon, H. (1957). *Models of man: Social and rational.* New York: Wiley.

Simon, H. (1983). *Reason in human affairs*. Stanford: Stanford University Press.

Sloman, S. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*(1), 3-22.

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning.* Mahwah, NJ: Lawrence Erlbaum Associates.

Stanovich, K. E. (2008). Higher-order preferences and the Master Rationality Motive. *Thinking and Reasoning, 14*(1), 111-127.

Stanovich, K. E., & West, R. F. (1998a). Cognitive Ability and Variation in Selection Task Performance. *Thinking and Reasoning, 4*(3), 193 - 230.

Stanovich, K. E., & West, R. F. (1998b). Individual Differences in Rational Thought. *Journal of Experimental Psychology, 127*(2), 161-188.

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

Sternberg, R., & Salter, W. (1982). *Handbook of human intelligence.* Cambridge, UK: Cambridge University Press.

Stupple, E. J. N., & Ball, L. J. (2007). Figural Effects in a Syllogistic Evaluation Paradigm: An Inspection-Time Analysis. *Experimental Psychology, 54*(2), 120-127.

Stupple, E. J. N., & Ball, L. J. (2008). Belief-logic conflict resolution in syllogistic reasoning: Inspection-time evidence for a parallel-process model. *Thinking and Reasoning, 14*(2), 168-181.

Thompson, V. A. (1994). Interpretational factors in conditional reasoning. *Memory and Cognition, 22*(6), 742-758.

Thompson, V. A. (2000). Interpretational factors in conditional reasoning. *Memory and Cognition, 22*, 742-758.

Thompson, V. A., Striemer, C. L., Reikoff, R., Gunter, R., & Campbell, J. I. D. (2003). Syllogistic reasoning time: Disconfirmation disconfirmed. *Psychonomic Bulletin and Review, 10*(1), 184-189.

Torrens, D., Thompson, V. A., & Cramer, K. M. (1999). Individual Differences and the Belief Bias Effect: Mental Models, Logical Necessity, and Abstract Reasoning. *Thinking and Reasoning, 5*(1), 1-28.

Vandierendonck, A. (Ed.). (2000). *Model construction and elaboration in spatial linear syllogisms.* New York: Lawrence Erlbaum Associates.

Vandierendonck, A., Doerckx, V., & Vooght, G. D. (2004). Mental model construction in linear reasoning: Evidence for the construction of initial annotated models. *The Quarterly Journal of Experimental Psychology, 57A*(8), 1369-1391.

Verschueren, N., Schaeken, W., & d'Ydewalle, G. (2005). A dual-process specification of causal conditional reasoning. *Thinking and Reasoning, 11*(3), 239-278.

Wason, P. C. (Ed.). (1966). *Reasoning*. Harmondsworth: Penguin Books.

Wetherick, N. E. (1989). Psychology and syllogistic reasoning. *Philosophical Psychology, 2*(1), 111-124.

Wetherick, N. E., & Gilhooly, K. J. (Eds.). (1990). *Syllogistic reasoning: effects of premise order*. Chichester: Wiley.

Wildman, T. M., & Fletcher, H. J. (1977). Developmental increases and decreases in solutions of conditional syllogism problems. *Developmental Psychology, 13*, 630-636.

Woodworth, R., & Sells, S. B. (1935). An Atmosphere effect in formal syllogistic reasoning. *Journal of Experimental Psychology, 18*, 451-460.

Appendices

All appendices are numbered in line with the chapter to which their contents refer

Appendix 2A
A complete set of syllogisms together with problem type, percentage endorsement rates previously recorded, figure, and conclusion direction (Evans et al., 1999)

First premise	Second premise	Conclusion	Problem	N*	P*	Figure	ac-ca
All R's are N's	All N's are B's	All R's are B's	N	73	80	1	
All P's are E's	No D's are E's	No P's are D's	N	83	87	3	
All P's are M's	Some A's are P's	Some M's are A's	N	87	87	2	a-c
All J's are E's	Some J's are not B's	Some E's are not B's	N	87	83	4	
All F's are T's	All N's are F's	All N's are T's	N	80	87	2	
All Q's are G's	No B's are G's	No B's are Q's	N	80	80	3	
Some M's are D's	All D's are P's	Some P's are M's	N	83	90	1	c-a
Some R's are not D's	All R's are K's	Some K's are not D's	N	87	90	4	
All M's are R's	All M's are K's	All R's are K's	PS	70	87	4	
No F's are E's	All E's are C's	No F's are C's	PS	77	83	1	
Some Q's are G's	All C's are G's	Some Q's are C's	PS	83	97	3	a-c
All D's are N's	Some P's are not D's	Some N's are not P's	PS	93	93	2	
All J's are T's	All T's are A's	All A's are J's	PS	77	83	1	
All B's are T's	No B's are D's	No D's are T's	PS	73	90	4	
Some T's are Q's	Some L's are T's	Some L's are Q's	PS	83	100	2	c-a
Some J's are not P's	All C's are P's	Some C's are not J's	PS	80	90	3	
All B's are L's	No B's are N's	All L's are N's	I	7	20	4	
All G's are K's	All J's are G's	No K's are J's	I	3	7	2	
All R's are M's	No E's are M's	Some R's are E's	I	7	20	3	a-c
All C's are P's	All P's are F's	Some C's are not F's	I	23	13	1	
No T's are D's	Some T's are L's	All L's are D's	I	2	10	4	
No A's are L's	Some G's are L's	All G's are A's	I	7	3	3	
Some T's are G's	All G's are K's	No K's are Ts	I	7	20	1	c-a
All J's are Q's	All F's are J's	Some F's are not Q's	I	7	20	2	
No N's are T's	All N's are G's	All T's are G's	PW	0	17	4	
Some G's are M's	All R's are G's	No M's are R's	PW	10	20	2	
No B's are K's	All K's are E's	Some B's are E's	PW	10	17	1	a-c
All A's are B's	All E's are B's	Some A's are not E's	PW	23	30	3	
Some G's are not K's	All K's are J's	All J's are G's	PW	7	10	2	
Some C's are L's	All M's are C's	No M's are L's	PW	7	20	2	
All T's are P's	All T's are G's	Some G's are P's	PW	27	17	4	c-a
All N's are D's	All F's are D's	Some F's are not N's	PW	17	30	3	

* Necessity or possibility instructions

Appendix 2B
Written instructions (necessity) presented to participants prior to the reasoning task

---

<u>Instructions (N)</u>

The purpose of this experiment is to investigate how people solve logical reasoning problems.

A number of problems will be presented on the screen one at a time. Each problem consists of two statements which describe the relationship between three letters, followed by a conclusion. Your task is to indicate whether the conclusion <u>necessarily</u> follows from the sentences that precede it. A <u>necessary</u> conclusion is one that <u>must be true</u>, given the truth of the preceding sentences.

Below are examples of the screen layouts. First you will be shown two statements, and you should press the space bar to indicate your understanding of these.

> *Given that*
>
> All of the A's are B's
> Some of the B's are C's
>
> *Press space bar to continue*

A conclusion will then be added, and your task is to decide whether this conclusion <u>must be true</u>. Using the keyboard, you should press 'yes' if you think the conclusion <u>necessarily</u> follows and 'no' if you think the conclusion does <u>not necessarily</u> follow.

> *Given that*
>
> All of the A's are B's
> Some of the B's are C's
> *Is it necessary that*
>
> All of the A's are C's
>
> press either 'yes' or 'no' on the keyboard

You will then be asked to press the space bar when you are ready to continue to the next problem.

Note: Initially, you will be given two practice problems, but *please ask* the experimenter at any time if you are unsure on how to proceed.

Appendix 2C
Written instructions (possibility) presented to participants prior to the reasoning task

---

<u>Instructions (P)</u>

The purpose of this experiment is to investigate how people solve logical reasoning problems.

A number of problems will be presented on the screen one at a time. Each problem consists of two statements which describe the relationship between three letters, followed by a conclusion. Your task is to indicate whether the conclusion <u>possibly</u> follows from the sentences that precede it. A <u>possible</u> conclusion is one that <u>could be true</u>, given the truth of the preceding sentences.

Below are examples of the screen layouts. First you will be shown two statements, and you should press the space bar to indicate your understanding of these.

