Evans et al. (1993) propose a three-part reasoning strategy consistent with mental logic theory. First, the reasoner uncovers the argument's logical form and represents this in an abstract fashion, possibly using tokenised symbols. Second, the reasoner engages a procedure which searches his or her mental repertoire for an appropriate rule, then applies it in order to construct a mental derivation or proof of a conclusion. Third, the reasoner translates the abstract conclusion back into the content of the given premisses. Advocates of mental logic ascribe non-logical reasoning errors to: sound logical reasoning but from misinterpreted premisses (Henle, 1962), the logical demands of a task exceeding the reasoner's logical competency (O'Brien, 1993), omission of relevant rules or commission of irrelevant rules (Rips, 1994), and physiological limitations of human information processing faculties (Gellatly, 1989).

The question of whether mental logic is an accurate theory of logical competence has been vehemently debated in the cognitive science community. Critics argue that there is no reason to suppose that the semantics of the components belonging to a person's internal medium of thought reliably map onto those of the logical connectives and quantifiers in formal systems, and point to experimental results which suggest that people do not reason via laws analogous to those prescribed in logic textbooks (Byrne, 1989; Cohen, 1981; Evans, 1993b; Johnson-Laird and Byrne, 1991). There are, unquestionably, numerous valid inferences which cannot be represented in standard logic, such as those based on: plausibility, probability, moral obligation, causation and belief (Manktelow and Over, 1990). So in order for any model built from formal logic to qualify as a credible account of human reasoning, it would have to combine the rules of inference from many different kinds of formalism, including standard, deontic and modal logic. Critics also argue that mental logic theory is poorly suited to explaining the apparent effects of problem content on reasoning performance, and point to experimental results consistent with the view that human reasoning processes do not abstract away from meaningful problem content, but are highly dependent upon it (Cheng and Holyoak, 1985; Evans et al., 1993; Garnham and Oakhill, 1994; Wason and Johnson-Laird, 1972).

3.3 Non-logical Errors in Formal Specification

A central tenet of this thesis is that the ways in which people reason about formal specifications are determined, at least in part, by the ways in which formal specifications are written and interpreted. This section points to some of the cognitive issues underlying the formal specification process and describes some relevant empirical questions facing the software specification community. It speculates on a range of conditions under which the users of formal methods may commit "non-logical" errors, that is, adhere to lines of reasoning, or accept conclusions, which deviate from the rules of logic. These discussions subsequently form the basis around which an initial exploratory investigation is formulated.

3.3.1 Writing Software Specifications

It is well known that it is possible for a single program to be written in one of an infinite number of ways and yet still be consistent with respect to its requirements (Sheppard and Ince, 1989).

> "Because computer programming is a highly creative endeavour - no less creative than composing a novel, for example - there is no single cognitive approach. Just as the novelist relies on inventiveness to get a story convincingly across to the reader, the programmer must rely on ingenuity to impart complete and unambiguous instructions to the computer. That is why no two programmers will write the same program in exactly the same way, and why there is so much room for introducing errors" (Jacky, 1989, p.24).

The process of writing a formal specification document is similar to programming in the sense that it forces a developer to make numerous conscious decisions about how best to communicate the desired system functionality. Although many of these decisions are trivial and made spontaneously, others are time consuming and require careful deliberation. Some common decisions that designers face include: choice of the level of abstraction at which to express the desired functionality, how to structure dependencies between modules, choice of identifier names, and choice of appropriate grammatical constructs. When making each of these decisions, it is likely that there will be several alternatives to choose from and that the same functionality may be described in different ways. An assessment of the "most appropriate" choice for a given situation may depend upon different, sometimes even conflicting, criteria. A designer's choice of notational constructs, for example, might be revised in view of the expertise of the specification's intended audience.

The criteria used to assess the validity of a designer's decisions can influence to a large extent the way in which a specification is expressed. It has, until now, been a central running argument of this thesis that a designer's decisions must be geared towards maximising the probability that the correct system requirements are communicated to readers in a form that supports reasoning. This argument appears to incorporate a key criterion against which design decisions might be assessed, namely, their probability of leading to reasoning errors. The validity of this argument will be examined during the forthcoming empirical studies.

This discussion has raised questions which would appear central to any study of the psychological complexity of formal specifications. In particular, do designers exhibit systematic preferences or biases in favour of certain types of notational construct? If so, are these preferences justified, that is, do designers always employ the "most appropriate" ones? If not, how far might this impair a reader's interpretation of a specification or their ability to reason about it correctly? One might, moreover, be inclined to seek an independent and objective means for determining whether a specification is expressed in terms of the most appropriate constructs and, hence, whether it is expressed in a manner which maximises the chances it will be correctly interpreted and reasoned about. One of the main assumptions underlying this thesis is that psychological experimentation holds a key to solving these kinds of empirical question, and that software metrics holds a key to providing an independent system for assessing designers' decisions.

3.3.2 Interpreting Formal Specifications

The interpretation of a formal specification is an inference intensive process, similar to that of interpreting ordinary discourse (Clark, 1977; Johnson-Laird, 1983; Noordman and Vonk, 1992). At its simplest level, readers must draw upon their knowledge of the formalised symbols, these symbols' real world referents, and grammatical laws of the notation. These separate bodies of knowledge must be related in such a way as to derive an understanding of the required system. It is important that this understanding is consistent with the system's requirements because design or implementation decisions are based on developers' interpretations of specifications.

> "Mathematical expressions have the advantage of being precise, because they rely only on a minimum of bases, and do not need any contextual information. All the information that is needed is formulated mathematically, and what is not needed is omitted" (Lightfoot, 1991, p.2).

One of the claims most commonly cited in favour of formal methods is that formal specifications are precise and unambiguous (Bowen, 1988; Liskov and Berzins, 1986; Thomas, 1993). The term "precise" is used here in the sense that a formal statement is open to only one interpretation according to the formal semantics underlying the notation in which it is expressed, and that a reader need only understand and apply these formal semantics in order to interpret it. It is debatable, however, whether this kind of precision can be guaranteed in a psychological sense because one cannot guarantee that the human audience of a formal specification will necessarily abide by the same set of formal (intensional) semantics during its interpretation. The problem of interpreting formal specifications is confounded by the fact that the semantics a designer wants a specification to have is not the same as the specification's mathematical semantics. A variable identifier, for example, often has an extensional referent which may be a physical object in the real world or a theoretical concept, and both the identifier itself and its surrounding context must necessarily cue the recall of relevant real world knowledge from readers' semantic memories in order that the statement may be meaningfully understood; it cannot be interpreted meaningfully by humans solely in terms of the mathematics underlying the grammar of the formal notation.

> "No formalism makes any sense in itself; no formal structure has a meaning unless it is related to an informal environment; to assume that the readers of a formal text have got their informal introduction already beforehand and will find precisely the right interpretation is a silly optimism" (Zemanek, 1979, p.14).

The identifiers in the statement "$\forall b : Book \bullet b \in shelved \vee b \in onloan$" appear to have real world referents, and when used together they suggest a specific real world context. Besides an interpretation of this statement in terms of its mathematical semantics, it is likely that readers will assume that the status of a book could be "reserved" or "lost", according to their real world knowledge of libraries, which is certainly not captured here formally. A formal notation's explicit, intensional semantics might therefore give rise to conflicting (extensional) interpretations by human audiences. It is for this reason that we will not be focussing on the formal semantics of the Z notation, but on the interpretations that people give to it.

One of the questions which the first exploratory study (see Chapter Four) aims to address is whether the real world referents of variable identifiers cue interpretations or reasoning strategies which conflict with the mathematical semantics of Z.

3.3.3 Reasoning About Formal Specifications

During the past three decades psychology has shown people to exhibit systematic tendencies towards error and bias when reasoning about natural language statements containing logical operators such as: "if" (Braine and O'Brien, 1991), "or" (Newstead et al., 1984), "and" (Lakoff, 1971), "not" (Johnson-Laird and Tridgell, 1972), "some" and "all" (Johnson-Laird, 1977). Such findings might warrant the concern of those natural language users who depend upon unimpeded deductive thought, because these operators help to determine the truth conditions of common everyday assertions (Johnson-Laird and Byrne, 1993). But despite obvious symbolic differences, most formal notations contain propositional connectives and predicate quantifiers with equivalent semantic definitions as these same natural language constructs: $\Rightarrow$, $\vee$, $\wedge$, $\neg$, $\exists$ and $\forall$. A question which the software engineering community should ask itself is: Do the same non-logical errors and biases that people exhibit when reasoning about natural language statements also occur when trained users are reasoning about the logically equivalent expressions in formal specifications?

> "It (logic) is justified in abstracting - indeed it is under obligation to do so - from all objects of knowledge and their differences, leaving the understanding nothing to deal with save itself and its form" (Kant, in Smith, 1993, p.18).

The idea that formal logic abstracts away all extraneous details, leaving reasoners free to concentrate purely on the underlying form of arguments, derives from Kantian philosophy. Although the influence of bias, intuition, prior belief and other non-logical lines of thought in everyday human reasoning are acknowledged, Kant theorises that these are seldom influential in formal reasoning which he regards as being guided predominantly by well defined mathematical axioms and deductive rules of inference (in Young, 1992). A similar stance appears to be held by contemporary cognitive theorists. It is claimed that the kinds of fallacy and bias which tend to pervade everyday human reasoning do not arise to the same extent in formal contexts (Perkins, 1989; Perkins et al., 1991). Advocates of this position would appear to have overlooked the possibility that the "form" of the formalisation might itself influence reasoning.

There are two reasons why this strand of Kantian philosophy should now be worthy of reassessment. First, in view of the increasingly critical applications of formal methods (MacKensie, 1992) and their increasing use in the software engineering community (Bowen and Hinchey, 1994) it is becoming more important that developers' logical lines of thought are not usurped by non-logical biases. Second, cognitive science has documented numerous empirical findings, especially during the past three decades, which suggest that the tendency toward non-logical reasoning is related to the degree of thematic content in problem material (Barston, 1986; Dominowski, 1995; Griggs and Cox, 1982; Van Duyne, 1974; Wason and Shapiro, 1971, Wilkins, 1928). The Kantian view thus represents an implicit claim in favour of the formal approach, but one that remains to be empirically tested in explicitly formal contexts.

Galotti (1989) states that formal reasoning is normally applied to those problems where all the relevant premisses and the procedures necessary to derive the correct conclusion are known in advance, whereas informal reasoning is normally applied to those less well defined problems where some of the relevant premisses are not supplied with the problem or the reasoner must determine which premisses are relevant before attempting to derive the correct conclusion. Formal reasoning thus appears well suited to the kinds of problem that occur in mathematics or logic, where correct conclusions can be reached by applying explicitly defined rules of inference to given premisses, and informal reasoning appears well suited to the kinds of problem that occur in everyday life, where reasoning heuristics based on belief or intuition are often sufficient for reaching correct decisions.

> "There are certainly many who see intuition or inspiration as central to the design process, and would argue that attempts to be scientific in design are dangerous because they cause designers to inhibit their intuitions" (Loomes et al., 1994, p.188).

It is argued that informal heuristics, such as intuition and guesswork, are central to the software development process because development decisions are typically based on such undeliberated forms of human judgement (Naur, 1985b). The natural tendency of software developers to reason according to these informal heuristics is perhaps being impaired, however, by an overriding obligation to conform with what is gradually becoming the purely formal dictates of computing science (Myers Jr., 1990; Naur, 1985a). It is argued, on the other hand, that it is fundamentally wrong to teach computing science as a discipline which relies heavily upon natural human intuition because this intuition is often "inadequate and misguided" (Gries, 1990). One of the aims of this research is to test whether intuitive, non-logical reasoning strategies are misapplied to formally defined problems in software engineering contexts. Given the frequency with which informal judgemental heuristics tend to be successfully employed in everyday life (Nisbett and Ross, 1980), it would not be surprising if software developers were sometimes found to use these in contexts which call for purely formal lines of thought. Such a finding would, however, warrant the concern of the software engineering community.

3.4 Summary

This chapter has described the central role of inference in human reasoning and decision making, and the importance of deductive reasoning in the context of formal specification. It has explored the possible relations that may exist between logic and human deductive competence, and described why human reasoning will always be susceptible to error and bias. It has also introduced the cognitive theory of mental logic, which argues that the processes underlying human reasoning are analogous to those underlying formal logic. In light of this theory and the various software engineering claims associated with formal methods, we have speculated several likely areas where software developers might err when using formal methods. These provide the basis for design of the initial exploratory experiment, reported in the next chapter, which aims to identify those variables which have a significant effect on reasoning performance in formalised contexts.

Chapter 4

Refining the Methodology

> "Rather than seeing failure and errors as things that exist, but can be avoided with the right methodology, we can view them as things that the designer brings about, and ask what behaviour causes this. If we understood better why designers make mistakes we might be able to suggest ways they can adjust their behaviour to minimise errors, or contain their impact on the process as a whole" (Loomes et al., 1994, p.186).

We have explored some of the claims made in favour of formal methods and speculated on why their users might err in ways similar to the users of informal methods. Until now our research hypotheses have been loosely supported by speculation and generalisation from cognitive studies of human reasoning in natural language contexts; our arguments have had no real scientific basis. The previous chapter reviewed some contemporary cognitive theories of human reasoning and described the possible relations that may exist between logic and human deductive competence. This knowledge was used to suggest areas where software developers are liable to err when using formal methods. This chapter reports an initial study aimed at exploring some of the cognitive issues involved in using formal specifications, including the kinds of systematic mistake likely to be made by trained users. This study was used as a basis around which the research aims and methodology were refined. The chapter concludes with a description of how the results from the initial study influenced the design of three more specialised studies of human reasoning in formalised contexts.

4.1 An Initial Investigation

The central hypothesis of this research concerns the question of whether the non-logical heuristics and biases that people exhibit when reasoning about logical problems expressed in natural language are liable to transfer over into the formal domain. An exploratory study (Vinter et al., 1996) was conducted to generate preliminary empirical evidence in support or refutation of this hypothesis. Design of the study was motivated, first, by hypotheses stemming from cognitive theories of human reasoning in informal contexts, second, by some of the software engineering community's claims pertaining to formal methods and, third, by the need to identify properties of the Z notation which correlate with human reasoning performance. The tasks in this study centered around five key cognitive processes to the formal specification process: reading, writing, understanding, translating and reasoning.

Twelve computing scientists from the University of Hertfordshire's Division of Computer Science, each with a working knowledge of the Z notation, volunteered to participate in the experiment: six members of staff and six students. Their mean age was 30.42 years (s = 10.13) and their mean length of experience with the Z notation was 2.60 years (s = 2.97). All participants were native English language speakers and were randomly selected. The study had a repeated measures design. Task sheets were computer generated and contained accompanying instructions, as shown in Appendix A. Task sheets were distributed to participants and completed anonymously then mailed back to the experimenter. Participants were asked to provide brief biographical details including: age, occupation, course and length of Z experience.

4.1.1 The Formalised Wason Selection Task

Wason's abstract selection task (Wason, 1966) is a problem requiring hypothesis testing and deductive reasoning based on conditional logic. Participants are shown four cards which have a letter on one side and a number on the other, and a conditional rule of the form *if p then q*. The facing values of the cards show one letter and number that match those in the rule, and one letter and number that differ, as shown in Figure 4.1. These correspond to the p, q, $\neg p$ and $\neg q$ cases for the conditional rule. Participants are asked to select the cards they would need to turn over in order to determine whether the rule is true or false.

> You are shown the following four cards, each of which has a letter on one side and a number on the other.
>
> $\boxed{A}$ $\boxed{4}$ $\boxed{S}$ $\boxed{7}$
>
> Here is a rule: If there is an A on one side of the card then there is a 4 on the other. Which cards would you need to turn over in order to determine whether the rule is true or false?