> *Given that*
>
> > None of the T's are D's
> > All of the D's are M's
>
> *Press space bar to continue*

A conclusion will then be added, and your task is to decide whether this conclusion <u>could be true</u>. Using the keyboard, you should press 'yes' if you think the conclusion is <u>possible</u> and 'no' if you think the conclusion is <u>impossible</u>.

> *Given that*
>
> > None of the T's are D's
> > All of the D's are M's
>
> *Is it possible that*
>
> > All of the T's are M's
>
> press either 'yes' or 'no' on the keyboard

You will then be asked to press the space bar when you are ready to continue to the next problem.

Note: Initially, you will be given two practice problems, but *please ask* the experimenter at any time if you are unsure on how to proceed.

Appendix 2D
A breakdown of the mean percentage endorsement rates and standard deviations for each reasoning problem; broken down into a-c and c-a direction conclusions

	Necessary				PS			
	a-c		c-a		a-c		c-a	
Necessary								
Low	83	*(19.86)*	73	*(30.75)*	71	*(28.69)*	77	*(27.02)*
High	79	*(27.92)*	76	*(2.71)*	68	*(30.90)*	68	*(38.92)*
Possibility								
Low	82	*(27.80)*	88	*(22.50)*	83	*(26.53)*	81	*(26.82)*
High	85	*(25.09)*	83	*(23.06)*	77	*(28.57)*	82	*(23.61)*

	Impossible				PW			
	a-c		c-a		a-c		c-a	
Necessary								
Low	22	*(26.86)*	23	*(25.72)*	23	*(31.44)*	41	*(25.83)*
High	13	*(17.04)*	10	*(16.87)*	18	*(21.71)*	33	*(21.92)*
Possibility								
Low	23	*(26.55)*	22	*(23.43)*	28	*(24.87)*	43	*(23.61)*
High	28	*(29.89)*	27	*(31.44)*	39	*(33.27)*	53	*(25.15)*

Appendix 2E
All ANOVA tables for experiment 1 (syllogistic reasoning)

Comparison of mean percentage endorsement rates between Necessary and PS problem types (within subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta^2
Instruction	4166.66	1	4166.67	9.02	.00	.14
Instruction*ability	41.66	1	41.67	.09	.77	.00
Error (instruction)	26807.29	58	462.20			
Problem	1760.42	1	1760.42	7.14	.01	.11
Problem*ability	260.42	1	260.42	1.06	.32	.02
Error (problem)	14307.29	58	246.68			
Instruction*problem	166.67	1	166.67	1.00	.32	.02
Instruction*problem*ability	41.66	1	41.67	.24	.63	.00
Error instruction*problem)	9713.54	58	167.48			

Comparison of mean percentage endorsement rates between low and high ability groups on Necessary and PS problem types (between subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta^2
Intercept	147266.66	1	142666.66	923.35	.00	.94
Ability	375.00	1	375.00	.24	.63	.00
Error	92505.21	58	1594.92			

Appendix 2E continued ......

Comparison of mean percentage endorsement rates between Impossible and PW problem types (within subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta²
Instruction	60000.00	1	60000.00	12.08	.00	.17
Instruction*ability	166.67	1	4166.67	8.39	.00	.13
Error (instruction)	28817.71	58	496.86			
Problem	11690.10	1	11690.10	48.40	.00	.46
Problem*ability	315.10	1	315.10	1.30	.26	.02
Error (problem)	14010.42	58	241.56			
Instruction*problem	260.42	1	260.42	1.35	.25	.02
Instruction*problem*ability	10.42	1	10.42	.05	.82	.00
Error instruction*problem)	11213.54	58	193.34			

Comparison of mean percentage endorsement rates between low and high ability groups on Impossible and PW problem types (between subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta²
Intercept	185648.44	1	185648.44	165.52	.00	.74
Ability	2.60	1	2.60	.00	.96	.00
Error	65052.08	58	1121.59			

Appendix 2E continued ......

Comparison of mean reasoning times between Necessary and PS problem types (within subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta²
Instruction	1749333.75	1	1749333.75	.06	.81	.00
Instruction*ability	98496.02	1	98496.02	.00	.95	.00
Error (instruction)	1.79	58	3.09			
Problem	304878.82	1	304878.82	.07	.80	.00
Problem*ability	1246176.82	1	1246176.82	.27	.61	.01
Error (problem)	2.73	58	4708196.54			
Instruction*problem	358981.35	1	358981.35	.08	.78	00
Instruction*problem*ability	899640.15	1	899640.15	.19	.67	.00
Error instruction*problem)	2.73	58	4747409.89			

Comparison of mean reasoning times between low and high ability groups on Necessary and PS problem types (between subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta²
Intercept	4.26	1	4.26	323.20	.00	.85
Ability	4.38	1	4.38	3.33	.07	.05
Error	7.64	58	1.31			

Appendix 2E continued ……

Comparison of mean reasoning times between Impossible and PW problem types (within subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta²
Instruction	4.22	1	4.22	2.00	.16	.03
Instruction*ability	3669932.02	1	3669932.02	.17	.68	.03
Error (instruction)	1.22	58	2.11			
Problem	5551650.01	1	5551650.01	.88	.35	.02
Problem*ability	76683.75	1	76683.75	.01	.91	.00
Error (problem)	3.66	58	6307501.81			
Instruction*problem	7828648.82	1	7828648.82	1.85	.18	.03
Instruction*problem*ability	1.40	1	1.40	3.32	.07	.05
Error instruction*problem)	2.50	58	4222831.14			

Comparison of mean reasoning times between low and high ability groups on Impossible and PW problem types (between subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta²
Intercept	3.93	1	3.93	368.43	.00	.86
Ability	5.43	1	5.43	5.09	.03	.08
Error	6.19	58	1.07			

Appendix 2F
A breakdown of the mean percentage endorsement rates and standard deviations for
a-c and c-a direction conclusions *(N = 60, SD in brackets)*

	N a-c		N c-a		PS a-c		PS c-a	
Necessary								
Low	83	*(29.86)*	73	*(30.75)*	71	*(28.68)*	77	*(20.02)*
High	79	*(27.92)*	76	*(29.71)*	68	*(30.90)*	68	*(38.92)*
M	81	*(24.08)*	75	*(30.00)*	69	*(29.60)*	72	*(33.53)*
Possibility								
Low	82	*(27.80)*	88	*(23.06)*	83	*(26.53)*	81	*(26.82)*
High	85	*(25.09)*	83	*(22.51)*	77	*(28.57)*	82	*(23.61)*
M	83	*(26.31)*	85	*(22.69)*	80	*(27.54)*	81	*(25.05)*

	I a-c		I c-a		PW a-c		PW c-a	
Necessary								
Low	21	*(26.86)*	23	*(25.72)*	2	*(31.44)*	41	*(25.83)*
High	13	*(17.04)*	10	*(16.87)*	18	*(21.71)*	33	*(21.92)*
M	18	*(22.69)*	16	*(22.47)*	21	*(26.91)*	37	*(24.12)*
Possibility								
Low	23	*(26.55)*	22	*(23.43)*	28	*(24.87)*	43	*(23.61)*
High	28	*(29.89)*	27	*(31.44)*	39	*(33.72)*	53	*(26.04)*
M	25	*(28.18)*	24	*(27.60)*	33	*(29.71)*	48	*(25.16)*

Appendix 3A
A full set of semantic descriptions of the 9 line relationships used in experiments 2 and 3, with instructions at the beginning

---

Instructions: *Enter the number of the diagram to which each sentence refers, in the box next to it:*