Figure 4.1: A variation of Wason's abstract selection task

Participants must project the possible consequences of turning over each card in order to arrive at the correct response combination. If participants incorrectly interpret the rule as a biconditional statement (that is, $p \Leftrightarrow q$) then all four cards should be turned over, however, logical deduction indicates that the correct combination is the A (p) and 7 ($\neg q$) cards. It is argued that reasoners must draw modus ponens (MP) and modus tollens (MT) logical inferences, as illustrated in Figure 4.2, in order to see the relevance of the p and $\neg q$ cases respectively (Griggs and Cox, 1982; Manktelow and Evans, 1979; Pollard and Evans, 1981).

$$\begin{array}{l} \textit{if } p \textit{ then } q \\ p \\ \hline q \end{array} \quad \text{Modus ponens} \qquad \begin{array}{l} \textit{if } p \textit{ then } q \\ \neg q \\ \hline \neg p \end{array} \quad \text{Modus tollens}$$

Figure 4.2: Logically valid conditional inferences

Most participants tend to select the p and q cards, or the p card alone, and fail entirely to select the $\neg q$ card. That only 4% of participants selected the correct combination during Wason's early trials (Wason and Johnson-Laird, 1972) suggests that people are frequently prone to error when reasoning about conditional statements and, more specifically, that the difference in selection rates for the two correct cards may be attributable to the relative difficulty in drawing affirming MP and denying MT inferences (Braine, 1978; Evans, 1977a; Pollard and Evans, 1981).

Figure 4.3 shows a variation of Wason's selection task, set within the formal grammar of the Z notation, which was presented to participants as part of the initial study. The logic of the conditional rule is implicit in Wason's natural language based version of the task, but made explicit by the presence of a logical implication operator in the formalised version. Operational "before" and "after" state variables expressing simple mathematical relations are used to correspond to the p, q, $\neg p$ and $\neg q$ cases. As in Wason's original version, the correct response is to select the p and $\neg q$ cases, which in this instance correspond to the input "$in? = A$" and output "$out! = 7$" respectively.[1]

[1] With the benefit of hindsight, the prompt used to obtain participants' responses could have been expressed clearer. Participants were asked "Which inputs and outputs would help you to test whether 'InOut' is working correctly?" If this prompt were interpreted literally then all of the inputs and outputs shown to participants could in fact "help" the testing process by providing feedback in the form of test results. A requirement of Wason's (1966) abstract selection task which is not spelt out explicitly in the instructions for the present task is to identify the minimum number of cases that would enable the participant to determine whether the given conditional rule is true or false. Perhaps a better prompt would therefore have been: "Select the minimum number of inputs and outputs which would enable you to determine whether operation 'InOut' is working correctly". The fact that no participants queried the task's instructions, and only one selected all four possible cases suggests that our task was, nevertheless, interpreted by participants in the manner intended and that its requirements were sufficiently clear from the instructions shown.

The requirements for a software operation 'InOut' are: "If the operation receives an 'A' as input then it will output a '4'." The following Z schema is the operation's formal specification.

$$\begin{array}{|l}
\multicolumn{1}{l}{\mathit{InOut}} \\
\hline
\mathit{in?} : \mathit{Letter} \\
\mathit{out!} : \mathbb{N} \\
\hline
(\mathit{in?} = A) \Rightarrow (\mathit{out!} = 4) \\
\hline
\end{array}$$

(a) $in? = A$ (b) $out! = 4$ (c) $in? = S$ (d) $out! = 7$

Which inputs and outputs would help you to test whether 'InOut' is working correctly? Please circle your choice(s).

Figure 4.3: The formalised selection task

Similarly low rates of correctness have been observed during numerous variations of the selection task (Manktelow and Evans, 1979; Reich and Ruth, 1982; Roberge and Antonak, 1979). Participants have been recruited during these studies, however, without regard for their prior experience in formal logic. It seems intuitively more probable that participants with backgrounds in formal logic would make the types of inference necessary to deduce the correct response combination for the formalised selection task, because their recognition of the logical conditional operator should cue their reasoning processes to endorse only those inferences which follow logically. The results, however, suggest that this is not the case.

Results

TABLE 4.1

Frequencies of response combinations selected during formalised ($N = 12$) and natural language ($N = 128$) versions of Wason's selection task

Study	$p, \neg q$	p, q	p	$p, q, \neg p$	*Others*
Formalised	0	7	3	1	1
Natural Language	5	59	42	9	13

Source: Wason and Johnson-Laird (1972, p. 182).

Although every participant correctly assessed the relevance of the p case, none appeared to see the relevance of the $\neg q$ case. This finding supports the hypothesis concerning the relative difficulties of drawing MP and MT inferences. The 0% success rate observed for the formalised version of the task was even lower than the 4% observed by Wason. There are, however, clear similarities in the patterns of results obtained. In both studies the frequency at which participants selected the p, q combination was highest, selection of the p case alone was second highest, and few or no participants selected the $\neg q$ case. These trends are particularly interesting because they suggest that participants in both studies used similar, language independent, reasoning processes to arrive at their selections. They also suggest that few participants adopted biconditional interpretations of the conditional rule.

The high rates of incorrect p, q selections given in response to Wason's selection task are sometimes attributed to "matching bias" (Evans, 1972b; 1983a; 1983b; Evans and Lynch, 1973). This theory claims that reasoners select, or evaluate as relevant, only those response options which contain one or more of the terms mentioned explicitly in the given rule; namely, p and q. The high rates at which these cases were selected in the present task, despite its explicitly formal context, suggests that participants succumbed to a similar non-logical bias and, hence, that their selections were based mainly on guesswork or intuition rather than logical deduction. This may appear somewhat of a surprising finding given participants' experience of using logic based notations and the clearly logical nature of the task.

4.1.2 The Translation Tasks

Advocates of formal methods often claim that there exists only one way to interpret a formal specification, owing to the precise and well defined nature of the semantics underlying formal notations (Barroca and McDermid, 1992; Bowen, 1988; Liskov and Berzins, 1986; Thomas, 1993). It is also claimed that the use of natural language specifications can give rise to ambiguous interpretations (Ince, 1992; Meyer, 1985; Norcliffe and Slater, 1991). In order to subject these claims to empirical scrutiny, two "translation tasks" were designed to test the ease with which people are able to interpret existing specifications and create new ones, whilst maintaining consistent logical meanings between formal and informal expressions. The tasks aimed to explore translation between natural language and formal grammars in both directions: the first task tested the translation process from Z to English, whilst the second tested it in the opposite direction.

Z to English Translation

For the Z to English translation task, participants were asked to translate part of the formal specification for a computerised library system (Potter et al., 1996) into an appropriate form in natural English. For the purposes of the task, however, the fourth Z predicate was modified to oppose people's conceptions of standard library procedures, as shown below.

> Original predicate: $\forall\, r : \mathit{readers} \bullet \#(\mathit{issued} \rhd \{r\}) \leq \mathit{maxloans}$
> The number of books that any reader borrows must be less than or equal to the maximum number of loans allowed.
>
> Revised predicate: $\neg\, \exists\, r : \mathit{readers} \bullet \neg(\#(\mathit{issued} \rhd \{r\}) > \mathit{maxloans})$
> The number of books that any reader borrows must be more than the maximum number of loans allowed.

Participants were asked to show their understandings of the specification shown in Figure 4.4 by expressing its predicates in natural language. Although it was expected that most would succeed in giving logically consistent English translations of the first three predicates, it was hoped that participants' efforts in translating the more complex fourth predicate would throw some light on the cognitive processes involved in interpreting and reasoning with formal specifications.

$$
\begin{array}{|l}
\multicolumn{1}{l}{\mathit{Library}} \\
\hline
\mathit{stock} : \mathit{Copy} \nrightarrow \mathit{Book} \\
\mathit{issued} : \mathit{Copy} \nrightarrow \mathit{Reader} \\
\mathit{shelved} : \mathbb{F}\,\mathit{Copy} \\
\mathit{readers} : \mathbb{F}\,\mathit{Reader} \\
\hline
\mathit{shelved} \cup \operatorname{dom} \mathit{issued} = \operatorname{dom} \mathit{stock} \\
\mathit{shelved} \cap \operatorname{dom} \mathit{issued} = \varnothing \\
\operatorname{ran} \mathit{issued} \subseteq \mathit{readers} \\
\neg\,\exists\, r : \mathit{readers} \bullet \neg(\#(\mathit{issued} \rhd \{r\}) > \mathit{maxloans}) \\
\hline
\end{array}
$$

Figure 4.4: The modified library specification

Natural language based studies of syllogistic reasoning suggest that people implicitly convert given premiss information, during its interpretation, into intuitive forms which are more amenable to mental representation or reasoning (Newstead, 1990; Revlis, 1975a). It was possible for participants to simplify the translation process in a similar manner for the present task by implicitly converting the fourth predicate into one of several logically equivalent forms, as shown below.

$$
\begin{array}{ll}
\neg\,\exists\, r : \mathit{readers} \bullet \neg(\#(\mathit{issued} \rhd \{r\}) > \mathit{maxloans}) & \equiv \\
\neg\,\exists\, r : \mathit{readers} \bullet \#(\mathit{issued} \rhd \{r\}) \leq \mathit{maxloans} & \equiv \\
\forall\, r : \mathit{readers} \bullet \neg(\#(\mathit{issued} \rhd \{r\}) \leq \mathit{maxloans}) & \equiv \\
\forall\, r : \mathit{readers} \bullet \#(\mathit{issued} \rhd \{r\}) > \mathit{maxloans} &
\end{array}
$$

Results

TABLE 4.2

Frequencies of valid Z to English translations ($N = 12$)

Predicate 1	*Predicate 2*	*Predicate 3*	*Predicate 4*
8	9	8	0

The results of the Z to English translation task suggest that most participants gave logically consistent English translations of the first three predicates, although the fact that at least one quarter erred in each case may be a cause for some concern. No participants' translations of the fourth predicate preserved the meaning of the given Z expression. This finding has two important implications. First, it suggests that important syntactic and semantic properties of formal specifications can be lost during their interpretation; a specification may have one formal semantics but is liable to be interpreted in many ways by its human audience. Second, the forms of participants' responses suggest that none attempted to simplify the original complex expression to a more intuitive form in order to ease their interpretations of the text. Instead, all appear to have relied upon guesswork based on associations implied by their prior knowledge of similar linguistic contexts in order to arrive at a plausible, but incorrect, understanding. As a possible explanation for these errors, it is hypothesised that all participants obtained the gist of the fourth predicate's meaning by relating its key linguistic components - namely, the variable identifiers - to their own preconceptions of real world library systems, regardless of the formal text. This hypothesis is supported by the fact that all responses were consistent with the form "No reader may borrow more books than the maximum number of loans allowed". Clearly here twelve readers interpreted a formal specification in a manner different to that of its author.

It is widely accepted that logic is truth-functional, and that "truth-functional statements are so called because their truth is determined entirely, and only, by the truth or falsity of their constituent statements" (Strawson, 1966, p.66). Our results suggest that the truth of statements written explicitly in symbolic form are sometimes interpreted according to people's personal beliefs towards the terms' referent concepts in the real world, rather than the intensional semantics of the logical terms. This may be attributable to the fact that formal grammars "touch ordinary usage at some vital points" (Strawson, 1966, p.58), and that non-logical heuristics can be cued at those points where logic and ordinary language do "touch".

This, as we have seen, appears to cause a divergence from purely truth-functional lines of thought. This finding suggests that the users of formal methods are liable to misinterpret even small scale specifications and opposes a commonly held belief in the software engineering community; namely, "there is only one way to interpret a formal specification, because of the well defined and unambiguous semantics of the specification language" (Liskov and Berzins, 1986, p.5).

English to Z Translation

The main aim of the second translation task was to test the common engineering claim that natural language specifications are prone to ambiguity (Gehani, 1986; Imperato, 1991). Participants were shown an English requirements description, "Operation *ComputeValue* outputs the sum of its two inputs squared", and asked to translate it into an appropriate form in Z. This description is problematic from a developer's perspective because it does not suggest a single, well defined computational algorithm. Specifically, it does not state whether the two inputs are to be squared before or after their addition. Responses resembling either of the two forms shown in Figure 4.5 could therefore satisfy the operation's requirements, despite the fact that they nearly always give different solutions for the same inputs. Although most participants were expected to recognise that more than one algorithm was possible, it was predicted that they would resolve this dilemma with recourse to knowledge of elementary mathematical principles; the rules of arithmetic precedence state that multiplication takes priority over addition wherever there is an absence of parentheses. It was therefore expected that the order of computational precedence in most responses would resemble the form of *ComputeValue*(a).

$$\begin{array}{|l}
\multicolumn{1}{l}{\textit{ComputeValue}(a)} \\
in1?, in2? : \mathbb{Z} \\
out! : \mathbb{Z} \\
\hline
out! = (in1? \times in1?) + \\
\qquad (in2? \times in2?) \\
\hline
\end{array}
\qquad
\begin{array}{|l}
\multicolumn{1}{l}{\textit{ComputeValue}(b)} \\
in1?, in2? : \mathbb{Z} \\
out! : \mathbb{Z} \\
\hline
out! = (in1? + in2?) \times \\
\qquad (in1? + in2?) \\
\hline
\end{array}$$

Figure 4.5: Consistent implementations of "ComputeValue"

Results

TABLE 4.3
English to Z translation methods ($N = 12$)

$(a \times a) + (b \times b)$	$(a + b) \times (a + b)$	*Other*
6	6	0

The results for the English to Z translation task suggest that there was an equally balanced difference of opinion regarding which method of computation is more appropriate. That each of the two predicted forms were selected by exactly half of participants, and only one participant sought clarification of the task requirements, supports the claim that natural language specifications are prone to ambiguity. Given that most participants did not appear to use arithmetic rules of precedence, as predicted, the results also suggest that designers must be careful about which aspects of their audience's prior knowledge are taken for granted.

Aside from the methods of computation prescribed in their responses, participants' varied use of the Z notation highlighted several further issues of relevance. It is claimed that languages with restricted grammars are less likely to constrain their users towards specific implementations (Bowen and Hinchey, 1995). Based on this assumption, one would expect users to be more constrained by their use of a natural rather than a formal language. The results lend support to this claim insofar as participants' use of Z resulted in a wide range of consistent solutions, despite the seemingly limited scope of the given requirements. In fact, no two responses were exactly the same. This illustrates an important, but often overlooked, issue; much can be implied by a set of requirements without being explicitly stated within it. The process of formalisation helps to explicate hidden functionality, which is then expressed by designers according to their own discretions and personalised styles of writing. Based on the responses for this task, most participants appeared to make conscious decisions involving the following issues: the valid and invalid use of Z, the choice of meaningful identifier names, the choice of data types assigned to variables, the use of parentheses to clarify operator precedence, the ordering of expressions and the use of extra variables for storing intermediate results.

Judging by the lack of prior research aimed at investigating the human aspects of using formal methods, it would appear that individual psychological differences exert little influence during the production of formal specifications. Yet participants' responses suggest that, in practice, rarely would any two designers arrive at exactly the same specification, even if this were based on the same set of requirements. The English to Z translation task shows that the process of writing a formal specification is far from being a completely automated or systematic exercise and that, despite containing much more restricted vocabularies than natural languages, formal notations are powerful enough to allow designers to exercise their own discretion, creativity and freedom of expression. Perhaps more importantly, this task suggests that the production of a formal specification is frequently guided by subjective human judgement, implicit linguistic conventions and undefined development procedures. It is therefore frequently prone to human error.

4.1.3 The Style Preference Task

The following claim, based on an informal "straw poll" of software engineers, suggests that audiences are more likely to understand clear (i.e. precise) specifications rather than brief (i.e. concise) specifications.

> "A succinct formulation of a certain property may seem adequate to one; while another will prefer a more verbose exposition of its consequences. To communicate clearly with the majority of readers, you should, in general, prefer clarity to brevity" (Gravell, 1991, p.139).

This seems like a reasonable assertion at first impression given that concise specifications tend to be presented at a high level of abstraction and the introduction of extra precision allows possible implementations to be more easily envisaged. An experimental task was designed to discover participants' linguistic style preferences and, at the same time, to subject Gravell's claim to empirical test. Participants were presented with an English description of an imaginary software operation ("Operation *Toggle* exchanges the current status of a switch"), a Z data type definition ("$SWITCH ::= on \mid off$") and four alternative Z specifications of the same operation as shown in Figure 4.6: (1) concise, (2) verbose, (3) precise and (4) imprecise (that is, wrong). Participants were asked to judge which implementation "best describes the operation's behaviour" and to justify their choices appropriately.

$$\begin{array}{|l}
\multicolumn{1}{l}{\textit{Toggle\_1}} \\ \hline
s, s' : SWITCH \\ \hline
s' \neq s \\ \hline
\end{array}
\qquad
\begin{array}{|l}
\multicolumn{1}{l}{\textit{Toggle\_2}} \\ \hline
s, s' : SWITCH \\ \hline
(s = off \land s' = on) \lor \\
(s = on \land s' = off) \\ \hline
\end{array}$$

$$\begin{array}{|l}
\multicolumn{1}{l}{\textit{Toggle\_3}} \\ \hline
s, s' : SWITCH \\ \hline
s = on \Rightarrow s' = off \\
s = off \Rightarrow s' = on \\ \hline
\end{array}
\qquad
\begin{array}{|l}
\multicolumn{1}{l}{\textit{Toggle\_4}} \\ \hline
s, s' : SWITCH \\ \hline
(s = on \lor s = off) \Rightarrow \\
(s' = on \lor s' = off) \\ \hline
\end{array}$$

Figure 4.6: Four ways of specifying *Toggle* in Z

Results

TABLE 4.4
Frequencies of style preferences ($N = 12$)

Concise	*Verbose*	*Precise*	*Imprecise*
4	4	4	0

Participants' responses suggest that preferences were equally divided amongst three of the four specifications, with one third selecting each of the concise, verbose and precise styles. Quite reasonably, none favoured the imprecise style. These results appear to contradict Gravell's claim that most readers find precise specifications easiest to understand. They suggest that, whilst precision might be highly desirable, it is not necessarily the most important factor. Upon closer inspection the results suggest strong links between participants' ages, levels of experience and their style preferences. It is noteworthy that the youngest and least experienced tended to prefer concision, whilst the eldest and most experienced tended to prefer precision. The way in which the task's prompt, "best describes", was interpreted could therefore have varied between participants. If this was the case then we would expect to see marked variation in the forms of justification offered by participants in support of their choices. Casual inspection reveals reference to a wide range of factors including: clarity, explicitness, intuitiveness, conciseness and the type of application being specified. This suggests that the criteria used to discriminate between the four styles did indeed vary and that the notion of what constitutes the "best" writing style for a formal specification is determined, at least in part, by the personal preferences of its audience.

It might be argued that the ideal level of abstraction at which one could write a specification would take into account its audience's prior knowledge and language expertise. In practice, however, a specification is normally written for different audiences with very different backgrounds: designers, programmers, managers, customers, quality controllers and technical authors (Wing, 1990). Given that the backgrounds of a specification's audience are often unknown in advance and it is normally impractical to produce a separate version for each group with a certain level of expertise, this might explain why precision is rarely compromised in practice and designers tend to specify as much explicit detail as possible. Whether this principle should be applied in all situations is debatable.

> "In spoken situation, the audience is given, but in writing, an audience must be imagined and is to some extent chosen. To write technically is to chose a learned audience" (Turner, 1986, p.190).

Considerate designers writing for novice language users might aim to specify the maximum detail clearly, using only the simplest of a notation's constructs, so as to leave nothing open to misinterpretation. This might enable all of an audience to comprehend the writer's intended meaning, without relying upon their knowledge of the notation's more obscure features. This has the unfortunate consequence that expert readers may find the document laborious to read. Designers writing for learned audiences, in contrast, might aim to specify the minimum detail necessary by freely using the full range of a notation's constructs, leaving readers to infer for themselves the other, implicit, properties of system functionality. In this case a writer would rely entirely upon the audience's expert knowledge of the notation and might presuppose their awareness of relevant information not explicitly stated in the specification. In this case novice readers may be unable to comprehend certain parts of the specification and might accept the first plausible meaning that appeals to their intuitions, as exemplified by participants' responses during the library translation task. This has the unfortunate consequence that readers may use their inaccurate interpretations as a basis for making incorrect development decisions.

4.1.4 Conditional Inference Tasks

Evans (1977a) sought to test whether the form of a logical rule, or the presence and absence of negative components, would affect the ability of reasoners to draw inferences about conditional statements expressed in natural language. The results suggest that conditional reasoning performance can be altered significantly by manipulating these two linguistic variables. Five "conditional inference tasks" expressed in the Z notation were designed to test whether formalisation would affect the rates at which reasoners drew the same valid inferences or succumbed to the same classical fallacies as those observed in Evans' natural language based study.

Premiss pairs containing one conditional and one equivalence were presented in the form of Z predicates. Participants were asked to say what followed from these premisses by selecting one from four possible response options. Three forms of valid inference needed to be drawn in order to deduce the correct conclusions: MP, MT and MT-N (modus tollens with negated antecedent). Each of these inferences have logically determinate conclusions. The tasks also required participants to avoid committing two forms of logical fallacy: Denying the Antecedent (DA) and Affirming the Consequent (AC). As no logically determinate conclusions can be deduced from DA and AC premisses, participants were required to indicate that nothing logically followed in these cases. Figure 4.7 shows the logical forms of these tasks.

Type	1^{st} *premiss*	2^{nd} *premiss*	*Conclusion*
MP:	$(shape = circle) \Rightarrow (colour = blue)$	$shape = circle$	$\therefore\ colour = blue$
MT:	$(shape = circle) \Rightarrow (colour = blue)$	$colour = red$	$\therefore\ shape \neq circle$
MT-N:	$\neg(shape = circle) \Rightarrow (colour = blue)$	$colour \neq blue$	$\therefore\ shape = circle$
DA:	$(shape = triangle) \Rightarrow (colour = red)$	$shape = square$	$\therefore\ colour \neq red^*$
AC:	$(shape = square) \Rightarrow (colour = green)$	$colour = green$	$\therefore\ shape = square^*$

Note: Conclusions marked with an asterisk are logical fallacies.

Figure 4.7: Conditional inference tasks

Results

TABLE 4.5
Frequencies of conditional inferences

Study	*n*	*MP*	*MT*	*MT-N*	*DA*	*AC*
Present study	12	12	4	4	1	2
Evans (1977a)	16	16	12	2	11	12

The results appear to throw some light on the difficulties that people experience when attempting to draw inferences, and their proneness to fallacies, when reasoning about formal conditional logic. Table 4.5 compares the results obtained from the present formalised study and Evans' (1977a) natural language based study. That perfect rates of correctness were observed for the MP inference in both studies suggests that people are highly adept at drawing MP inferences, independent of the linguistic context in which they are presented. The fact that a much lower rate of correct MT inferences was observed in the present study, however, suggests that formalisation can, under some circumstances at least, lead to a degradation in reasoning performance. That the rate of correctness observed for MT was much lower than that observed for MP suggests that the two forms may have different levels of psychological complexity. It might also begin to explain the same participants' poor performance on the formalised Wason selection task where it is necessary to make an MT inference in order to evaluate the $\neg q$ case as being relevant.

In Evans' study the presence of a negative operator in the antecedent of a conditional rule appears to have a detrimental effect on participants' ability to draw MT inferences. The equal rates of success observed for MT and MT-N in the present study suggests that participants were not distracted by the introduction of a negative operator. This conclusion cannot be drawn conclusively from the results, however, in view of the extremely low levels of performance observed for these inference types. The fact that a much lower rate responded correctly to these inferences than in Evans' study, nevertheless, suggests that greater difficulty is experienced when drawing MT inferences in formalised contexts. It is noteworthy that the rates at which participants committed the DA and AC fallacies were much lower in the present study because this suggests that people may be less susceptible to these logical fallacies in formalised contexts.

4.2 Outcomes of the Initial Investigation

The initial study was helpful both in illuminating some of the main cognitive processes underlying the formal specification process and in identifying grammatical properties of the Z notation that may correlate with human reasoning performance. Perhaps the most noteworthy findings to come from this study are: first, that the users of formal methods are susceptible to many of the same non-logical conditional reasoning heuristics which have been exhibited during equivalent natural language based formulations of logically equivalent tasks, second, that reasoners are liable to favour guesswork or pragmatic heuristics in contexts which clearly call for logical thought, third, that significant properties of formal specifications can be lost or illicitly converted during their interpretation, fourth, that the formalisation of informal requirements relies heavily upon human judgement and ingenuity and, fifth, that the linguistic style preferences of a specification's audience are liable to vary according to their levels of experience. These findings point to the fallibility of human reasoning in formalised contexts and appear to contradict several popular software engineering claims made in favour of formal methods.

The nature of the errors committed during the initial study would have important consequences for the software development process if they were replicated in the community at large, or in more comprehensive experiments. The remainder of this chapter explores the implications of the results for the cognitive and computing communities, and discusses how the initial study helped to refine the research methodology by focussing on specific lines of inquiry. There are two aspects with regard to the latter discussion: first, the study was used to select a small subset of the range of topics which seemed likely to yield robust results and, second, lessons were learnt which influenced the design and conduct of the main experiments.

4.2.1 Implications for Cognitive Science

Taken together with results from natural language based studies, the present study suggests that many of the erroneous logical inferences which people make about implicit logic in natural language are also liable to occur when trained users are reasoning about explicit logic in formal specifications. Despite their experience in using logic based notations and the explicitly logical nature of the tasks set before them, the users of formal methods appear to stray from rudimentary rules of logic when reasoning about formal specifications, even when the appropriate inference rules are well known to them. Although tentative, this finding should be of particular interest to the cognitive science community because it suggests that control of an individual's reasoning is liable to be usurped by higher order, language independent heuristics which lead human judgement away from what is strictly valid under the truth-functional dictates of formal logic. Such results lend credence to the Kantian claim that logic is concerned with objective ideals; with how people ought to think, rather than how they actually do think (Kant, in Smith, 1993).

> "In logic, however, the question is not one of contingent but of necessary rules, not how we think but how we ought to think. The rules of logic, therefore, must be taken not from the contingent but from the necessary use of the understanding, which one finds, without any psychology, in oneself" (Kant, in Young, 1992, p.592).

The results appear to contradict the claims of traditional mental logic theory, which argues that the semantics of the logical components belonging to a person's internal medium of thought map reliably onto those of the logical operators found in formal logic (Braine, 1978; Inhelder and Piaget, 1958; Osherson, 1975; Rips, 1983). Mental logic theory might argue that participants would have identified the $\neg q$ case as being relevant in the formalised selection task if they had already committed the MT rule to their mental repertoires of inference rules. Results from the five conditional inference tasks, however, suggest that at least one third of the same participants knew how to perform both the MT and MT-N forms of inference. The question of why the same number did not draw the MT inference necessary for the formalised selection task might be answered by the possibility that, in practice, people's deductive performance often does not reflect their deductive competence and that reasoning can be impaired or facilitated merely by changing the way in which a problem is presented.

4.2.2 Implications for Software Engineering

The results of the initial study are important from a computing perspective because they suggest that several of the software engineering community's widely held beliefs about formal methods might, in fact, be misconceptions. Although the study did not set out to test whether it is easier to reason about specifications expressed in formal logic than natural language, its results do suggest that, under certain conditions, human reasoning is just as error prone. The findings from the conditional inference and formalised selection tasks, for example, suggest that trained computing scientists do not necessarily find it easier to reason with explicit conditionals in formal logic than untrained laymen with implicit conditionals in natural language. This may be attributable to the use of similar, non-logical reasoning processes which lead to people's downfalls in formal and informal versions of the same tasks.

> "There is no better means of communication than good, spoken, natural language. Whoever engages in formalization and formal languages should never let this basic truth disappear in his mind" (Zemanek, 1979, p.14)

The results demonstrate that the precision of a formal specification cannot be guaranteed in an absolute sense. Although a formal specification might be precise and unambiguous in the sense that a notation can be assigned a formal semantics, certifiably error-free communication based on this principle presupposes that readers abide by the same semantics during its comprehension. This is clearly not the case for human audiences whose faculties for language comprehension and reasoning are invariably guided by informal processes, such as pragmatics (Levinson, 1983) and intuitive heuristics (Kahneman et al., 1991), which can lead to interpretations or judgements that conflict with those prescribed by a notation's formal semantics.

> "People place a premium on being rational and cognitively consistent, and so they are reluctant to simply disregard pertinent evidence in order to see what they expect to see and believe what they expect to believe. Instead, people subtly and carefully 'massage' the evidence to make it consistent with their expectations" (Gilovich, 1991, p.53).

The results of the Z to English translation task demonstrate that one particular time at which a reader's faculty for logical reasoning is likely to give way to heuristic processes or associative guesswork is when meaningful information is encountered in formal specifications, especially where this information runs counter to prior belief or expectation. The fact that all of the observed responses were consistent with what participants considered to be the designer's intended meaning, as opposed to the specification's formal meaning, implies that they had adopted pragmatic interpretations. This result suggests that, under such circumstances, trained developers are prone to postulate possible meanings and accept the most plausible based on the surrounding context and their prior beliefs about the type of system being specified. So despite the best efforts of designers, even seemingly "precise and unambiguous" (Bowen, 1988, p.164) specifications are liable to be interpreted in different ways by different users.

4.3 Designing the Main Experiments

A comprehensive investigation of the non-logical tendencies exhibited by users of formal methods might take into account every stage where specifications influence system development. This would include an analysis of the cognitive processes employed by designers when creating formal specifications from first principles, and those employed by developers when interpreting and reasoning about existing specifications. Although such an investigation might touch upon some of the most intriguing intellectual factors underlying software development, it is not a practical proposition for a research project of this scale to pursue all of these avenues of inquiry. The remainder of this chapter describes which possible lines of inquiry were rejected and those selected for further consideration.

Although the initial study appeared to identify several kinds of conscious decision made by designers during the production of formal specifications, it also exposed the difficulties of designing and controlling experiments involving the production of written texts. Whilst it is relatively easy to exercise experimental control over material to be comprehended, it is usually much more difficult to constrain an individual's production of language (Eysenck and Keane, 1990). This was evident in the wide variety of responses offered for the English to Z translation task, no two of which were exactly the same. Given the lack of prior empirical groundwork in the area of formal language production and the likelihood that any theories relating to this idea would be difficult to prove even under "laboratory" conditions, the decision was taken not to pursue this line of inquiry in the main experiments. This remains, therefore, an area in need of further empirical investigation.

Given the forms of error and bias exhibited by participants during the initial study, the scope of the main experiments was refined around testing the cognitive processes involved in interpreting and reasoning about formal specifications. These cognitive processes are, after all, central to many of the psychological claims associated with formal methods and are well documented in the cognitive literature. The results of the initial study suggest, in particular, that the users of formal methods are susceptible to error when reasoning about formal conditional rules. These results are supported by cognitive studies of conditional reasoning in natural language which report numerous forms of error and bias, including those observed during the initial study (Braine and O'Brien, 1991; Evans, 1983a; Evans et al., 1995; Wason, 1966). The non-negligible differences in performance observed for the formalised conditional inferences and Evans' (1977a) equivalent natural language forms, in particular, prompted the decision to investigate the extent to which people commit these logical errors under a wider range of formalised linguistic conditions.