The red line overlaps the blue line from the left

The red line equals the blue line

The red line is surrounded by the blue line

The red line surrounds the blue line

The red line touches the blue line at the right

The red line lies to the left of the blue line

The red line lies to the right of the blue line

The red line touches the blue line at the left

The red line overlaps the blue line from the right

Appendix 3B
A full set of the 9 line relationships used in experiments 2 and 3, which was presented
to participants with the semantic descriptions shown in appendix 3A
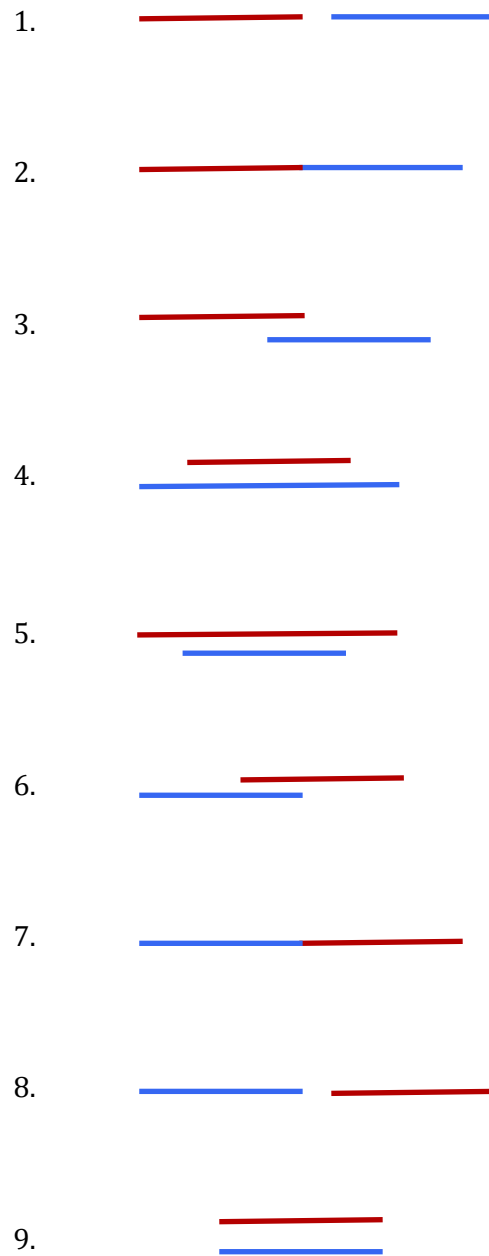
Appendix 3C
Written instructions (necessity) presented to participants, for experiment 2 and 3,
prior to the transitive inference

Instructions (N)

A number of problems will be presented on the screen one at a time. Each problem consists of two statements describing the relationship between a red and a green line, and a blue and a green line, followed by a conclusion indicating the relationship between the red and the blue line. Your task is to say whether the conclusion <u>necessarily</u> follows from the sentences that precede it. A <u>necessary</u> conclusion is one that <u>must be true</u>, given the truth of the preceding sentences.

Below are examples of the screen layouts. First you will be shown two statements, and you should press the space bar to indicate your understanding of these.

> *Given that*
> The red line surrounds the green line
> The blue line lies to the left of the green line
>
> press the space bar to continue

A conclusion will then be added to the screen, and your task is to decide whether this conclusion <u>must be true.</u> Using the keyboard, you should press 'yes' if you think the conclusion <u>necessarily</u> follows and 'no' if you think the conclusion does <u>not necessarily</u> follow. .For example:

> *Given that*
> The red line surrounds the green line
> The blue line lies to the left of the green line
> *Is it necessary that*
> The red line lies to the left of the blue line
>
> press either 'yes' or 'no' on the keyboard

You will then be asked to press the space bar when you are ready to continue to the next problem. Initially, you will be given four practice problems, but please ask the experimenter at any time if you are unsure on how to proceed.

Note: Responses are timed, and it is important that you answer the questions both carefully and accurately.

Appendix 3D
Written instruction (possibility) presented to participants, for experiment 2 and 3,
prior to the transitive inference

<u>Instructions (P)</u>

A number of problems will be presented on the screen one at a time. Each problem
consists of two statements describing the relationship between a red and a green line,
and a blue and a green line; followed by a conclusion indicating the relationship
between the red and the blue line. Your task is to say whether the conclusion <u>possibly</u>
follows from the sentences that precede it. A <u>possible</u> conclusion is one that <u>could be
true</u>, given the truth of the preceding sentences.

Below are examples of the screen layouts. First you will be shown two statements, and
you should press the space bar to indicate your understanding of these.

> *Given that*
> The red line surrounds the green line
> The green line touches the blue line at the left
>
> press the space bar to continue

A conclusion will be added to the screen, and your task is to decide whether this
conclusion <u>could be true</u>. Using the keyboard, you should press 'yes' if you think the
conclusion is <u>possible</u> and 'no' if you think the conclusion is <u>impossible</u>. For example:

> *Given that*
> The red line surrounds the green line
> The green line touches the blue line at the left
> *Is it possible that*
> The red line surrounds the blue line
>
> press either 'yes' or 'no' on the keyboard

You will then be asked to press the space bar when you are ready to continue to the
next problem. Initially, you will be given four practice problems, but please ask the
experimenter at any time if you are unsure on how to proceed.

Note: Responses are timed, and it is important that you answer the questions both
carefully and accurately.

Appendix 3E
All ANOVA tables for experiment 2 (transitive inference)

Comparison of mean percentage endorsement rates between Necessary and PS problem types (within subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta^2
Instruction	78.78	1	78.78	.29	.59	.00
Instruction*ability	52.73	1	52.73	.19	.66	.00
Error (instruction)	15766.93	58	271.84			
Problem	4271.48	1	4271.48	11.93	.00	.17
Problem*ability	235.03	1	235.03	.66	.42	.01
Error (problem)	20766.93	58	358.05			
Instruction*problem	287.11	1	287.11	1.28	.26	.02
Instruction*problem*ability	406.90	1	406.90	1.81	.18	.03
Error (instruction*problem)	13016.93	59	224.43			

Comparison of mean percentage endorsement rates between low and high ability groups on Necessary and PS problem types (between subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta^2
Intercept	1156828.78	1	1156828.78	1237.36	.00	.95
Ability	1094.40	1	1094.40	1.17	.28	.02
Error	54225.26	58	934.92			

Appendix 3E continued ......

Comparison of mean percentage endorsement rates between Impossible and PW problem types (within subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta^2
Instruction	13146.84	1	13146.84	22.75	.00	.28
Instruction*ability	1503.50	1	1503.50	2.60	.11	.04
Error (instruction)	33514.30	58	577.83			
Problem	23028.05	1	23028.05	33.78	.00	.37
Problem*ability	443.09	1	443.09	.64	.43	.01
Error (problem)	40014.75	58	689.91			
Instruction*problem	14631.26	1	14631.26	32.44		.36
Instruction*problem*ability	258.55	1	258.55	.57		.01
Error (instruction*problem)	261.58	58	451.01			

Comparison of mean percentage endorsement rates between low and high ability groups on Impossible and PW problem types (between subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta^2
Intercept	356456.46	1	356456.46	663.97	.00	.92
Ability	2.88	1	2.88	.01	.94	.00
Error	31137.80	58	36.86			

Appendix 3E continued ......

Comparison of mean reasoning times between Necessary and PS problem types (within subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta^2
Instruction	4.77	1	4.77	.56	.46	.01
Instruction*ability	1.61	1	1.61	1.37	.25	.02
Error (instruction)	4.93	58	8.50			
Problem	2.78	1	2.78	1.59	.21	.03
Problem*ability	1.19	1	1.19	6.77	.01	.11
Error (problem)	1.02	58	1.75			
Instruction*problem	4.83	1	4.83	4.10	.05	.07
Instruction*problem*ability	2899052.11	1	2899052.11	2.50	.62	.00
Error (instruction*problem)	6.83	58	1.18			

Comparison of mean reasoning times between low and high ability groups on Necessary and PS problem types (between subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta^2
Intercept	9.64	1	9.64	899.83	.00	.94
Ability	7.84	1	7.84	.73	.40	.01
Error	6.22	58	1.07			

Appendix 3E continued ......

Comparison of mean reasoning times between Impossible and PW problem types (within subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta^2
Instruction	2.43	1	2.43	1.91	.66	.00
Instruction*ability	8.02	1	8.02	6.32	.43	.01
Error (instruction)	7.37	58	1.27			
Problem	1.17	1	1.17	7.93	.38	.01
Problem*ability	4622040.15	1	4622040.15	3.10	.58	.00
Error (problem)	8.60	58	1.48			
Instruction*problem	1.89	1	1.89	.90	.35	.01
Instruction*problem*ability	4151126.31	1	4151126.31	.20	.66	.00
Error (instruction*problem)	1.22	58	2.10			

Comparison of mean reasoning times between low and high ability groups on Impossible and PW problem types (between subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta^2
Intercept	9.93	1	9.93	780.83	.00	.93
Ability	1.67	1	1.67	1.31	.26	.02
Error	7.38	58	1.27			

Appendix 3F
All ANOVA tables for experiment 3 (transitive inference)

Comparison of mean percentage endorsement rates between Necessary and PS problem types (within subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta^2
Instruction	2490.38	1	2490.38	7.62	.01	.12
Instruction*ability	4134.98	1	4134.98	12.65	.00	.18
Error (instruction)	18637.80	57	326.98			
Problem	8740.41	1	8740.41	23.34	.00	.29
Problem*ability	1926.32	1	1926.32	5.14	.03	.08
Error (problem)	21344.33	57	374.46			
Instruction*problem	3102.14	1	3102.14	14.64	.00	.20
Instruction*problem*ability	488.26	1	488.26	2.31	.14	.04
Error (instruction*problem)	12075.30	57	211.85			

Comparison of mean percentage endorsement rates between low and high ability groups on Necessary and PS problem types (between subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta^2
Intercept	1172080.49	1	1172080.49	1619.66	.00	.97
Ability	1198.61	1	1198.61	1.66	.20	.03
Error	41248.43	57	723.66			

Appendix 3F continued ......

Comparison of mean percentage endorsement rates between Impossible and PW problem types (within subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta²
Instruction	27316.32	1	27316.32	58.33	.00	.51
Instruction*ability	3011.36	1	3011.36	6.43	.01	.10
Error (instruction)	26692.39	57	468.29			
Problem	20949.38	1	20949.38	31.76	.00	.36
Problem*ability	16.50	1	16.50	.03	.88	.00
Error (problem)	37600.67	57	659.66			
Instruction*problem	22213.66	1	22213.66	.00	.00	.48
Instruction*problem*ability	3345.69	1	3345.69	.01	.01	.12
Error (instruction*problem)	24101.87	57	422.84			

Comparison of mean percentage endorsement rates between low and high ability groups on Impossible and PW problem types (between subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta²
Intercept	265753.56	1	265753.56	427.32	.00	.88
Ability	298.95	1	298.95	.48	.49	.01
Error	35448.81	57	621.91			

Appendix 3F continued ......

Comparison of mean reasoning times between Necessary and PS problem types (within subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta^2
Instruction	5.73	1	5.73	1.03	.32	.02
Instruction*ability	2.85	1	2.85	.51	.48	.01
Error (instruction)	3.17	57	5.56			
Problem	7.98	1	7.98	3.81	.06	.06
Problem*ability	9.02	1	9.02	4.31	.04	.07
Error (problem)	1.19	57	2.09			
Instruction*problem	1.38	1	1.38	.90	.35	.02
Instruction*problem*ability	39111.41	1	39111.41	.96	.96	.00
Error (instruction*problem)	8.81	57	1.55			

Comparison of mean reasoning times between low and high ability groups on Necessary and PS problem types (between subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta^2
Intercept	9.34	1	9.34	1.03	.00	.89
Ability	7.97	1	7.97	.51	.05	.07
Error	1.16	57	2.03			

Appendix 3F continued ......

Comparison of mean reasoning times between Impossible and PW problem types (within subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta²
Instruction	5.25	1	5.25	1.36	.25	.02
Instruction*ability	2.17	1	2.17	.57	.46	.01
Error (instruction)	2.19	57	3.85			
Problem	808.75	1	808.75	00	1.00	.00
Problem*ability	6839648.08	1	6839648.08	.35	.59	.01
Error (problem)	1.12	57	1.97			
Instruction*problem	2.08	1	2.08	1.83	.18	.03
Instruction*problem*ability	4058195.57	1	4058195.57	.55	.55	.01
Error (instruction*problem)	6.48	57	1.14			

Comparison of mean reasoning times between low and high ability groups on Impossible and PW problem types (between subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta²
Intercept	8.67	1	8.67	544.22	.00	.91
Ability	4.77	1	4.77	2.93	.09	.05
Error	9.08	57	1.59			

Appendix 4A
A complete set of conditional inference problem, showing problem category and
argument form, premise one/premise two, and conclusion

N (MP)	if the letter is an A then the number is a 2, the letter is an A	the number is a 2
	if the letter is a B then the number is not an 8, the letter is a B	the number is a not an 8
	if the letter is not an N then the number is a 5, the letter is not an N	the number is a 5
	If the letter if not a J then the number if not a 2, the letter is not a J	the number is not a 2
N (MT)	if the letter is a C then the number is an 8, the number is not an 8	the letter is not a C
	If the letter is a T then the number is not a 9, the number is a 9	the letter is not a T
	if the letter is not a G then the number is a 3, the number is not a 3	the letter is a G
	if the letter is not a D then the number is not a 6, the number is a 6	the letter is a D
I (MP)	if the letter is a Z then the number is a 3, the letter is a Z	the number is not a 3
	if the letter is a K then the number is not a 2, the letter is a K	the number is a 2
	if the letter is not an M then the number is a 6, the letter is not an M	the number is not a 6
	if the letter is not an H then the number is not a 3, the letter is not an H	the number is a 3
I (MT)	if the letter is a B then the number is a 9, the number is not a 9	the letter is a B
	if the letter is a V then the number is not an 8, the number is a 8	the letter is a V
	if the letter is not an H then the number is a 2, the number is not a 2	the letter is not an H
	if the letter is not an F then the number is not a 4, the number is a 4	the letter is not an F
PS (AC)	If the letter is a C then the number is a 7, the number is a 7	the letter is a C
	if the letter is a P then the number is not a 4, the number is not a 4	the letter is a P
	if the letter is not a Z then the number is a 7, the number is a 7	the letter is not a Z
	if the letter is not an F then the number is not a 5, the number is not a 5	the letter is not an F
PS (DA)	if the letter is an L then the number is a 5, the letter is not an L	the number is not a 5
	if the letter is a D then the number is not a 7, the letter is not a D	the number is a 7
	if the letter is not an L then the number is a 7, the letter is an L	the number is not a 7
	if the letter is not a J then the number is not a 9, the letter is a J	the number is a 9
PW (AC)	if the letter is a W then the number is a 6, the number is a 6	the letter is not a W
	if the letter is an R then the number is not a 3, the number is not a 3	the letter is not an R
	if the letter is not an A then the number is a 9, the number is a 9	the letter is an A
	if the letter is not an E then the number is not a 6, the number is not a 6	the letter is an E
PW (DA)	if the letter is a K then the number is an 8, the letter is not a K	the number is an 8
	if the letter is a Y then the number is not a 4, the letter is not a Y	the number is a not a 4
	If the letter is not an E then the number is a 5, the letter is an E	the number is a 5
	if the letter is not a G then the number is not a 4, the letter is a G	the number is not a 4

Appendix 4B
Written instructions (necessity) presented to participants prior to the reasoning task

---

Instructions (N)

A number of problems will be presented one at a time. For each problem you will be shown a statement that you must consider to be true. Following this you will be given a second statement, and your task is to indicate whether the conclusion <u>necessarily</u> follows from the sentence that precedes it. A <u>necessary</u> conclusion is one that <u>must be true</u>, given the truth of the preceding sentence. Below are examples of the screen layouts.

> *Fact:*
> If the letter is a T then the number is a 4
> The letter is a T
>
> press the space bar to continue

A conclusion will then be added to the screen, and your task is to decide whether this conclusion <u>must be true.</u> Using the keyboard, you should press 'yes' if you think the conclusion <u>necessarily</u> follows and 'no' if you think the conclusion does <u>not necessarily</u> follow. For example:

> *Fact:*
> If the letter is a T then the number is a 4
> The letter is a T
> *Is it necessary that*
> The number is a 4
>
> Press either 'yes' or no' on the keyboard

You will then be asked to press the space bar when you are ready to continue to the next problem. Initially, you will be given four practice problems, but please ask the experimenter at any time if you are unsure on how to proceed.

Note: Responses are timed, and it is important that you answer the questions both carefully and accurately.