Our review of the cognitive literature in the previous chapter suggests that people are susceptible to systematic errors and biases when reasoning about natural language statements containing: conditionals, disjunctives, conjunctives, negatives and quantifiers. The results of the initial study suggest that non-logical reasoning heuristics are liable to transfer over to the formal domain for conditional inferences. The possibility that non-logical heuristics might also do so for inferences involving the remaining constructs was considered a genuine reason for concern. Given the research aim to discover grammatical properties of the Z notation which correlate with reasoning performance it was decided, therefore, to design the project's main empirical studies around these grammatical constructs which have been shown to elicit reasoning errors and biases in alternative linguistic domains.

The influence of term polarity (i.e. p or $\neg p$) on reasoning performance is a well documented phenomenon in the cognitive literature, where it is argued that the presence or absence of negative terms in task information can cause reasoners to suppress valid inferences and succumb to logical fallacies (Evans, 1972a; 1977a; Johnson-Laird and Tridgell, 1972; Roberge, 1974; 1976b; Wales and Grieve, 1969; Wason, 1959). Given the degree of support for these findings in the natural language domain and in light of our participants' apparent difficulties in drawing the modus tollens based inferences, where the negation of a conditional rule's consequent results in the negation of its antecedent, term polarity was made a focus of concern in all of the main experiments.

Although some of the material presented in the library system translation task was misleading, the fact that no participants gave logically consistent responses suggests that the presence of meaningful identifiers can exert a marked influence on the abilities of individuals to reason about formal specifications. The lessons learnt from this task influenced the design of the main experiments in two ways. First, it prompted the design of tasks containing abstract material (i.e. letters, colours or shapes) and thematic material (i.e. meaningful terms with real world referents). This design decision is supported by findings from cognitive studies which suggest that the degree of meaningful problem content can significantly influence logical reasoning performance (Dominowski, 1995; Griggs and Cox, 1982; Johnson-Laird and Wason, 1970; Van Duyne, 1974; Wason and Shapiro, 1971; Wilkins, 1928). Second, the strength of the result prompted the decision to investigate the effects of intuitive and counter intuitive information on reasoning performance, which was tested during the study of quantified reasoning. This design decision is also well supported by the cognitive literature, which suggests that inferences can be suppressed or facilitated according to the believability of conclusions to be inferred (Evans et al., 1983; Morgan and Morton, 1944; Revlin and Leirer, 1980).

4.4 Summary

In this chapter we have described how the research methodology was refined via an exploratory investigation into several of the cognitive processes deemed central to the formal specification process. The results of this investigation suggest that: reasoning can be affected markedly by changes to the degree of meaningful content or the polarities of logical terms, significant properties of formal specifications can be lost during their interpretation, the precision of a formal specification cannot be guaranteed in an absolute sense, and that the users of formal methods are liable to employ non-logical heuristics and biases in reasoning about formal specifications. The study also helped to identify specific linguistic properties of the Z notation which appear to correlate with human reasoning performance, thereby paving the way for the design of more specialised studies.

Chapter 5

Conditional Reasoning

> "*If* is a two-letter word that has fascinated philosophers for centuries and has stimulated equal interest in the more recently developed disciplines of linguistics and cognitive psychology. The conditional construction *if* ... *then* seems to epitomise the very essence of reasoning. The use of the conditional *if* requires the listener to make suppositions, to entertain hypotheses or to consider once or future possible worlds. If some particular condition was, is, could be or might one day be met, then some particular consequence is deemed to follow" (Evans et al., 1993, p.29).

One of the main implications to arise from our initial study was the possibility that the users of formal methods are prone to systematic forms of error and bias when reasoning about conditional statements expressed in the Z notation. Moreover, the ways in which participants were observed to err is supported by cognitive studies of conditional reasoning in natural language contexts. The initial study concentrated on four classical types of conditional inference: modus ponens (MP), modus tollens (MT), denial of the antecedent (DA) and affirmation of the consequent (AC). These inferences were not tested exhaustively, however, in the sense that the polarities of the antecedent and consequent terms were systematically varied. Furthermore, all of the initial study's conditional inference tasks were presented in the guise of abstract colour and shape scenarios, which were unlikely to have elicited strong connotations with participants' prior beliefs. It might also be argued that the use of a relatively low sample size renders the data suspect. This chapter reports a much more comprehensive study of formalised conditional reasoning designed to identify specific combinations of linguistic properties that are liable to cause human error and bias when trained users are reasoning about Z specifications.

5.1 Error and Bias in Conditional Reasoning

Psychological studies of the conditional rule are perhaps more likely to provide pointers to the more complex cognitive processes that people undergo in deductive reasoning because, unlike the other propositional connectives, a conditional introduces the concepts of hypothesis and supposition (Braine and O'Brien, 1991; Evans et al., 1995). That is, successful interpretation of a conditional rule requires the presupposition of its antecedent as the necessary precondition for the truth of its consequent in some hypothetical world. Cognitive studies suggest people are prone to numerous forms of error and bias when reasoning about conditional statements (see for example: Braine and O'Brien, 1991; Evans, 1977a; O'Brien and Overton, 1982; Taplin and Staudenmayer, 1973), however, all such studies have been conducted within the confines of natural language based contexts alone. Such studies have typically scrutinised the ways in which people reason with abstract sentences, such as "If the letter is a vowel then the number is even", or sentences containing more realistic content, such as "If the man is drinking beer then he must be over 18 years of age". The findings suggest that there are dominant causes for peoples' departure from logical rules of reasoning, and that erroneous responses to conditional tasks are attributable to the influence of non-logical biases and heuristics (Braine and O'Brien, 1991; Evans, 1993a; Pollard and Evans, 1980). The purpose of the first main experiment was to test whether or not the same non-logical tendencies are also liable to occur when the trained users of formal methods are reasoning about logically equivalent conditional statements in formal specifications.

Matching Bias

The theory of "matching bias" claims that reasoners are liable to select, or evaluate as relevant, only those conclusions which contain one or more of the terms mentioned explicitly in given premisses (Evans, 1972b; 1983a; 1983b; Evans and Lynch, 1973). The conditional statements "If A then not 4" and "If not A then 4", for example, both appear to concern the same topic; the letter "A" and the number "4". The theory predicts that, when a response option fails to contain one or both of these terms, reasoners will tend to judge that option as irrelevant, regardless of its actual logical validity. It is suggested that reasoners adopt the matching heuristic only as a last resort, when they do not see which logical rules will lead to a definitively correct solution or they fail to see how they can be applied to the task in hand (Manktelow and Evans, 1979). Linguistic factors play a major part in determining when matching bias occurs because they direct attention to relevant or irrelevant problem information. The high rates at which responses tend to coincide with the predictions of the theory, however, suggest that matching may be part of a higher level reasoning process which is exercised whenever reasoners are unwilling to expend the mental effort necessary for the full logical analysis of a task.

Polarity Effects

It is argued that people exhibit a general implicit bias towards positive information which derives from that convention of everyday discourse stating that the use of a negative presupposes the reason to believe a positive (Wason, 1959). Negatives are normally used to deny prior positives but positives are rarely used to deny prior negatives (Evans, 1972b; 1972c; 1983a; 1983b). For example, "not p" is used to deny "p", but "p" is rarely used to deny "not p", so in both cases attention is directed towards the positive "p". According to linguistic convention, the topic of a positive sentence is the positive itself, but the topic of a negative sentence is the positive which is denied. As people learn to use such conventions to great effect in everyday discourse, it should hardly be surprising that they appear so reluctant to violate them under experimental conditions. The tendency for people to perceive "not p" as the converse of "p" without perceiving "p" as the converse of "not p" in everyday experience should therefore be quite understandable. There is an almost universal tendency in everyday discourse not to recognise a double negative as an affirmative. Given a negated description of an object, "not blue", it can sometimes be difficult for an individual to see how a further negation "not not blue" could result in the object becoming any less blue than it already is. Similarly, the emphatic tone in which statements such as "You don't know nothing" are articulated in everyday communication seem almost to encourage a negative interpretation of the form "You know nothing" (Cohen, 1971). This frequent disinclination in everyday discourse to convert a doubly negated proposition into an affirmative might also account for many of the errors observed under strictly logical experimental conditions.

The influence of component polarity on reasoning performance is a well documented phenomenon in the cognitive literature. Evans (1993a) points to the existence of two seemingly unconscious biases with respect to conditional reasoning.

First, the theory of "negative conclusion bias" claims that people are more inclined to endorse inferences whose conclusions are negative rather than affirmative. The errors observed by Pollard and Evans (1980) are consistent with this theory, and are attributed to participants' misapplication of an everyday heuristic which maximises the chances of making statements that are unlikely to be disproved. Affirmative conclusions normally have particular referents, whereas negative conclusions have multiple referents, so cautious reasoners are liable to favour statements that make non-specific negative predictions over specific affirmative predictions, which are more likely to be refuted. The experimenters claim that the effects of this bias may be lessened by the presence of familiar problem content which directs reasoning towards specific affirmative conclusions. Second, Evans' (1993a) theory of "affirmative premiss bias" claims that individuals are more inclined to endorse determinate conclusions from premisses that do not contain any negative components. That there is little support for this theory in the literature might be due to the generality of its predictions and the fact that determinate responses to affirmative premisses can usually be explained in terms of more specific causes.

Content Effects and Belief Bias

The theory of "facilitation by realism" (Gilhooly and Falconer, 1974) argues that realistic, as opposed to symbolic, task content can have a strong facilitatory effect on human reasoning. But whether this "effect" is in fact genuine and the extent to which it seemingly improves reasoning performance has been the subject of much contention. The debate has yet to be resolved mainly because of the difficulties involved in distinguishing between those occasions when meaningful content aids the process of reasoning and those occasions when it simply cues the direct recall of a response from memory with little or no reasoning having taken place. It is a well supported finding that conclusions conforming with prior convictions are more likely to be endorsed than those running counter, although such inferences are often endorsed at the expense of logical necessity (Barston, 1986; Evans et al., 1983; Janis and Frick, 1943; Morgan and Morton, 1944; Oakhill et al., 1990; Wilkins, 1928).

Paradigms similar to that originally established by Wason (1966) have been used to compare conditional reasoning performance under abstract and thematic conditions. The "selection task" has been administered in guises containing degrees of realistic material. These range from totally abstract scenarios, describing relations between meaningless symbols, through to much more realistic scenarios, describing: locations and transportation methods, letters and postage rates, foods and beverages, bars and drinking clientele. Some studies suggest that significantly improved performance is ascribable to the use of realistic material (Gilhooly and Falconer, 1974; Griggs and Cox, 1982; Pollard and Evans, 1981; Van Duyne, 1974 1971). Other studies report little or no evidence of improved performance (Manktelow and Evans, 1979; Reich and Ruth, 1982; Roberge and Antonak, 1979). Even repetitions of studies have failed to replicate the originally observed trends. Griggs and Cox (1982), for example, were unable to reproduce the findings of Wason and Shapiro (1971) and Johnson-Laird et al. (1972). Taken together, these results suggest that the facilitatory effects caused by thematic material are, at best, unreliable and may depend upon specific factors pertaining to both the task and the reasoner.

It is argued that the relatively poor performance of reasoners in certain abstract reasoning studies is attributable to the fact that people are much more accustomed to reasoning with realistic material on a frequent basis in everyday life, and that this additional experience accounts for the differential when compared with their performances in thematic studies (Wason and Shapiro, 1971). Unlike the normally sporadic rates of performance seen in purely thematic reasoning tasks, performance rates for abstract reasoning tasks tend to be either consistently low or consistently high for different types of logical problem. This phenomenon might be explained by the possibility that individuals adopt a fixed strategy when evaluating abstract arguments, and so the reasoner avoids contemplating alternative possible interpretations of the various terms involved; a process which frequently leads to inconsistent responses in thematic reasoning (Staudenmayer, 1975). The real problem for cognitive science, however, arises when reasoners respond inconsistently to different forms of abstract problem because, in these situations, it can be difficult to determine when performance is being affected by the logical requirements of the task or by the reasoner's attempts to generate concrete instances of the abstract symbols and reason by analogy.

Perhaps the main conclusion that can be drawn from cognitive studies comparing abstract and thematic reasoning is that any form of improved performance, whether by enhanced reasoning or cued information recall, is specific to linguistic properties of the task, such as content and context, or cognitive properties of the reasoner, such as prior beliefs and reasoning expertise. Moreover, those tasks which elude close associations with information stored in reasoners' semantic memories often seem to elicit responses that accord with prior belief, albeit at the expense of logical necessity. The findings of Manktelow and Evans (1979) and Wason and Shapiro (1971), in particular, have important implications for the design of future cognitive studies because they suggest that a "reasoning task" can no longer be classified as such if its content becomes so meaningful to the reasoner that it simply cues the correct solution to be "read off" from information stored in memory with no element of reasoning having occurred.

5.2 Aims and Methodology

The results of a study conducted by Evans (1977a) suggest that people are liable to reason in significantly different ways about logically equivalent "if p then q" and "p only if q" statements. This finding suggests that two conditional statements might be logically equivalent, yet psychologically disparate. The first aim of the present study was to test, in a similar fashion, whether the logically equivalent formal expression "$p \Rightarrow q$" and the natural language statement, "if p then q" are psychologically equivalent. Cognitive research suggests that the type of inference to be drawn can exert a significant influence on reasoning performance (Evans, 1977a; 1983a; 1983b; Evans et al., 1995; Staudenmayer, 1975; Taplin, 1971; Taplin and Staudenmayer, 1973). The second aim, therefore, was to test the ease with which reasoners are able to draw specific types of logically valid conditional inference and their proneness to classical fallacies. In light of findings which suggest that thematic content can facilitate sound reasoning for natural language based conditionals (Cheng and Holyoak, 1985; Griggs and Cox, 1982; Wason and Shapiro, 1971), the third aim was to investigate the extent to which reasoning performance can be influenced by manipulating the levels of realistic content in task material. In light of findings which suggest that reasoning can be impaired by the presence of negative components in premiss information (Evans, 1972c; 1977a; 1993; Wason, 1959; Johnson-Laird and Tridgell, 1972), the fourth aim was to test how far reasoning performance can be affected by varying the polarities of logical terms in conditional rules.

5.2.1 Participants

A total of sixty computing scientists volunteered to take part in the experiment. These comprised staff and students from academic institutions and computing professionals from industrial software companies, all of whom were recruited by personal invitation. All participants were native English language speakers and were randomly selected. Participants were divided equally into three linguistic groups: Abstract Natural Language (ANL), Abstract Formal Logic (AFL) and Thematic Formal Logic (TFL). The groups were loosely matched, firstly, according to participants' personal ratings of Z expertise and, secondly, according to their lengths of Z experience. The ANL group comprised 16 students, 2 staff, 1 professional and 1 other. Their mean age was 27.00 years ($s = 9.41$) and 13 had studied a system of formal logic beforehand, such as the propositional or predicate calculus, Boolean algebra or Higher Order Logic. The AFL group comprised 11 students, 7 staff and 2 professionals. Their mean age was 32.75 years ($s = 13.06$) and 13 had studied a system of formal logic beforehand. Their mean level of Z experience was 3.82 years ($s = 4.09$). According to participants' personal ratings of expertise, the group comprised 9 novice, 8 proficient and 3 expert users of the Z notation. The TFL group comprised 4 students, 12 academic staff and 4 software professionals. Their mean age was 31.15 years ($s = 6.54$) and 18 had studied a system of formal logic beforehand. Their mean level of Z experience was 3.34 years ($s = 3.16$) and the group comprised 7 novice, 8 proficient and 5 expert users.