Appendix 4C
Written instructions (possibility) presented to participants prior to the reasoning task

---

Instructions (P)

A number of problems will be presented one at a time. For each problem you will be shown a statement that you must consider to be true. Following this you will be given a second statement, and your task is to indicate whether the conclusion <u>possibly</u> follows from the sentence that precedes it. A <u>possible</u> conclusion is one that <u>could be true</u>, given the truth of the preceding sentence. Below are examples of the screen layouts.

> *Fact:*
> If the letter is a B then the number is a 6
> The number is a 6
>
> press the space bar to continue

A conclusion will be added to the screen. Your task is to decide whether this conclusion <u>could be true</u>. Using the keyboard, you should press 'yes' if you think the conclusion <u>is possible</u> and 'no' if you think the conclusion is <u>impossible</u>. For example:

> *Fact:*
> If the letter is a B then the number is a 6
> the number is a 6
> *Is it possible that*
> The letter is a B
>
> press either 'yes' or 'no' on the keyboard

You will then be asked to press the space bar when you are ready to continue to the next problem. Initially, you will be given four practice problems, but please ask the experimenter at any time if you are unsure on how to proceed.

Note: Responses are timed, and it is important that you answer the questions both carefully and accurately.

Appendix 4D
Breakdown of conditional inference endorsement rates into MP, MT, AC and DA argument forms. There were 30 participants in each ability group (*SD* shown in brackets)

	Necessary				Possible strong			
	MP		MT		AC		DA	
Necessary								
Low	89	*(19.35)*	50	*(34.14)*	80	*(27.38)*	53	*(34.34)*
High	95	*(12.06)*	45	*(30.37)*	65	*(36.32)*	36	*(33.27)*
Possibility								
Low	92	*(15.16)*	74	*(24.11)*	98	*(11.24)*	72	*(21.51)*
High	98	*(7.63)*	81	*(20.34)*	98	*(6.34)*	87	*(19.40)*

	Impossible				Possible weak			
	MP*		MT*		AC*		DA*	
Necessary								
Low	13	*(14.31)*	22	*(21.51)*	17	*(26.53)*	18	*(17.55)*
High	9	*(17.96)*	18	*(20.91)*	7	*(17.29)*	14	*(21.46)*
Possibility								
Low	13	*(18.08)*	36	*(24.29)*	33	*(33.57)*	42	*(36.16)*
High	15	*(20.34)*	49	*(24.11)*	50	*(35.96)*	61	*(35.77)*

*Conclusion presented in opposite direction

Appendix 4E
All ANOVA tables for experiment 4 (abstract conditionals)

Comparison of mean percentage endorsement rates between Necessary and PS problem types (within subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta²
Instruction	30940.10	1	30940.10	47.09	.00	.45
Instruction*ability	3760.42	1	3760.42	5.72	.02	.09
Error (instruction)	38111.98	58	657.10			
Problem	1500.00	1	1500.00	8.34	.01	.13
Problem*ability	585.94	1	585.94	3.26	.08	.05
Error (problem)	10414.06	58	179.55			
Instruction*problem	2502.61	1	2502.61	9.84	.00	.15
Instruction*problem*ability	1500.00	1	1500.00	5.90	.02	.09
Error (instruction*problem)	14747.40	58	254.27			

Comparison of mean percentage endorsement rates between low and high ability groups on Necessary and PS problem types (between subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta²
Intercept	1365041.67	1	1365041.67	2758.07	.00	.97
Ability	2.60	1	2.60	.01	.94	.03
Error	28705.72	58	494.93			

Appendix 4E continued ......

Comparison of mean percentage endorsement rates between Impossible and PW problem types (within subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta^2
Instruction	30940.10	1	30940.10	44.39	.00	.43
Instruction*ability	5041.67	1	5041.67	7.23	.01	.11
Error (instruction)	40424.48	58	696.97			
Problem	4166.67	1	4166.67	14.73	.00	.20
Problem*ability	210.94	1	210.94	7.45	.39	.01
Error (problem)	16403.65	58	282.82			
Instruction*problem	5752.60	1	5752.60	18.58	.00	.24
Instruction*problem*ability	666.67	1	666.67	1.49	.15	.05
Error (instruction*problem)	17955.73	58	309.58			

Comparison of mean percentage endorsement rates between low and high ability groups on Impossible and PW problem types (between subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta^2
Intercept	160166.67	1	160166.67	279.24	.00	.83
Ability	940.10	1	940.10	1.64	.21	.03
Error	33268.23	58	573.59			

Appendix 4E continued ......

Comparison of mean reasoning times between Necessary and PS problem types (within subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta²
Instruction	3635389.35	1	3635389.35	.96	.33	.02
Instruction*ability	3227700.23	1	3227700.23	.85	.36	.01
Error (instruction)	2.20	58	3789614.54			
Problem	2348677.35	1	2348677.35	2.30	.14	.04
Problem*ability	1119541.90	1	1119541.90	1.09	.30	.02
Error (problem)	5.95	58	1025115.48			
Instruction*problem	1276.51	1	1276.51	.00	9.74	.00
Instruction*problem*ability	674266.00	1	674266.00	.55	.46	.01
Error (instruction*problem)	7.10	58	1223367.59			

Comparison of mean reasoning times between low and high ability groups on Necessary and PS problem types (between subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta²
Intercept	6.88	1	6.88	657.43	.00	.99
Ability	1.95	1	1.95	1.87	.17	.03
Error	6.07	58	1.05			

Appendix 4E continued .....

Comparison of mean reasoning times between Impossible and PW problem types (within subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta²
Instruction	629888.22	1	629888.22	.21	.65	.00
Instruction*ability	5694227.25	1	5694227.25	1.87	.18	.03
Error (instruction)	1.77	58	3053881.75			
Problem	501260.45	1	501260.45	1.16	.29	.02
Problem*ability	2992255.01	1	2992255.01	4.65	.04	.07
Error (problem)	2.50	58	430855.68			
Instruction*problem	3343.20	1	3343.20	.00	.95	.00
Instruction*problem*ability	531406.23	1	531406.23	.63	.43	.01
Error (instruction*problem)	4.92	58	848212.61			

Comparison of mean reasoning times between low and high ability groups on Impossible and PW problem types (between subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta²
Intercept	7.14	1	7.14	623.56	.00	.92
Ability	1.38	1	1.38	1.20	.28	.02
Error	6.65	58	1.15			

Appendix 5A
The eight problem structures with set A for Necessary and PS, and set B for Impossible and PW (experiment 5)
* Indicates that the conclusion was presented in the opposite direction

	Premises	Conclusion
N (MP)	Few If butter is heated it will melt, the butter was heated Many If a stone if kicked it will move, the stone was kicked	the butter melted the stone moved
N (MT)	Few If butter is heated it will melt, the butter did not melt Many If a stone is kicked it will move, the stone did not move	the butter was not heated the stone was not kicked
PS (AC)	Few If Simon cuts his finger it will bleed, Simons finger bled Many If the brake is pressed the car will slow down, the car slowed down	Simon cut his finger The brake was pressed
PS (DA)	Few If Simon cuts his finger it will bleed, Simon did not cut his finger Many If the brake is pressed the car will slow down, the brake was not pressed	Simon's finger did not bleed The car did not slow down
I (MP)	Few If the paperclip touches the magnet it will stick to it, the paper clip touched the magnet Many If the window is open the room will be cold, the window is open	The paper clip did not stick to the magnet* The room was not cold*
I (MT)	Few If the paper clip touches the magnet it will stick to it, the paper clip did not stick to it, Many If the window is open the room will be cold, the room was not cold	The paper clip did not touch the magnet* The window was opened*
PW (AC)	Few If water is frozen it will become ice, the water became ice Many If the mug is dropped it will break, the mug broke	The water was not frozen* The mug was not dropped*
PW (DA)	Few If water is frozen it will become ice, the water was not frozen Many If the mug is dropped it will break, the mug was not dropped	The water became ice* The mug broke*

Appendix 5B
The eight problem structures with set B for Necessary and PS, and set A for Impossible and PW (experiment 5)
* Indicates that the conclusion was presented in the opposite direction

		Premises	Conclusion
N (MP)	Few	If the paperclip touches the magnet it will stick to it, the paper clip touched the magnet	The paper clip stuck to the magnet
	Many	If the window is open the room will be cold, the window is open	The room was cold
N (MT)	Few	If the paper clip touches the magnet it will stick to it, the paper clip did not stick to it,	The paper clip touched the magnet
	Many	If the window is open the room will be cold, the room was not cold	The window was not opened
PS (AC)	Few	If water is frozen it will become ice, the water became ice	The water was frozen
	Many	If the mug is dropped it will break, the mug broke	The mug was dropped
PS (DA)	Few	If water is frozen it will become ice, the water was not frozen	The water did not became ice
	Many	If the mug is dropped it will break, the mug was not dropped	The mug did not break
I (MP)	Few	If butter is heated it will melt, the butter was heated	the butter did not melt*
	Many	If a stone if kicked it will move, the stone was kicked	the stone did not move*
I (MT)	Few	If butter is heated it will melt, the butter did not melt	the butter was heated*
	Many	If a stone is kicked it will move, the stone did not move	the stone was kicked*
PW (AC)	Few	If Simon cuts his finger it will bleed, Simons finger bled	Simon did not cut his finger*
	Many	If the brake is pressed the car will slow down, the car slowed down	The brake was not pressed*
PW (DA)	Few	If Simon cuts his finger it will bleed, Simon did not cut his finger	Simon's finger bled*
	Many	If the break is pressed the car will slow down, the brake was not pressed	The car slowed down*

Appendix 5C
Written instructions (necessity) presented to participants, for experiment 5, prior to the everyday conditional inference task

---

<u>Instructions (N)</u>

A number of problems will be presented one at a time. For each problem you will be shown two statements that you must consider to be true. Following this you will be given a conclusion, and your task is to indicate whether the conclusion <u>necessarily</u> follows from the sentences that precede it. A <u>necessary</u> conclusion is one that <u>must be true</u>, given the truth of the preceding sentence. Below are examples of the screen layouts.

> *Fact:*
>
> if an egg is boiled, then it will become solid
>
> press space bar to continue

A conclusion will then be added to the screen, and your task is to decide whether this conclusion <u>must be true.</u> Using the keyboard, you should press 'yes' if you think the conclusion <u>necessarily</u> follows and 'no' if you think the conclusion does <u>not necessarily</u> follow. For example:

> *Fact:*
>
> if an egg is boiled, then it will become solid
> the egg was boiled
> is it necessary that:
> the egg was solid
>
> press either 'yes' or 'no' on the keyboard

You will then be asked to press the space bar when you are ready to continue to the next problem. Initially, you will be given two practice problems, but please ask the experimenter at any time if you are unsure on how to proceed.

Note: Responses are timed, and it is important that you answer the questions both carefully and accurately.

Appendix 5D
Written instructions (possibility) presented to participants for experiment 5 prior to the everyday conditional inference task

_____

<u>Instructions (P)</u>

A number of problems will be presented one at a time. For each problem you will be shown two statements that you consider to be true. Following this you will be given a conclusion, and your task is to indicate whether the conclusion <u>possibly</u> follows from the sentences that precede it. A <u>possible</u> conclusion is one that <u>could be true</u>, given the truth of the preceding sentence. Below are examples of the screen layouts.

Fact:
if the gong is struck, then it will sound
press the space bar to continue

A conclusion will be added to the screen. Your task is to decide whether this conclusion <u>could be true</u>. Using the keyboard, you should press 'yes' if you think the conclusion <u>necessarily</u> follows and 'no' if you think the conclusion does <u>not necessarily</u> follow. For example:

Fact:
if the gong is struck, then it will sound
the gong was struck
is it possible that:
the gong sounded
press either 'yes' or 'no' on the keyboard

You will then be asked to press the space bar when you are ready to continue to the next problem. Initially, you will be given two practice problems, but please ask the experimenter at any time if you are unsure on how to proceed.

Note: Responses are timed, and it is important that you answer the questions both carefully and accurately.

Appendix 5E
Breakdown of conditional inference endorsement rates into MP, MT, AC and DA argument forms for experiment 5. There were 30 participants in each ability group (*SD* shown in brackets)

	N (few)				N (many)			
	MP		MT		MP		MT	
Necessity								
Low	83	(33.05)	80	(33.73)	96	(20.34)	87	(21.30)
High	95	(15.26)	83	(33.05)	97	(12.68)	88	(25.20)
Possibility								
Low	100	(0)	95	(15.26)	98	(9.13)	100	(0)
High	100	(0)	98	(9.13)	98	(9.13)	97	(12.69)

	N (few)				N (many)			
	AC		DA		AC		DA	
Necessity								
Low	78	(33.95)	82	(27.80)	63	45.36	65	(41.83)
High	82	(38.25)	78	(38.69)	70	42.75	68	(38.25)
Possibility								
Low	100	(0)	95	(15.26)	98	(9.13)	92	(18.90)
High	100	(0)	95	(15.26)	100	(0)	97	(12.69)

Appendix 5E continued ......

	I (few)				I (many)			
	MP*		MT*		MP*		MT*	
Necessity								
Low	18	(30.75)	12	(28.42)	27	(36.52)	27	(38.80)
High	2	(9.13)	2	(9.18)	7	(25.37)	7	(25.37)
Possibility								
Low	13	(26.04)	13	(22.07)	41	(43.71)	40	(44.34)
High	3	(12.69)	7	(17.29)	38	(48.57)	37	(43.42)

*Opposite direction conclusion

	PW (few)				PW (many)			
	AC*		DA*		AC*		DA*	
Necessity								
Low	3	(18.26)	10	(20.34)	10	(30.51)	5	(15.26)
High	17	(37.91)	5	(15.26)	7	(25.37)	5	(20.13)
Possibility								
Low	13	(34.58)	10	(20.34)	47	(50.74)	42	(41.70)
High	17	(37.90)	30	(38.51)	50	(50.86)	53	(43.42)

*Opposite direction conclusion

Appendix 5F
All ANOVA tables for experiment 5 (everyday causal conditionals)

Comparison of mean percentage endorsement rates between Necessary and PS inferences (within subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta^2
Instruction	36750.00	1	36750.00	32.14	.00	.36
Instruction*ability	520.83	1	520.83	.46	.50	.01
Error (instruction)	66322.92	58	1143.50			
Problem	6750.00	1	6750.00	16.79	.00	.22
Problem*ability	83.33	1	83.33	.21	.65	.00
Error (problem)	23322.92	58	402.12			
Few/many	1020.83	1	1020.83	5.90	.02	.09
Few/many*ability	20.83	1	20.83	.12	.73	.00
Error (few/many)	10052.08	58	173.31			
Instruction*problem	4687.50	1	4687.50	9.84	.00	.15
Instruction*problem*ability	333.33	1	333.33	.70	.41	.00
Error(instruction*problem)	27635.42	58	476.47			
Instruction*few/many	750.00	1	750.00	4.79	.03	.08
Instruction (few/many*ability)	20.83	1	20.83	.13	.72	.00
Error (instruction*few/many)	972.92	58	156.43			
Problem*few/many	2083.33	1	2083.33	18.62	.00	.24
Problem*few/many*ability	333.33	1	333.33	2.98	.09	.05
Error (problem*few/many)	6489.58	58	111.89			
Instruction*problem*few/many	1687.50	1	1687.50	18.32	.00	.24
Instruction*problem*few/many*ability	.00	1	.00	.00	1.00	.00
Error (instruction*problem*few/many	5373.75	58	92.13			

Appendix 5F continued ......

Comparison of mean percentage endorsement rates between low and high ability groups on Necessary and PS problem types (between subjects)

	Sum of Squares	df	F	Sig.	Partial Eta²
Intercept	3798520.83	1	3406.37	.00	.98
Ability	1020.83	1	.92	.34	.02
Error	64677.08	58			

Appendix 5F continued ......