5.2.2 Design

The study had a three factor mixed design. The first, between groups, factor was the language in which the problem material was presented: ANL, AFL and TFL. It should be noted that the inclusion of a natural language based group was motivated by the need to demonstrate the feasibility of the methodology for testing people's abilities to reason about logically equivalent tasks in different notations, rather than to provide a basis from which metrics could be derived. The second, repeated measures, factor was the type of inference to be drawn and had four levels: MP, MT, DA and AC. It should be noted that the MP and MT inferences lead to determinate conclusions, whereas AC and DA are fallacious inferences in which nothing can be deduced logically. The third, repeated measures, factor was the polarity of the premiss pairs and had four levels: AA, AN, NA and NN (where A and N correspond to the position of affirmative and negative components in the conditional premisses respectively). Three sets of logically equivalent tasks were designed and presented to the three experimental groups. The underlying logical forms of the tasks corresponded to the sixteen possible combinations shown in Table 5.1, that is, four different types of inference each with four different types of premiss polarity.

TABLE 5.1

Logical forms of the conditional inference tasks

Polarity	*MP*	*MT*	*DA*	*AC*
AA	if p then q, p, $\therefore q$	if p then q, $\neg q$, $\therefore \neg p$	if p then q, $\neg p$, $\therefore \neg q$	if p then q, q, $\therefore p$
AN	if p then $\neg q$, p, $\therefore \neg q$	if p then $\neg q$, q, $\therefore \neg p$	if p then $\neg q$, $\neg p$, $\therefore q$	if p then $\neg q$, $\neg q$, $\therefore p$
NA	if $\neg p$ then q, $\neg p$, $\therefore q$	if $\neg p$ then q, $\neg q$, $\therefore p$	if $\neg p$ then q, p, $\therefore \neg q$	if $\neg p$ then q, q, $\therefore \neg p$
NN	if $\neg p$ then $\neg q$, $\neg p$, $\therefore \neg q$	if $\neg p$ then $\neg q$, q, $\therefore p$	if $\neg p$ then $\neg q$, p, $\therefore q$	if $\neg p$ then $\neg q$, $\neg q$, $\therefore \neg p$

5.2.3 Materials

The premisses and conclusions of the tasks were expressed in the form of natural English for the ANL group, and in the form of Z predicate expressions for the AFL and TFL groups. The linguistic content of both sets of abstract tasks was confined to describing relations between colours and shapes, so as to minimise the possible interference of real world content. A series of sixteen different scenarios were designed to elicit associations with participants' prior beliefs and intuitions for the thematic group. These related to imaginary but realistic computing applications including: a library database system, a flight reservation system, a missile guidance system, a video lending system, and a vending machine operation. So as to minimise any potential conflict between logic and prior belief, all tasks were designed to lead to believable conclusions, that is, to plausible conceptions of the corresponding real world applications. More than one plausible conclusion was included in the available response options in order to avoid the correct answers simply being "read off" from memory with no recourse to reasoning processes. The experimental materials are exemplified in Figures 5.1, 5.2 and 5.3. These show the AC-AN task presented to the AFL, ANL and TFL groups respectively. All task sheets were computer generated and contained accompanying instructions, as shown in Appendix A.

If the shape is a circle then the colour is not blue.
The colour is not blue.
Based on the above description, what can you say about shape?

(a)	The shape is not a rectangle	(c)	The shape is not a circle
(b)	The shape is a circle	(d)	Nothing

Figure 5.1: Task AC-AN presented to the ANL group

If $colour' \neq blue$ after its execution, what can you say about the value of $shape$ before operation $SetColour$ has executed?

$SetColour$

$\Delta ShapeAndColour$

$(shape = circle) \Rightarrow (colour' \neq blue)$
$shape' = shape$

(a)	$shape \neq rectangle$	(c)	$shape \neq circle$
(b)	$shape = circle$	(d)	Nothing

Figure 5.2: Task AC-AN presented to the AFL group

If $\neg(reactor\_status! = Ok)$ after its execution, what can you say about $coolertemp$ before operation $ReactorTempCheck$ has executed?

$ReactorTempCheck$

$\Xi NuclearPlantStatus$
$reactor\_status! : Report$

$coolertemp > Maxtemp \Rightarrow \neg(reactor\_status! = Ok)$

(a)	$coolertemp \leqslant Maxtemp$	(c)	$coolertemp > Mintemp$
(b)	$coolertemp > Maxtemp$	(d)	Nothing

Figure 5.3: Task AC-AN presented to the TFL group

5.2.4 Procedure

Before starting the experiment, the ANL group was asked to provide biographical details including: occupation, age, organisation, course, division, year of study, and a description of any systems of formal logic studied beforehand (such as the propositional or predicate calculus, Boolean algebra or Higher Order Logic). The AFL and TFL groups were asked to provide the following additional information: number of years' Z experience, a list of other formal notations known, and a personal rating of their Z expertise (novice, proficient or expert). The two formal logic based groups were then shown the instructions below.

> "In each of the tasks that follow, you will be shown a Z operational schema and a description of the operation's execution. You will be asked to determine which one of four given statements follow from the information given. Please circle the letter of your choice. You will also be asked to give a confidence rating, which should indicate how far you believe your answer to be correct. Please complete all tasks to the best of your ability, without reference to textbooks. The experiment should take no longer than 30 minutes to complete."

The same instructions were shown to the ANL group with the exception of only the first sentence; "In each of the tasks that follow, you will be shown a description of a colours and shapes scenario". The AFL group was also told that they may assume the global Z type definitions shown in Figure 5.4, which clarified the range of values that could be assigned to variables.

$$\begin{array}{l} \mathit{SHAPE} ::= \mathit{square} \mid \mathit{circle} \mid \mathit{triangle} \mid \mathit{rectangle} \\ \mathit{COLOUR} ::= \mathit{red} \mid \mathit{green} \mid \mathit{blue} \mid \mathit{white} \end{array}$$

ShapeAndColour

$$\begin{array}{|l} \mathit{shape} : \mathit{SHAPE} \\ \mathit{colour} : \mathit{COLOUR} \\ \hline \end{array}$$

Figure 5.4: Global Z type definitions

For each task participants were shown: two Z predicates representing the premisses of a conditional inference, three Z predicates representing possible determinate conclusions, labelled "(a)" to "(c)", and a fourth predicate, labelled "(d)", representing a possible indeterminate conclusion. Participants were asked to select the one conclusion that followed from the given premiss pair by circling the appropriate letter, then to give a rating of the extent to which they believed their response was correct by ticking one of the corresponding boxes shown below. These boxes were coded 1, 2 and 3 respectively for the purposes of analysis. Task sheets were distributed to participants and completed anonymously then mailed back to the experimenter. All participants were tested on an individual basis.

Confidence rating: □ Not confident □ Guess □ Confident

5.3 Results

Valid Inferences

The rates of valid inferences endorsed by the three linguistic groups are shown in Figure 5.5. An analysis of variance revealed a significant effect of group type on correctness ($F_{(2,57)} = 7.19, p < 0.01$). A comparison of group correctness revealed a rank order as follows: ANL ($\bar{x} = 67\%$) < AFL ($\bar{x} = 79\%$) < TFL ($\bar{x} = 90\%$). A Scheffe post hoc comparison revealed no significant differences in performance between the ANL and AFL groups, nor between the AFL and TFL groups, but a significant difference between the ANL and TFL groups ($p < 0.01$).

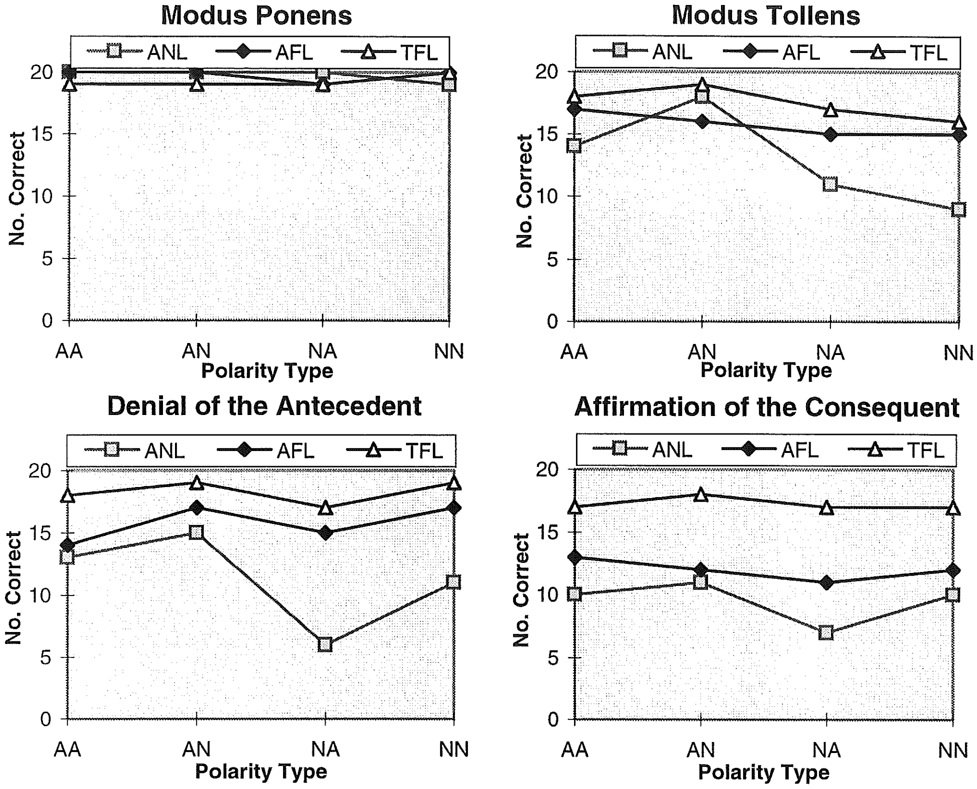

Figure 5.5: Frequencies of valid conditional inferences endorsed ($n = 20$)

Fallacious Inferences

The frequencies of fallacious DA and AC inferences endorsed are shown in Figure 5.6. Analyses of variance revealed significant effects of group type ($F_{(2,57)} = 7.33, p < 0.01$), polarity type ($F_{(3,171)} = 7.73, p < 0.01$), and inference type ($F_{(1,57)} = 11.79, p < 0.01$) on participants' susceptibility to the fallacious inferences. A comparison of group susceptibility to the fallacies revealed a rank order as follows: TFL ($\bar{x} = 8\%$) < AFL ($\bar{x} = 29\%$) < ANL ($\bar{x} = 46\%$). A Scheffe post hoc comparison of group susceptibility to the fallacious inferences revealed no significant differences in performance between the ANL and AFL groups, nor between the AFL and TFL groups, but a significant difference between the ANL and TFL groups ($p < 0.01$).

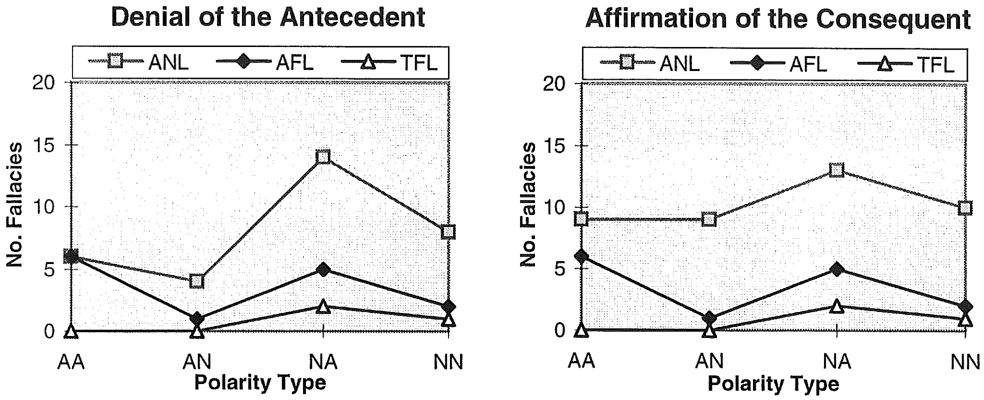

Figure 5.6: Frequencies of fallacious conditional inferences endorsed ($n = 20$)

Inference and Polarity Type

An analysis of variance revealed a significant effect of inference type ($F_{(3,171)} = 16.61, p < 0.01$) and a significant effect of premiss polarity on correctness ($F_{(3,171)} = 11.18, p < 0.01$). An analysis of variance revealed a significant interaction between inference and group type ($F_{(6,171)} = 2.73, p = 0.01$). This interaction is consistent with the finding that the TFL group outperformed the AFL group which, in turn, outperformed the ANL group for three of the four inference types. A further analysis of variance revealed a significant interaction between inference and polarity type ($F_{(9,513)} = 2.57, p < 0.01$). This interaction might be attributed to the fact that most errors were observed for the NA and NN polarities across three of the four inference types. A comparison of correctness for inference type revealed a rank order as follows: AC ($\bar{x} = 65\%$) < DA ($\bar{x} = 75\%$) < MT ($\bar{x} = 77\%$) < MP ($\bar{x} = 98\%$). A comparison of correctness for premiss polarity revealed a rank order as follows: NA ($\bar{x} = 73\%$) < NN ($\bar{x} = 77\%$) < AA ($\bar{x} = 80\%$) < AN ($\bar{x} = 85\%$).

Experience and Expertise

A linear regression analysis revealed significant correlations between participants' length of Z experience and correctness (Adjusted $R^2 = 0.15, F_{(1,39)} =$

$7.82, p < 0.01$), and between their Z expertise rating and correctness (Adjusted $R^2 = 0.18, F_{(1,39)} = 9.40, p < 0.01$). These correlations suggest that those participants with relatively high levels of experience and expertise with the Z notation reasoned more logically than those without. This finding is supported by numerous cross-cultural studies which attribute improved cognitive performance to increased language familiarity (see for example: Brown et al., 1980; Kiyak, 1982; Okonji, 1971).

Confidence Ratings

An analysis of variance revealed no significant effect of group type on confidence, but a significant effect of inference type on confidence ($F_{(3,171)} = 2.90, p = 0.04$), and an effect of polarity type on confidence approaching significance ($F_{(3,171)} = 2.55, p = 0.06$). A comparison of the mean confidence ratings for group type revealed a rank order as follows: TFL ($\bar{x} = 2.81$) < AFL ($\bar{x} = 2.83$) < ANL ($\bar{x} = 2.84$). A comparison of the mean confidence ratings for inference type revealed a rank order as follows: MT ($\bar{x} = 2.78$) < DA ($\bar{x} = 2.82$) < AC ($\bar{x} = 2.83$) < MP ($\bar{x} = 2.87$). A comparison of the mean confidence ratings for polarity type revealed a rank order as follows: NN ($\bar{x} = 2.79$) < AN ($\bar{x} = 2.82$) < NA ($\bar{x} = 2.83$) < AA ($\bar{x} = 2.85$).

Few clear trends are evident in the observed confidence ratings, perhaps owing to participants' overconfidence. It is interesting to note, however, isolated correspondences between confidence and correctness. For example, all participants expressed a maximum confidence rating for the ANL group's MP-AA inference which corresponds with the perfect success rate observed for this task. Whereas a low mean confidence rating was observed for the ANL group's MT-NN inference, which corresponds with the fact that more than half of participants erred on this task. Figure 5.7 reflects the high levels of confidence declared by participants.

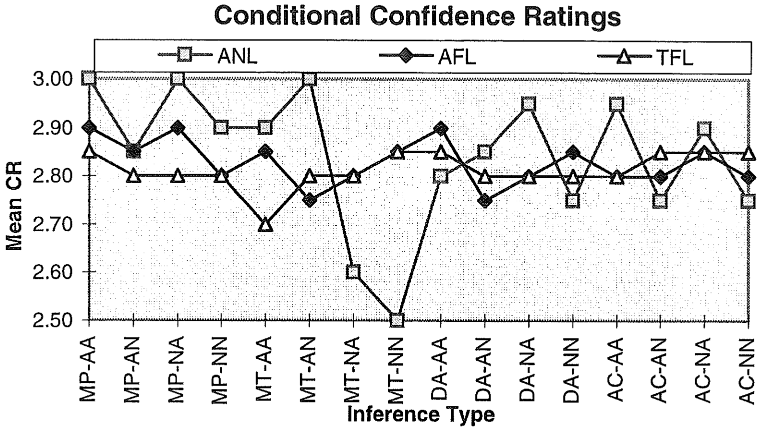

Figure 5.7: Confidence ratings for conditional inferences ($1 \leqslant CR \leqslant 3$)

5.4 Discussion

Valid Inferences

The results show that participants frequently erred in both abstract groups, but signs of significantly improved performance are evident in the thematic group. With regard to the two valid inference types, few participants appeared to experience difficulties in drawing MP inferences, with near ceiling levels observed for all premiss polarity combinations across all three groups. This finding is supported by natural language based studies which suggest that reasoners rarely err when drawing MP inferences, irrespective of the term polarities (Evans, 1972b; 1977a; Evans et al., 1995; Taplin, 1971) and the degree of realistic content involved (Griggs and Cox, 1982; Manktelow and Evans, 1979; Pollard and Evans, 1987). These high rates of performance under experimental conditions might be attributed to the fact that people are accustomed to drawing affirming MP inferences on a frequent basis in everyday life. The lower rate of correct MT inferences is also supported by natural language based studies (Evans, 1977a; Taplin, 1971; Taplin and Staudenmayer, 1973). A planned comparison for the MP and MT inference types revealed this difference to be significant ($F_{(1,15)} = 48.22, p < 0.01$). The rates of incorrect MT inferences were particularly noticeable in the ANL group where up to 55% erred.

For the Wason selection task, it is argued that reasoners need to draw MP and MT inferences in order to see the relevance of the p and $\neg q$ cases respectively (Griggs and Cox, 1982; Manktelow and Evans, 1979; Pollard and Evans, 1981). In abstract variants of the task, most reasoners select the p case, but nearly all fail to select the $\neg q$ case, and this trend is often attributed to the relative difficulty of drawing MP and MT inferences. Although signs of improved performance are reported for realistic variants of the task (Gilhooly and Falconer, 1974; Johnson-Laird et al., 1972; Van Duyne, 1974; Wason and Shapiro, 1971), the rates of correct MT inferences are generally much higher in the present study. This difference might be explained by the possibility that the logical demands of the selection task are in fact more complex than they are commonly purported to be. O'Brien (1993; 1995) argues that the logical structure of the selection task goes way beyond drawing simple MP and MT inferences because its successful resolution depends upon participants reasoning via additional principles of logic, such as reductio ad absurdum. Assuming that the logical structure of the selection task and the additional lines of reasoning required to solve it are indeed too complex for most participants to grasp, within the confines of the time and mental effort that they are willing to expend on the task, this would seem to account for the difference in rates at which its participants tend to select the $\neg q$ case and the rates at which the present study's participants were observed to draw correct MT inferences.

Fallacious Inferences

With regard to the two fallacious inference types, the results suggest that participants experienced substantial difficulties in drawing DA inferences, where up to 70% committed the fallacy, and AC inferences, where up to 65% committed the fallacy. The high rates at which the abstract groups succumbed to these fallacies is widely supported in the cognitive literature (Evans, 1972b; 1983; Evans et al., 1995; Taplin, 1971; Taplin and Staudenmayer, 1973).

The high rates of DA and AC fallacies endorsed by the ANL group suggest that many of these participants adopted a more symmetrical, biconditional, interpretation of the conditional rules than that adopted by the formal logic groups. This might be attributed to two factors. First, it is possible that the clearly unidirectional appearance of the arrow in the formal operator "$\Rightarrow$" led participants away from biconditional interpretations of the conditional rules and into an increased appreciation of their asymmetrical nature - a trend which was born out in the results of our initial study. Second, biconditional interpretations of "if ... then" statements are often adopted in everyday discourse where they lead reasoners to conclusions that are both pragmatically sanctionable and sufficient for their purpose (Evans et al., 1993; Geis and Zwicky, 1971). From the statements "If the switch is up then the light is on" and "The light is on", for example, one might be inclined to infer that "The switch is up", despite the fact that this inference does not follow logically. Attempts to apply the same principles that govern everyday discourse to strictly logical tasks, however, invariably lead to error. Although the DA and AC inferences endorsed by the ANL and AFL groups were unlikely to have been influenced significantly by pragmatic associations, owing to the abstract nature of the material involved, it appears that this did not prevent participants from adopting pragmatic biconditional interpretations of the conditional rules.

Content Effects

The theory of "facilitation by realism" claims that thematic, as opposed to abstract, problem content can significantly improve reasoning performance (Gilhooly and Falconer, 1974; Griggs and Cox, 1982; Johnson-Laird et al., 1972; Van Duyne, 1974). To determine whether participants succumbed to this bias, we conduct our analysis at the group level. The fact that the TFL group significantly outperformed both abstract groups supports this theory. It is argued that reasoning tasks which elude associations with information stored in reasoners' semantic memories are more likely to elicit responses that accord with prior belief, albeit sometimes at the expense of logical necessity (Barston, 1986; Evans et al., 1983; Henle and Michael, 1956; Oakhill et al., 1990). This form of "belief bias" seems most prominent when speculative conclusions do not run contrary to existing convictions, but conform with reasoners' beliefs (Dominowski, 1995). The fact that all thematic tasks in the present study were designed to lead to believable conclusions may therefore have contributed to the high levels of performance observed.

Polarity Effects

If the presence of negative components can impair reasoning performance (Wason, 1959), then one might reasonably expect the difficulty of an inferential task to increase along with the number of negative terms in its logical premisses. Analyses of the rates of correct MT, DA and AC inferences, however, revealed no such simple relation in the results. Instead, the results for all three groups suggest complex relations between reasoning performance, inference type and premiss polarity. The rates of correct MT inferences endorsed by the ANL group support Evans' (1972b) hypothesis that negating the antecedent of the conditional premiss renders MT inferences more difficult, whereas negating the consequent alone has little effect. The results also suggest that DA and AC fallacies were most likely to be committed by the ANL group where the antecedent of the conditional premiss was negative and the consequent was affirmative. This finding appears to contradict Evans' (1993a) theory of affirmative premiss bias because it appears that participants were more inclined to draw determinate conclusions from negative premisses.

The theory of affirmative premiss bias argues that reasoners are more inclined to endorse determinate conclusions from premisses that do not contain any negative components (Evans, 1993a). To determine whether participants succumbed to this bias, we focus our attention on the responses to those tasks whose premisses were both affirmative. Comparison of the scores for MP-AA with the other MP tasks suggests that participants were no more likely to draw the logical conclusion simply because both premisses were affirmative. Similarly, comparison of the scores for AC-AA with the other AC tasks suggests that participants were no more likely to commit this fallacy simply because both premisses were affirmative. The fact that these trends suggest participants exhibited no signs of affirmative premiss bias in the present study is supported by Evans et al. (1995).

The theory of negative conclusion bias argues that reasoners are more inclined to endorse inferences whose conclusions are negative rather than affirmative (Evans, 1972c; 1977a; 1993a). Had participants succumbed to this bias one would expect to see four trends in the results. First, one would expect to see more correct MP inferences for conditionals with negative rather than affirmative consequents. Given that the correct responses for MP inferences approach ceiling levels in all three groups, however, the presence or absence of negatives in MP conclusions would appear to have had little effect. Second, one would expect to see more correct MT inferences for conditionals with affirmative rather than negative antecedents. Comparison of the correct response rates for MT-AA and MT-AN with those for MT-NA and MT-NN reveals that such a trend is born out in the results, particularly in the ANL group. This result replicates the findings of Evans et al. (1995), which suggest that reasoners are particularly prone to the bias when drawing MT inferences about problems in abstract natural language. Third, one would expect to see more fallacious DA inferences for conditionals with affirmative rather than negative consequents. Comparison of the correct response rates for DA-AA and DA-NA with those for DA-AN and DA-NN reveals that such a trend is only evident in the AFL group. This suggests that the bias can be evoked in abstract formal contexts for DA inferences; a finding replicated in abstract informal contexts by Evans et al. (1995). Fourth, one would expect to see more fallacious AC inferences for conditionals with negative rather than affirmative antecedents. Comparison of the correct response rates for AC-NA and AC-NN with AC-AA and AC-AN reveals that such a trend is apparent only in the ANL group. This suggests that the tendency to endorse non-logical conclusions was again strongest for tasks couched in abstract terms.

It is argued that people rarely recognise a double negative as an affirmative in everyday discourse, but instead disregard the extra negatives (Wason, 1959; Evans, 1972b; 1972c; 1983a; 1983b). If this bias had occurred one would have expected members of the TFL group, for example, to fail to see how "$\neg\neg(status! = Success)$" could result in "$status! = Success$". To determine whether participants succumbed to this bias, we focus on those tasks in which double negation leads to a valid conclusion: MT-AN, MT-NA, MT-NN. The results suggest that all three groups experienced no difficulty in drawing MT-AN inferences, but that some difficulties were experienced in drawing MT-NA and MT-NN inferences, particularly in the AFL group. The high rates of MT-NN errors might be attributed to the fact that this task requires a double negation of both the consequent and the antecedent of the conditional rule in order to reach the correct solution: *if not p then not q*; *q* (*not not q*); therefore *p* (*not not p*). Whilst this might account for the difference in scores for the MT-NN and MT-AN inferences, however, it does not explain the difference in scores for MT-NA and MT-AN, where only one double negation is required. Insofar as the MT-AN inference gives rise to a negative conclusion and the MT-NA and MT-NN inferences give rise to affirmative conclusions, the difference in scores for these inferences appears to corroborate more with the theory of negative conclusion bias. Aside from these isolated cases, it should be noted that participants successfully converted double negatives in numerous other tasks, even in the case of DA and AC inferences where it seemingly led them to fallacious conclusions.

Matching Bias

Matching bias theory claims that reasoners select only those conclusions which contain one or more of the terms explicitly mentioned in the given premisses (Evans, 1972b; 1983a; 1983b). To determine whether participants succumbed to this bias, we focus on those inferences where the correct conclusion requires making explicit a logical term not explicitly stated in the premisses: MT, DA and AC. For matching bias to have occurred for the MT inferences, reasoners would have had to select responses containing the same antecedent from the conditional premiss. No such trend is visible in the results; of those participants who erred on MT tasks, most gave indeterminate responses. When a reasoner commits a DA fallacy the response contains a negation of the consequent, and when a reasoner commits an AC fallacy the response contains a positive antecedent. Judging by the high rates of these fallacies committed, which are consistent with biconditional interpretations, it might be argued that matching bias was evident in both abstract groups.

Confidence Ratings

It is argued that reasoners who tend not to rate MT inferences as more difficult than MP inferences do not appreciate the asymmetrical nature of the logical conditional (O'Brien and Overton, 1982). Casual inspection reveals a clear correspondence between confidence and correctness for the MP and MT inferences in the ANL group; the higher the confidence rating, the higher the score. The strength of this correspondence declines as one inspects participants' confidence ratings for these inferences in the AFL group, however, and disappears altogether in the TFL group. On this basis, it might be claimed that participants' appreciation of the relative difficulty of MP and MT inferences was strongest in the natural language group but weakest in the two formal logic groups, where participants were justifiably confident throughout. Inspection of the levels of correctness for the DA and AC inferences, however, prevents us from making any such claim. One would expect reasoners who do not appreciate the one way nature of the conditional to give responses which conform with a biconditional interpretation when they are presented with DA and AC inferences. This hypothesis is well supported insofar as a biconditional interpretation appears to have led many participants to endorse these fallacies, particularly in the two abstract groups. So perhaps the most that can be inferred, therefore, is that participants appreciated the relative complexities of these types of inference, but failed to recognise that the asymmetrical nature of the conditional was at least partly responsible for this difference in complexity.[2]

That participants frequently erred, despite their high levels of confidence, suggests that they were often overconfident in the correctness of their responses. This might be attributed to two factors. First, the seemingly trivial nature of some tasks may have instilled a false sense of security which participants retained even when completing the more complex tasks. Second, that participants appeared reluctant to admit any form of doubt may be due to the fact that the English language based tasks were presented exclusively to people whose native language was English, and who were accustomed to reasoning with English based arguments in everyday life. Similarly, members of the two formal groups were either studying or teaching Z as part of some university degree course, or applying Z to provide business solutions to industrial problems. Possibly for political reasons, therefore, one might have expected members of the two formal groups in particular to proclaim high levels of confidence throughout, irrespective of the difficulty of the tasks.

[2] Some limitations of the way in which confidence ratings were extrapolated during the three main experiments are discussed in the final chapter of this thesis.

5.5 Conclusions

There would appear to be considerable differences in the ways that people reason about English statements containing "if ... then" and formal statements containing "⇒", despite their logical equivalences. The results of the present study suggest that the users of natural language and the users of formal methods are liable, under certain conditions, to succumb to similar non-logical strategies when reasoning about conditional rules. These strategies appear to include some of the reasoning heuristics and biases that natural language based studies have shown people to adopt in favour of logical rules: matching bias (Evans, 1972b), negative conclusion bias (Evans, 1993a), illicit conversion of double negatives (Evans, 1972c), and biconditional interpretations of conditionals (Geis and Zwicky, 1971). That the rates of correctness were highest in the thematic group for nearly all inference types across all premiss polarities suggests that these participants were less reliant on everyday heuristics than on logical principles. This may be somewhat of a surprising finding given that one might expect thematic material to cue the reasoner into using everyday lines of thought which often diverge from the laws of logic. The fact that both formal logic groups outperformed the natural language group might be interpreted in favour of the formalists' claim that it is easier to reason about formal logic than natural language (Thomas, 1995), however, there were still numerous occasions on which even the formal reasoners appeared prone to systematic error and bias.

5.6 Summary

This chapter began with a review of some common errors committed by people without logical training in natural language based studies of conditional reasoning. It has shown how the results and hypotheses emerging from these studies, together with the results obtained from the initial investigation, served as a basis for design of the first main experiment. The findings of this experiment suggest that, despite clear differences in language symbology, several of the conditional reasoning errors and biases that people commit in natural language based contexts are also liable to occur when trained users of formal methods are reasoning about logically equivalent statements in formal specifications. Analysis of the results suggests that participants' reasoning performance was influenced significantly by each of the main experimental variables: the type of inference to be drawn, the meaningfulness of task content, and the polarity of premiss terms. We shall investigate how far different combinations of such variables are liable to influence people's reasoning with other forms of logical rule in the forthcoming chapters.

Chapter 6

Disjunctive and Conjunctive Reasoning

> "*Or* is in some respects like *and* in its behaviour, in others like *but*, and in still others, different from both. *Or* is like the other two in that it requires a common topic between the two conjuncts, and in that this common topic may be overtly present or derivable by presupposition and deduction" (Lakoff, 1971, p.142).

One might expect the users of formal methods to exhibit similar heuristics when reasoning about conditionals, disjunctives and conjunctives, given that these rules share a number of common logical properties. Our study of conditional reasoning suggests that many of the non-logical heuristics and biases that have been demonstrated in natural language based studies of conditional reasoning are also liable to occur when the users of formal methods are reasoning about the equivalent statements in formal specifications. Based on cognitive theories of human reasoning in natural language contexts, this chapter reports an empirical study designed to test the extent to which this apparent "transfer effect" occurs for disjunctives and conjunctives in formalised contexts.