Comparison of mean percentage endorsement rates between Impossible and PW inferences (within subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta^2
Instruction	41255.21	1	41255.21	25.49	.00	.31
Instruction*ability	2520.83	1	2520.83	1.56	.22	.03
Error (instruction)	93880.21	58	1618.62			
Problem	520.83	1	520.83	.73	.40	.01
Problem*ability	7921.88	1	7921.88	11.10	.00	.16
Error (problem)	41401.04	58	713.81			
Few/many	34171.88	1	34171.88	44.02	.00	.43
Few/many*ability	333.33	1	333.33	.43	.52	.01
Error (few/many)	45026.04	58	776.31			
Instruction*problem	5671.88	1	5671.88	10.19	.00	.15
Instruction*problem*ability	833.33	1	833.33	.15	.70	.00
Error(instruction*problem)	32838.54	58	566.18			
Instruction*few/many	22687.50	1	22687.50	35.26	.00	.38
Instruction (few/many*ability)	255.21	1	255.21	.40	.53	.01
Error (instruction*few/many)	37213.54	58	641.61			
Problem*few/many	630.21	1	630.21	1.70	1.98	.03
Problem*few/many*ability	187.50	1	187.50	.51	.48	.01
Error (problem*few/many)	21526.04	58	371.14			
Instruction*problem*few/many	1020.83	1	1020.83	3.69	.06	.06
Instruction*problem*few/many*ability	255.21	1	255.21	.92	.02	.02
Error (instruction*problem*few/many	16067.71	58	277.03			

Appendix 5F continued ......

Comparison of mean percentage endorsement rates between low and high ability groups on Necessary and PS inferences (between subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta²
Intercept	180187.50	1	180187.50	129.55	.00	.69
Ability	1171.88	1	1171.88	.84	.36	.01
Error	80671.88	58	1390.89			

Appendix 5F continued ……

Comparison of mean reasoning times between Necessary and PS inferences types (within subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta^2
Instruction	2.08	1	2.08	3.36	.07	.06
Instruction*ability	6841858.85	1	6841858.85	1.10	.30	.02
Error (instruction)	3.59	58	2.09			
Problem	139571.30	1	139571.30	.06	.81	.00
Problem*ability	7181211.50	1	7181211.50	3.04	.09	.05
Error (problem)	1.37	58	6841858.85			
Few/many	304688.20	1	304688.20	.10	.75	.00
Few/many*ability	1774752.02	1	1774752.02	.60	.44	01
Error (few/many)	1.71	58	2965861.02			
Instruction*problem	86322.85	1	86322.85	.04	.84	.00
Instruction*problem*ability	1183358.10	1	1183358.10	.60	.44	.01
Error(instruction*problem)	1.14	58	1960752.68			
Instruction*few/many	204476.35	1	204476.35	.08	.77	.00
Instruction (few/many*ability)	264751.10	1	264751.10	.11	.74	.00
Error (instruction*few/many)	1.41	58	2424040.02			
Problem*few/many	7736586.92	1	7736586.92	2.53	.12	.04
Problem*few/many*ability	390621.35	1	390621.35	.13	.72	.00
Error (problem*few/many)	1.77	58	3054563.65			
Instruction*problem*few/many	1025362.97	1	1025362.97	.52	.48	.10
Instruction*problem*few/many*ability	557262.55	1	557262.55	.28	.60	01
Error (instruction*problem*few/many)	1.15	58	1982234.93			

Appendix 5F continued ......

Comparison of mean reasoning times between low and high ability groups on Necessary and PS inferences (between subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta^2
Intercept	2.02	1	2.02	963.47	.00	.94
Ability	1.93	1	1.93	.91	.35	.02
Error	1.23	58	2.12			

Appendix 5F continued ......

Comparison of mean reasoning times between Impossible and PW inferences
(within subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta²
Instruction	1.20	1	1.20	.88	.35	.02
Instruction*ability	884856.00	1	884856.00	.06	.80	.00
Error (instruction)	7.98	58	1.38			
Problem	655862.60	1	655862.60	.11	.75	.00
Problem*ability	660.35	1	660.35	.00	.99	.00
Error (problem)	3.57	58	6165916.63			
Few/many	5.85	1	5.85	13.15	.00	.18
Few/many*ability	1462468.80	1	1462468.80	.33	.60	.01
Error (few/many)	2.58	58	4452560.17			
Instruction*problem	20206.30	1	20206.30	.01	.94	.00
Instruction*problem*ability	3653506.52	1	3653506.52	.89	.35	.02
Error(instruction*problem)	2.39	58	4119495.32			
Instruction*few/many	7623756.35	1	7623756.35	1.68	.20	.03
Instruction (few/many*ability)	1037043	1	1037043	2.29	.64	.00
Error (instruction*few/many)	2.64	58	4546430.94			
Problem*few/many	430980.60	1	430980.60	.08	.77	.00
Problem*few/many*ability	27015.00	1	27015.00	.01	.94	.00
Error (problem*few/many)	3.00	58				
Instruction*problem*few/many	5480763.92	1	5480763.92	1.28	.26	.02
Instruction*problem*few/many*ability	4.10	1	4.10	9.58	.00	.14
Error (instruction*problem*few/many	2.48	58				

Appendix 5F continued ......

Comparison of mean percentage endorsement rates between low and high ability groups on Necessary and PS inferences (between subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta²
Intercept	3.05	1	3.05	937.67	.00	.94
Ability	1.33	1	1.33	4.07	.05	.07
Error	1.89	58	3.26			

Appendix 5G
The mean ratings for all of the 64 inferences used in the pilot study for experiment 6
(the shaded inferences were selected for experiment 6, as per selection criteria)

N	if water temperature falls below zero	it will freeze	97.20
N	if butter is heated	it will melt	96.20
N	if a person dies	they will stop breathing	94.60
N	if the bucket is lowered into a dry well	it will come up empty	91.40
N	if the balloon is pricked	it will burst	91.40
N	if the dog has fleas	it will scratch itself	91.20
N	if the Pope dies	a new pope will be elected	90.90
N	if it rains heavily	the streets will get wet	90.60
N	if black paint is added to white paint	it will turn grey	90.50
N	if the water is carbonated	it will get bubbles	90.10
N	if it has warm blood	it will be a mammal	89.90
N	if Stephanie has her hair cut	it will be shorter	89.60
N	if toast is overcooked	it will be black	88.30
N	if a ruler is used	the line will be straight	85.00
N	if the bananas are allowed to become over-ripe	they will turn brown	80.60
N	if the car runs into a brick wall	it will stop	73.90

I	if water is heated to 100 degrees c	it will be cold	0
I	if a ball is placed at the bottom of a slope	it will roll upwards	0
I	if plastic touches a magnet	it will stick to it	4
I	if the car brake is pressed	the car will go faster	4.25
I	if it is night time	it will be sunny	7.00
I	if she has forgotten her passport	she will be able to take her flight	8.10
I	if the print cartridge is empty	the printer will print	8.35
I	if oil is added to water	they will mix	10.00
I	if the TV is on standby	it will not be using electricity	10.20
I	if she is a vegetarian	she will eat steak	10.70
I	if he is awake	he will have a bad dream	10.90
I	if the door is locked and bolted	it will open	11.50
I	if the lock is empty	the boat will float	17.00
I	if the canoe has a hole in it	it will float	17.00
I	if it is a lemon	it will taste sweet	21.60
I	if the oxygen runs out	the fire will burn	21.90

Appendix 5G continued ......

PS	if it is foggy	visibility will be poor	91.90
PS	if he swims in the Plymouth Sound in the winter	he will feel cold	87.50
PS	if Camilla eats an ice-lolly	her mouth will get cold	86.90
PS	if the weather is hot	people will perspire	86.60
PS	if Simon cuts his finger with a sharp knife	it will bleed	86.30
PS	if there is a power cut	the lights will go out	86.00
PS	if the cat it content	She will purr	84.50
PS	if a baby is hungry	he will cry	82.50
PS	if the plant receives water and sunlight	it will grow	81.10
PS	if a glass is dropped on a stone floor	it will break	79.40
PS	if the aeroplane crashes	people will die	70.30
PS	if the train breaks down	it will be late	69.20
PS	if Sarah peels an onion	her eyes will water	64.50
PS	if he has a cold	he will cough or sneeze	61.40
PS	if there is no water in the vase	The flowers will wilt	59.96
PS	if blotting paper gets wet	it will tear	59.00