6.1 Error and Bias in Disjunctive Reasoning

The logical connective "or" is used liberally in everyday communication to join and express choice between statements. Many writers take for granted, however, the complex linguistic rules and hidden conventions that govern its use. Hurford (1974) argues that disjunction is misused in the English language wherever one disjunct entails the other, so sentences such as "John is British or American" should be considered legal, whereas "John is British or a Londoner" should be avoided. This is not to suggest, however, that two disjunctive sentences should be completely unrelated as in, for example, "The car is red or London is the capital of England". The overwhelming consensus is that any two conjoined sentences in the English language must share a common topic (Lakoff, 1971; Fillenbaum, 1974).

Aside from its common misuse by writers, disjunctive statements can evoke ambiguous interpretations in readers. Besides what is written explicitly, extra linguistic factors such as context, register and intonation can provide additional clues to a speaker's intended meaning and help to resolve ambiguities (Turner, 1986). For decades psychologists and linguists have been trying to establish the precise conditions under which readers are obliged to draw inclusive or exclusive interpretations. Normally when we are offered refreshment, for example, the question "Tea or coffee?" requires an exclusive interpretation, whereas "Milk or sugar?" requires an inclusive interpretation, and we can determine this with the help of contextual clues. The resolution of disjunctive ambiguities in everyday language has been strongly debated, with some arguing that "or" in English is generally inclusive (Pelletier, 1977), some claiming that it is generally exclusive (Lakoff, 1971), and others arguing that the correct interpretation depends upon linguistic factors such as the context or form of the sentence (Hurford, 1974; Newstead and Griggs, 1983a). The truth table in Figure 6.1 shows that the only condition under which this ambiguity can arise is when both disjuncts are true; under an inclusive interpretation the whole sentence would be true, but under an exclusive interpretation it would be false.

P	Q	Inclusive Or	Exclusive Or
F	F	F	F
F	T	T	T
T	F	T	T
T	T	T	F

Figure 6.1: Truth tables for inclusive and exclusive disjunction

Where this form of ambiguity arises, cognitive studies suggest that people generally prefer to draw exclusive interpretations, although the strength of preference has been found to vary according to the context in which the disjunctive is presented (Newstead and Griggs, 1983a; Newstead et al., 1984). Developmental studies suggest that people begin childhood with a strong preference for inclusive interpretations, and gradually develop a preference for exclusive interpretations (Sternberg, 1979; Braine and Rumain, 1981). One factor that might confound the results of such studies is the possibility that young children respond to disjunctive statements as if they were conjunctives, thereby giving the misleading impression that they are adopting inclusive interpretations. A similar phenomenon has been observed under experimental conditions, whereby adult reasoners appeared to confuse the principles of "and" and "or" by using the terms synonymously (Leahey, 1980; Newstead et al., 1984; Roberge, 1977; Wason and Johnson-Laird, 1969).

Once the correct interpretation has been adopted, cognitive research suggests that people find it easier to reason about exclusive, rather than inclusive, disjunctives (Newstead et al., 1984; Newstead and Griggs, 1983a; Roberge, 1977; 1978). Two explanations for this difference in complexity are offered. First, it might be that people are more adept at reasoning with exclusive disjunctives simply because these are the more common form in everyday usage. Second, it might be due to the fact that exclusive disjunctives lead to symmetrical inferences, and by knowing the truth value of one disjunct the truth value of the other can be inferred. This contrasts with inclusive disjunctives, where simply knowing the truth value of one disjunct is not sufficient for inferring the truth value of the other. It would appear that reasoning performance, nevertheless, can be improved significantly when it is clarified to reasoners which interpretation is to be drawn. This has typically been achieved by adding qualifying instructions to the disjunctive rules; "*p or q* (or both)" for inclusive disjunction, "*p or q* (but not both)" for exclusive disjunction.

Mathematical logic appears to have developed an almost universal bias against exclusive disjunction, despite its frequent occurrence in everyday language, and tends to favour the inclusive form alone. The extent of the bias is typified by standard propositional logic which contains a formal operator for expressing inclusive disjunction but no corresponding operator for exclusive disjunction. Although the propositional operator "$\vee$" was derived from the Latin term for inclusive disjunction, *vel*, it seems odd that there was no corresponding operator derived from its term for exclusive disjunction, *aut*, in a similar fashion.[3] Newstead and Griggs (1983a) offer two possible explanations. First, it is advantageous from a parsimonious perspective because all logical operations can be defined in terms of inclusive disjunction and negation. Second, the inclusive operator complements the set union operator, "$\cup$", because both refer to either one of two propositions or sets, and possibly both.

The first argument shown in Figure 6.2 is commonly referred to as a "denial inference" because the major premiss specifies two disjunctive terms, one of which is explicitly denied in the minor premiss, resulting the affirmation of the other term. The inference is logically valid under either an inclusive or an exclusive interpretation of the disjunctive rule. The second argument in Figure 6.2 is referred to as an "affirmation inference" because the minor premiss affirms one of the terms in the

[3] Several texts on the Z notation compensate for this by introducing non-standard symbols. Diller (1994), for example, introduces the "$\|$" symbol to denote exclusive disjunction.

major premiss, resulting in the denial of the other. It is valid only under an exclusive interpretation of the disjunctive rule, however, and would be indeterminate under an inclusive interpretation because both disjuncts might be true.

$$\begin{array}{ll} p \text{ or } q & \\ \text{not } p & \\ \hline & \textit{Or}\ \text{denial} \\ q & \end{array} \qquad \begin{array}{ll} p \text{ or } q & \\ p & \\ \hline & \textit{Or}\ \text{affirmation} \\ \text{not } q & \end{array}$$

Figure 6.2: Denial and affirmation inferences for *Or*

Evans et al. (1993) report that the denial inference was made correctly by 84% and 80% of participants for exclusive and inclusive disjunctives respectively. The same study reports that 83% of participants drew the affirmation inference correctly for exclusive disjunctives, but that 36% persisted in drawing it for inclusive disjunctives where it constitutes a logical fallacy. The experimenters attribute poor performance on the affirmation task for inclusive disjunctives to the fact that the correct conclusion is indeterminate and that people have a general preference for drawing determinate, true or false, conclusions.

6.2 Error and Bias in Conjunctive Reasoning

There has been notably less cognitive research aimed at exploring conjunctive reasoning than that directed towards conditional and disjunctive reasoning. This might be due to the possibility that the linguistic rules and conventions governing the use of "and" are relatively simple and that people may be less prone to error and bias when drawing conjunctive inferences. Lakoff (1971) argues that the principles governing the use of disjunction and conjunction in the English language are rather similar, however, because both require a common topic between two terms and this may be explicit, or implicit but inferable. Lakoff points to the existence of a hierarchy of sentences conjoined by "and" with varying strengths of relation between their common topics. At the top of this hierarchy are sentences like "John eats apples and he eats pears", where the meaning of one conjoined sentence complements the meaning of the other. At the bottom are sentences like "John is a strict vegetarian and he eats lots of meat", where one of the conjoined sentences directly opposes the meaning of the other.

Cohen (1971) argues that an implicit dependence between conjuncts can be conveyed in everyday communication simply by changing the order in which they are spoken. The causal chain of events appears much clearer in, say, "The king has died and a republic has been declared" than in "A republic has been declared and the king has died". A considerate speaker aiming to help his or her audience's interpretation would normally prefer the former form. This view is supported by Lakoff who argues that the successful interpretation of a conjunctive sentence usually relies upon the presupposition of the first conjunct in order to facilitate understanding of the second. This contrasts markedly with disjunction, where the truth of the first disjunct is never presupposed, although its negation might be presupposed in order for the second disjunct to be considered true.

Those systems of formal logic defined in terms of the principles underlying Gentzen's logical deductive calculus (in Szabo, 1969) include two kinds of inference rule for connecting logical chains of reasoning; those for introducing and those for eliminating propositional connectives. Figure 6.3 shows the introduction and elimination rules for conjunction. Cognitive science has been slow, however, to question how far people adhere to these formal rules of inference when reasoning about conjunctive statements in everyday communication.

$$\frac{\begin{array}{l} p \\ q \end{array}}{p \text{ and } q}\;\textit{And}\text{ intro} \qquad \frac{p\ \textit{and}\ q}{p}\;\textit{And}\text{ elim 1} \qquad \frac{p\ \textit{and}\ q}{q}\;\textit{And}\text{ elim 2}$$

Figure 6.3: Introduction and elimination rules for *And*

Probability theory states that the likelihood of a conjunction, *p and q*, cannot exceed the likelihood of one of its constituent outcomes, *p* or *q*. A series of studies conducted by Tversky and Kahneman (1983) sought to uncover conflicts between logic and intuition when reasoning about conjunctives. The experimenters aimed specifically to test whether participants' systematic violations of this principle persisted across a variety of different contexts. Their results suggest that violation is likely whenever reasoners depart from principles of logic and adhere to intuitive heuristics, such as those based on "representativity" and "availability". The representativeness heuristic states that people are likely to judge an overall conjunction as more representative of a particular category than its individual constituents. The availability heuristic states that instances of a more inclusive category are easier to imagine and retrieve than those of an individual category. An empirical question which cognitive science has yet to address is whether these kinds of intuitive heuristic are also adopted by trained logicians and, hence, whether the conjunctive fallacy transfers into the formal domain.

Figure 6.4 shows De Morgan's laws (Diller, 1994; Lemmon, 1993). These are used in formal reasoning to convert disjunctions into conjunctions and vice versa. Although they are perhaps not as commonly used as propositional logic's more basic rules, such as the introduction and elimination rules for propositional connectives, it would be reasonable to expect most trained logicians to have developed an appreciation of them. The question of whether they see the relevance of these rules in a given problem situation, however, is another matter.

$$\frac{\textit{not (p or q)}}{\textit{not p and not q}}\ \textit{Not}\text{ over }\textit{Or} \qquad \frac{\textit{not (p and q)}}{\textit{not p or not q}}\ \textit{Not}\text{ over }\textit{And}$$

Figure 6.4: De Morgan's laws for disjunctives and conjunctives

The first argument shows how De Morgan's law can be applied to a negated disjunction to derive two separately negated conjuncts. The second argument shows how De Morgan's law can be applied to a negated conjunction to derive two separately negated disjuncts. It would be reasonable to expect that people who have acquired a fair degree of deductive competence, particularly those with prior training in mathematical logic, would be capable of combining De Morgan's laws with other rules of inference in a chain of logical reasoning. From the premisses "$\neg(p \wedge q)$" and "p", for example, it is possible to deduce the logical conclusion "$\neg q$". This is achieved by applying De Morgan's law over conjunction and then applying the rule for disjunctive elimination. Cognitive science has yet to test the ability of reasoners to draw explicitly formal, multiple stage, inferences of this kind.

6.3 Aims and Methodology

It is hard to envisage a language that could be much more precisely defined than one whose syntax and semantics are defined in terms of explicit, mathematical rules. Yet this is precisely how many formal notations are defined. The use of "$\vee$" and "$\wedge$" in formal logic is not constrained by the same linguistic rules and conventions that govern the use of "or" and "and" in the English language. Formal operators can be used to connect any two terms, for example, regardless of whether they share a common topic. Moreover, in certain branches of formal logic, such as the standard propositional calculus, the concept of exclusive disjunction is not even defined.

> "It is tempting to attribute the difficulty of disjunctive concepts to the fact that the word 'or' is so ambiguous in the English language ... One would expect that in languages where the word for disjunction is less ambiguous, performance on disjunctive concepts should be better" (Newstead and Griggs, 1983a, p.100).

One might not expect users of formal methods to encounter the kinds of linguistic problem experienced by those people who use disjunctives in natural language, such as drawing an exclusive interpretation when an inclusive interpretation is called for, or vice versa. The apparently strong bias which people exhibit towards exclusive interpretations under experimental conditions and in everyday life, however, suggests that people might still reason according to exclusive principles when they encounter formalised inclusive disjunctives. If this is the case then are the users of formal methods being asked to distort their natural thought processes and abandon intuitive judgement in favour of reasoning based on formal rules alone? These are the kinds of issue that the present study is aimed at illuminating.

6.3.1 Participants

A total of forty computing scientists volunteered to take part in the experiment. These comprised staff and students from academic institutions and computing professionals from industrial software companies, all of whom were recruited by personal invitation. All participants were native English language speakers and were randomly selected. Participants were divided equally into two linguistic groups: Abstract Formal Logic (AFL) and Thematic Formal Logic (TFL). The groups were loosely matched, first, according to participants' personal ratings of Z expertise and, second, according to their lengths of Z experience. The AFL group comprised 14 staff, 3 students and 3 professionals. Their mean age was 34.50 years ($s = 10.53$) and all had studied a system of formal logic beforehand. Their mean level of Z experience was 5.69 years ($s = 4.47$). According to participants' personal ratings of expertise, the group comprised 4 novice, 9 proficient and 7 expert users of the Z notation. The TFL group comprised 11 staff, 4 students and 5 software professionals. Their mean age was 35.00 years ($s = 9.85$) and 19 had studied a system of formal logic beforehand. Their mean level of Z experience was 4.67 years ($s = 3.80$), and the group comprised 4 novice, 9 proficient and 7 expert users.

6.3.2 Design

The study had a four factor mixed design. The first, between groups, factor was the degree of realistic material, abstract or thematic, corresponding to the two linguistic groups, AFL and TFL. The second, repeated measures, factor was the type of inference to be drawn and had six levels: disjunctive denial, disjunctive affirmation, conjunctive elimination, conjunctive introduction, De Morgan's over disjunction with conjunctive elimination, and De Morgan's over conjunction with disjunctive elimination. The third, repeated measures, factor was the polarity of the premiss pairs and had four levels for each of the denial, affirmation and elimination inferences: AA, AN, NA and NN (where A and N correspond to the position of affirmative and negative terms in the disjunctive or conjunctive premisses respectively). This factor also had two levels for the introduction inferences: A and N. The fourth, repeated measures factor, was the position of the logical term denied, affirmed, eliminated or introduced, and had two levels: first or second.

The disjunctive tasks involved the denial (DD) or the affirmation (DA) of a term from the major premiss by the minor premiss. The polarity of terms in the major premiss and the position of the term affirmed or denied in the minor premiss were varied systematically. Table 6.1 shows the logical forms of the sixteen disjunctive tasks. It should be noted that the conclusions shown for the denial inferences are logically sanctionable, but those shown for the affirmation inferences are fallacious under a logical, inclusive, interpretation of the Z notation's "$\vee$" operator.

TABLE 6.1
Logical forms of the disjunctive denial and affirmation tasks

Polarity	*Term Denied or Affirmed*	*DD*	*DA*
AA	1	$p \vee q,\ \neg p\ \therefore q$	$p \vee q,\ p\ \therefore \neg q$
AA	2	$p \vee q,\ \neg q\ \therefore p$	$p \vee q,\ q\ \therefore \neg p$
AN	1	$p \vee \neg q,\ \neg p\ \therefore \neg q$	$p \vee \neg q,\ p\ \therefore q$
AN	2	$p \vee \neg q,\ q\ \therefore p$	$p \vee \neg q,\ \neg q\ \therefore \neg p$
NA	1	$\neg p \vee q,\ p\ \therefore q$	$\neg p \vee q,\ \neg p\ \therefore \neg q$
NA	2	$\neg p \vee q,\ \neg q\ \therefore \neg p$	$\neg p \vee q,\ q\ \therefore p$
NN	1	$\neg p \vee \neg q,\ p\ \therefore \neg q$	$\neg p \vee \neg q,\ \neg p\ \therefore q$
NN	2	$\neg p \vee \neg q,\ q\ \therefore \neg p$	$\neg p \vee \neg q,\ \neg q\ \therefore p$

Note: The following abbreviation refers to the disjunctive inferences: DD/DA-<*Major premiss polarity*>-<*Term denied or affirmed*>.