PW	if he drinks 6 pints of beer	he will pass a breathalyser test	6.85
PW	if the Queen dies	Beatrice will become Queen	11.30
PW	if Victoria Beckham makes a movie	she will win an Oscar	15.80
PW	if his parachute fails to open	the parachutist will survive	18.80
PW	if she follows a low calorie diet	she will gain weight	21.30
PW	if there is an in class test	all students will get 100%	24.30
PW	if she is poor	she will own a BMW	24.70
PW	if Tim Henman recovers from injury	he will win Wimbledon	31.30
PW	if the lecture is cancelled	the students will be disappointed	31.90
PW	if Oasis release a new CD	sales will be low	31.90
PW	if it is stormy weather	the oil tanker will sink	36.40
PW	if the dog falls into the canal	it will drown	36.50
PW	if she falls down the stairs	she will break her elbow	37.50
PW	if Steve drives through red traffic lights	he will be arrested	47.50
PW	if Beckham plays for England again	he will be captain	51.40
PW	if the aeroplane crashes on take off	some passengers will die	59.90

Appendix 5H
A full set of inferences used for experiment 6

_____

Necessary	if water temperature falls below zero, it will freeze
	if butter is heated, it will melt
	if a person dies, they will stop breathing
	if the bucket is lowered into a dry well, it will come up empty
	if the balloon is pricked, it will burst
	if the dog has fleas, it will scratch itself
	if the Pope dies, a new pope will be elected
	if it rains heavily, the streets will get wet
PS	if a cat is content, it will purr
	if a baby is hungry, he will cry
	if the plant receives water and sunlight, it will grow
	if a glass is dropped on a stone floor, it will break
	if the aeroplane crashes, people will die
	if the train breaks down, it will be late
	if Sarah peels an onion, her eyes will water
	if he has a cold, he will cough or sneeze
Impossible	if water is heated to 100 degrees centigrade, it will be cold
	if a ball is placed at the bottom of a slope, it will roll upwards
	if plastic touches a magnet, it will stick to it
	if the car brake is pressed, the car will go faster
	if it is night time, it will be sunny
	if has forgotten her passport, she will be able to take her flight
	if the print cartridge is empty, the printer will print
	if oil is added to water, they will mix
PW	if there is an in class test, all students will get 100%
	if she is poor, she will own a BMW
	if Tim Henman recovers from injury, he will win Wimbledon
	if the lecture is cancelled, the students will be disappointed
	if Oasis release a new CD, sales will be low
	if it is stormy weather, the oil tanker will sink
	If the dog falls into the canal it will drown
	she falls down the stairs, she will break her elbow

Appendix 5I
Written instruction (necessity) presented to participants, for experiment 6, prior to the inference task

---

<u>Instructions (N)</u>

A number of problems will be presented one at a time. For each problem you will be shown a statement that you must consider to be true. Following this you will be given a second statement, and your task is to indicate whether the conclusion <u>necessarily</u> follows from the sentence that precedes it. A <u>necessary</u> conclusion is one that <u>must be true</u>, given the truth of the preceding sentence. Below are examples of the screen layouts.

> Given that
> > It is a lemon
>
> > Press space bar to continue

A conclusion will then be added to the screen, and your task is to decide whether this conclusion <u>must be true.</u>

> Given that
> > It is a lemon
> *is it necessary that*
> > It will taste sweet
>
> > press either 'yes' or 'no' on the keyboard

You will then be asked to press the space bar when you are ready to continue to the next task. Initially, you will be given two practice problems, but *please ask* the experimenter at any time if you are unsure on how to proceed.

Note: Responses are timed, and it is important that you answer the questions both carefully and accurately

Appendix 5J
Written instruction (possibility) presented to participants, for experiment 6, prior to
the inference task

Instructions (P)

A number of problems will be presented one at a time. For each problem you will be
shown a statement that you must consider to be true. Following this you will be given a
second statement, and your task is to indicate whether the conclusion <u>possibly</u> follows
from the sentence that precedes it. A <u>possible</u> conclusion is one that <u>could be true</u>,
given the truth of the preceding sentence. Below are examples of the screen layouts.

> Given that
> > He cuts his finger
>
> > Press space bar to continue

A conclusion will be added to the screen. Your task is to decide whether this conclusion
<u>could be true</u>.

> Given that
> > He cuts his finger
> *is it possible that*
> > It will bleed
>
> > press either 'yes' or 'no' on the keyboard

You will then be asked to press the space bar when you are ready to continue to the
next task. Initially, you will be given two practice problems, but *please ask* the
experimenter at any time if you are unsure on how to proceed.

Note: Responses are timed, and it is important that you answer the questions both
carefully and accurately.

Appendix 5K
All ANOVA tables for experiment 6

Comparison of mean percentage endorsement rates between Necessary and PS inferences (within subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta^2
Instruction	57041.68	1	57041.68	103.08	.00	.64
Instruction*ability	2.60	1	2.60	.01	.95	.00
Error (instruction)	32096.35	58	57041.67			
Problem	18815.10	1	18815.10	125.88	.00	.69
Problem*ability	93.75	1	93.75		.43	.01
Error (problem)	8669.27	58	149.47	.63		
Instruction*problem	20627.60	1	20627.60	133.39	.00	.70
Instruction*problem*ability	166.67	1	166.67	1.08	.30	.02
Error (instruction*problem)	8971.35	58	154.68			

Comparison of mean percentage endorsement rates between low and high ability groups on Necessary and PS inferences (between subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta^2
Intercept	1666666.67	1	1666666.67	3065.49	.00	.98
Ability	315.10	1	315.10	.58	.45	.01
Error	31533.85	58	543.69			

Appendix 5K continued ......

Comparison of mean percentage endorsement rates between Impossible and PW inferences (within subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta^2
Instruction	112666.67	1	112666.67	352.83	.00	.86
Instruction*ability	375.00	1	375.00	1.17	.28	.02
Error (instruction)	18520.83	58	319.33			
Problem	69190.10	1	69190.10	342.29	.00	.86
Problem*ability	23.48	1	23.48	.12	.74	.00
Error (problem)	11723.96	58	202.14			
Instruction*problem	70041.67	1	70041.67	495.54	.00	.90
Instruction*problem*ability	41.67	1	41.67	.30	.59	.01
Error (instruction*problem)	8197.92	58	141.34			

Comparison of mean percentage endorsement rates between low and high ability groups on Impossible and PW inferences (between subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta^2
Intercept	156315.10	1	156315.10	680.77	.00	.92
Ability	210.94	1	210.94	.92	.34	.02
Error	13317.71	58	229.62			

Appendix 5K continued ......

Comparison of mean evaluation times between Necessary and PS inferences (within subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta^2
Instruction	2.63	1	2.63	25.61	.00	.31
Instruction*ability	2994146.43	1	2994146.43	2.90	.09	.05
Error (instruction)	5.97	58	1029229.77			
Problem	453.41	1	453.41	.00	.97	.00
Problem*ability	243389.74	1	243389.74	.44	.51	.01
Error (problem)	3.23	58	556867.30			
Instruction*problem	1190992.79	1	1180992.79	3.52	.07	.06
Instruction*problem*ability	9186.89	1	9186.89	.03	.87	.00
Error (instruction*problem)	1.95	58	335571.78			

Comparison of mean evaluation times between low and high ability groups on Necessary and PS inferences (between subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta^2
Intercept	2.82	1	2.82	709.38	.00	.92
Ability	3.69	1	3.69	9.28	.00	.19
Error	2.30	58	3970643.61			

Appendix 5K continued ......

Comparison of mean evaluation times between Impossible and PW inferences (within subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta^2
Instruction	837240.47	1	837240.47	.88	.35	.02
Instruction*ability	562166.00	1	562166.00	.59	.45	.01
Error (instruction)	5.53	58	953752.47			
Problem	220493.13	1	220493.13	.23	.63	.00
Problem*ability	107.67	1	107.67	.00	.99	.00
Error (problem)	5.48	58	955951.17			
Instruction* problem	1878589.68	1	1878589.68	2.27	.14	.04
Instruction*problem*ability	379314.38	1	379314.38	.46	.50	.01
Error (instruction*problem)	4.80	58	826958.89			

Comparison of mean evaluation times between low and high ability groups on Impossible and PW inferences (between subjects)

	Sum of Squares	df	Mean Square	F	Sig.	Partial Eta^2
Intercept	3.89	1	3.89	1379.81	.00	.96
Ability	5.92	1	5.92	20.99	.00	.27
Error	1.64	58	2819823.15			