The conjunctive reasoning tasks involved the elimination (CE) and the introduction (CI) of terms. The polarity of terms in premisses and the position of the term introduced or eliminated were manipulated. Table 6.2 shows the logical forms of the eight conjunctive inference tasks. It should be noted that the conclusions shown for the elimination tasks are logically sanctionable, but those shown for the introduction tasks are fallacious.

TABLE 6.2
Logical forms of the conjunctive elimination and introduction tasks

Polarity	*Term Eliminated*	*CE*	*Polarity*	*Term Introduced*	*CI*
AA	2	$p \wedge q\ \therefore p$	A	1	$p\ \therefore p \wedge q$
AN	1	$p \wedge \neg q\ \therefore \neg q$	A	2	$q\ \therefore p \wedge q$
NA	2	$\neg p \wedge q\ \therefore \neg p$	N	1	$\neg p\ \therefore \neg p \wedge q$
NN	1	$\neg p \wedge \neg q\ \therefore \neg q$	N	2	$\neg q\ \therefore p \wedge \neg q$

Note: The following abbreviation refers to the conjunctive inferences: CE/CI-<*Premiss polarity*>-<*Term Eliminated or Introduced*>.

In addition to the application of disjunctive or conjunctive rules in isolation, participants were given tasks requiring the application of two logical rules of inference: De Morgan's over disjunction followed by conjunction elimination (DMDCE), and De Morgan's over conjunction followed by disjunctive elimination (DMCDE). Owing to the additional complexity of these inferences, the polarities of the premiss terms were held constant and only the type of term eliminated was varied. The underlying structures of these tasks are illustrated in Table 6.3.

TABLE 6.3
Logical forms of the De Morgan's based tasks

Term Eliminated	*DMDCE*	*DMCDE*
1	$\neg(p \vee q) \ \therefore \neg q$	$\neg(p \wedge q),\ p \ \therefore \neg q$
2	$\neg(p \vee q) \ \therefore \neg p$	$\neg(p \wedge q),\ q \ \therefore \neg p$

Note: The following abbreviation refers to the De Morgan's inferences: DMDCE/DMCDE-<*Term eliminated*>.

6.3.3 Materials

The premisses and conclusions of the tasks were expressed as Z predicate expressions for both the AFL and TFL groups. The linguistic content of the abstract tasks was confined to describing relations between colours and shapes, so as to minimise the possible interference of realistic content. A series of twenty eight different scenarios were designed to elicit associations with participants' prior beliefs and intuitions for the thematic tasks. These related to imaginary but realistic computing applications such as: a missile guidance system, a live event's television coverage, a telephone network and a hotel reservation system. So as to minimise any potential conflict between logic and prior belief, all tasks were designed to lead to believable conclusions, that is, to plausible conceptions of the corresponding real world applications. More than one plausible conclusion was included in the available response options in order to avoid the correct answers simply being "read off" from memory with no recourse to reasoning processes. Figures 6.5 to 6.7 exemplify the tasks presented to participants. All task sheets were computer generated and contained accompanying instructions, as shown in Appendix A.

If $\neg(colour! = white)$ what can you say about $shape!$ in operation $GetShapeColour$?

$$
\begin{array}{|l}
\hline
\mathit{GetShapeColour} \\
\mathit{shape!} : \mathit{SHAPE} \\
\mathit{colour!} : \mathit{COLOUR} \\
\hline
\mathit{colour!} = \mathit{white} \lor \neg(\mathit{shape!} = \mathit{rectangle}) \\
\hline
\end{array}
$$

(a) $shape \neq rectangle$

(b) $shape = circle$

(c) $shape \neq circle$

(d) Nothing

Figure 6.5: Abstract DD-AN-1 task

What can you say about the effect of operation $HireVideo$ on its after-state variables?

$$
\begin{array}{|l}
\hline
\mathit{HireVideo} \\
\Delta \mathit{VideoShop} \\
\hline
\neg(\mathit{film'} \in \mathit{FilmsOnShelf}) \land \mathit{report'} = \mathit{OnLoan} \\
\hline
\end{array}
$$

(a) $film' \in FilmsOnShelf \land report' = OnLoan$

(b) $\neg(report' = OnLoan)$

(c) $\neg(film' \in FilmsOnShelf)$

(d) Nothing

Figure 6.6: Thematic CE-NA-2 task

If $line\_status' = Unconnected$ after the execution of operation $ConnectNewUser$, what can you say about $user'$?

$$
\begin{array}{|l}
\hline
\mathit{ConnectNewUser} \\
\Delta \mathit{TelephoneNetwork} \\
\hline
\neg(\mathit{user'} \notin \mathit{ConnectedUsers} \land \mathit{line\_status'} = \mathit{Unconnected}) \\
\hline
\end{array}
$$

(a) $\neg(user' \notin ListedUsers)$

(b) $\neg(user' \notin ConnectedUsers)$

(c) $user' \notin ConnectedUsers$

(d) Nothing

Figure 6.7: Thematic DMCDE-2 task

6.3.4 Procedure

Before starting the experiment, all participants were asked to provide brief biographical details including: occupation, age, organisation, course, number of years' Z experience, a list of other formal notations known, a personal rating of their Z expertise (novice, proficient or expert), and a description of any systems of formal logic studied beforehand. Participants were then shown the following instructions.

> "In each of the tasks that follow, you will be shown a Z operational schema and a description of the operation's execution. You will be asked to determine which one of four given statements follow from the information given. Please circle the letter of your choice. You will also be asked to give a confidence rating, which should indicate how far you believe your answer to be correct. Please complete all tasks to the best of your ability, without reference to textbooks. The experiment should take no longer than 30 minutes to complete."

The AFL group was also told that they may assume the global Z type definitions shown in Figure 6.8, which clarified the range of values that could be assigned to variables.

$$\begin{array}{lll} \mathit{SHAPE} & ::= & \mathit{square} \mid \mathit{circle} \mid \mathit{triangle} \mid \mathit{rectangle} \\ \mathit{COLOUR} & ::= & \mathit{red} \mid \mathit{green} \mid \mathit{blue} \mid \mathit{white} \end{array}$$

Figure 6.8: Global Z type definitions

For each task participants were shown: two Z predicates representing the premisses of a disjunctive or conjunctive inference, three Z predicates representing possible determinate conclusions, labelled "(a)" to "(c)", and a fourth predicate, labelled "(d)", representing a possible indeterminate conclusion. Participants were asked to select the one conclusion that followed from the given premisses by circling the appropriate letter, then to give a rating of the extent to which they believed their response was correct by ticking one of the corresponding boxes shown below. These boxes were coded 1, 2 and 3 respectively for the purposes of analysis. Task sheets were distributed to participants and completed anonymously then mailed back to the experimenter. All participants were tested on an individual basis.

Confidence rating: □ Not confident □ Guess □ Confident

6.4 Results

Group Correctness

Analyses of variance revealed no significant effect of group type on correctness with the disjunctive inferences, a significant effect of group type on correctness with the conjunctive inferences approaching significance ($F_{(1,38)} = 3.85, p = 0.06$), and a significant effect of group type on correctness with the De Morgan's based inferences ($F_{(1,38)} = 5.08, p = 0.03$). A comparison of group correctness revealed rank orders as follows: TFL ($\bar{x}$ = 88%) < AFL ($\bar{x}$ = 93%) for disjunctive inferences, TFL ($\bar{x}$ = 89%) < AFL ($\bar{x}$ = 98%) for conjunctive inferences, and TFL ($\bar{x}$ = 78%) < AFL ($\bar{x}$ = 96%) for De Morgan's based inferences. A comparison of each group's susceptibility to logical fallacy revealed rank orders as follows: AFL ($\bar{x}$ = 1%) < TFL ($\bar{x}$ = 6%) for the disjunctive inferences, and AFL ($\bar{x}$ = 6%) < TFL ($\bar{x}$ = 10%) for the conjunctive inferences.

Inference Type

The frequencies of correct disjunctive affirmation and denial inferences drawn by the two linguistic groups are shown in Figure 6.9 respectively. An analysis of variance failed to reveal any significant effects of disjunctive inference type on correctness. A comparison of participants' correctness for the disjunctive inference types revealed a rank order as follows: DA ($\bar{x}$ = 89%) < DD ($\bar{x}$ = 92%).

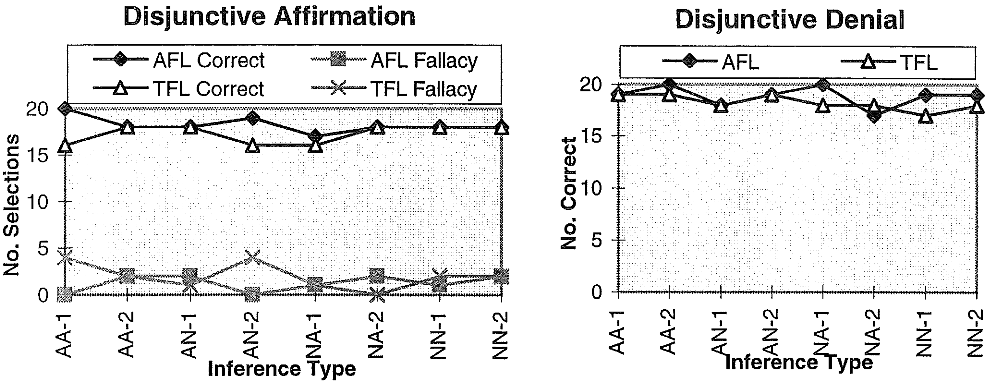

Figure 6.9: Frequencies of disjunctive inferences endorsed ($n = 20$)

The rates of correct conjunctive elimination and introduction inferences drawn by the two linguistic groups are shown in Figure 6.10. An analysis of variance reveal no significant effect of conjunctive inference type on participants' correctness. A comparison of participants' correctness for the conjunctive inference types revealed a rank order as follows: CE ($\bar{x} = 91\%$) $<$ CI ($\bar{x} = 93\%$).

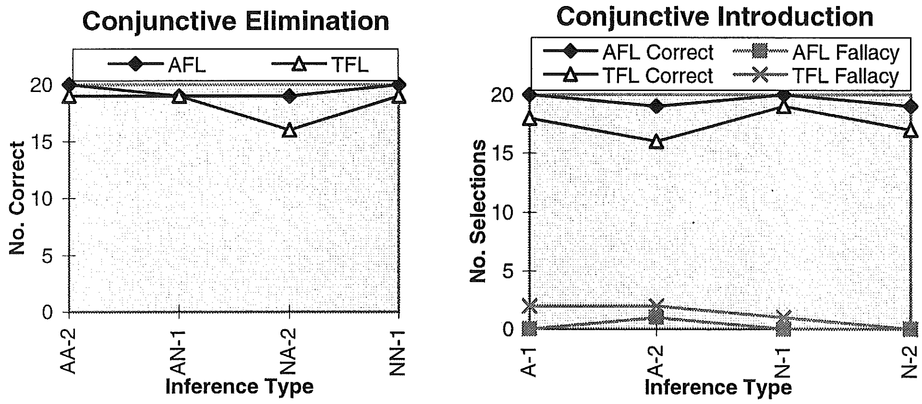

Figure 6.10: Frequencies of conjunctive inferences endorsed ($n = 20$)

The rates of correct De Morgan's based inferences endorsed by the two groups are shown in Figure 6.11. An analysis of variance revealed a significant effect of inference type on correctness for the De Morgan's based tasks ($F_{(1,38)} = 9.04, p = 0.05$). A comparison of participants' correctness for the De Morgan's based tasks revealed a rank order as follows: DMDCE ($\bar{x} = 91\%$) $<$ DMCDE ($\bar{x} = 93\%$).

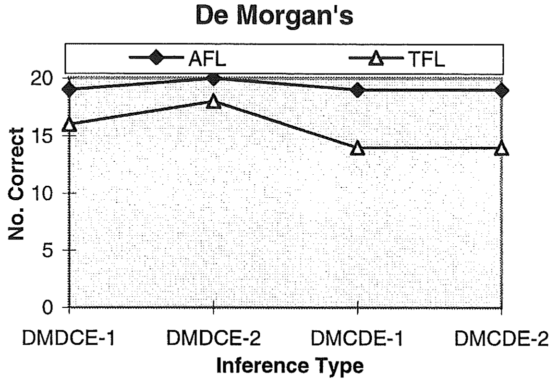

Figure 6.11: Frequencies of De Morgan's inferences endorsed ($n = 20$)

Polarity Type

Analyses of variance revealed no significant effects of premiss polarity on participants' correctness or their proneness to fallacies for either the disjunctive or conjunctive inferences. Further analyses of variance revealed significant interactions between polarity and inference type in participants' conjunctive reasoning performance ($F_{(3,114)} = 2.93, p = 0.04$), and between polarity and group type in participants' susceptibility to the disjunctive affirmation fallacies ($F_{(3,114)} = 3.17, p = 0.03$). A comparison of participants' correctness across premiss polarity types revealed rank orders as follows: AN ($\bar{x}$ = 90%) < NA ($\bar{x}$ = 91%) = NN ($\bar{x}$ = 91%) < AA ($\bar{x}$ = 96%) for the denial inferences, NA ($\bar{x}$ = 86%) < AN ($\bar{x}$ = 90%) = NN ($\bar{x}$ = 90%) = AA ($\bar{x}$ = 90%) for the affirmation inferences, AN ($\bar{x}$ = 85%) < NN ($\bar{x}$ = 90%) < NA ($\bar{x}$ = 95%) = AA ($\bar{x}$ = 95%) for the elimination inferences, and N ($\bar{x}$ = 90%) < A ($\bar{x}$ = 96%) for the introduction inferences.

Term Ordering

An analysis of variance revealed that the order of logical terms denied, affirmed, introduced and eliminated had no significant effects on participants' reasoning performance. A comparison of participants' correctness across term orderings revealed rank orders as follows: first ($\bar{x}$ = 91%) < second ($\bar{x}$ = 93%) for denial inferences, first ($\bar{x}$ = 89%) = second ($\bar{x}$ = 89%) for affirmation inferences, second ($\bar{x}$ = 88%) < first ($\bar{x}$ = 95%) for elimination inferences, first ($\bar{x}$ = 93%) < second ($\bar{x}$ = 94%) for introduction inferences, first ($\bar{x}$ = 85%) < second ($\bar{x}$ = 89%) for De Morgan's based inferences.

Experience and Expertise

A linear regression analysis revealed a significant correlation between participants' ratings of expertise and their levels of correctness (Adjusted $R^2 = 0.10$, $F_{(1,39)} = 5.08, p = 0.03$), a correlation between participants' years of Z experience and their levels of correctness approaching significance (Adjusted $R^2 = 0.06, F_{(1,39)} = 3.55, p = 0.07$), but no correlation between participants' ages and their levels of correctness. Taken together, these results suggest that it was participants' increased usage and familiarity with the Z notation, rather than their increased ages, that was responsible for their high levels of correctness.

Confidence Ratings

Figure 6.12 suggests that most participants were highly confident in the correctness of their responses for all task types. Analyses of variance revealed no significant effects of the main variables on participants' confidence ratings. Casual inspection of the results, however, suggests that the abstract group was consistently more confident across all inference types, polarities and term orderings. A comparison of mean confidence ratings for group type revealed a rank order as follows: TFL ($\bar{x}$ = 2.84) < AFL ($\bar{x}$ = 2.95). A comparison of mean confidence ratings for inference type revealed a rank order as follows: CI ($\bar{x}$ = 2.86) < DE-A ($\bar{x}$ = 2.89) = DMDCE ($\bar{x}$ = 2.89) < DMCDE ($\bar{x}$ = 2.90) < CE ($\bar{x}$ = 2.91) < DE-D ($\bar{x}$ = 2.92).

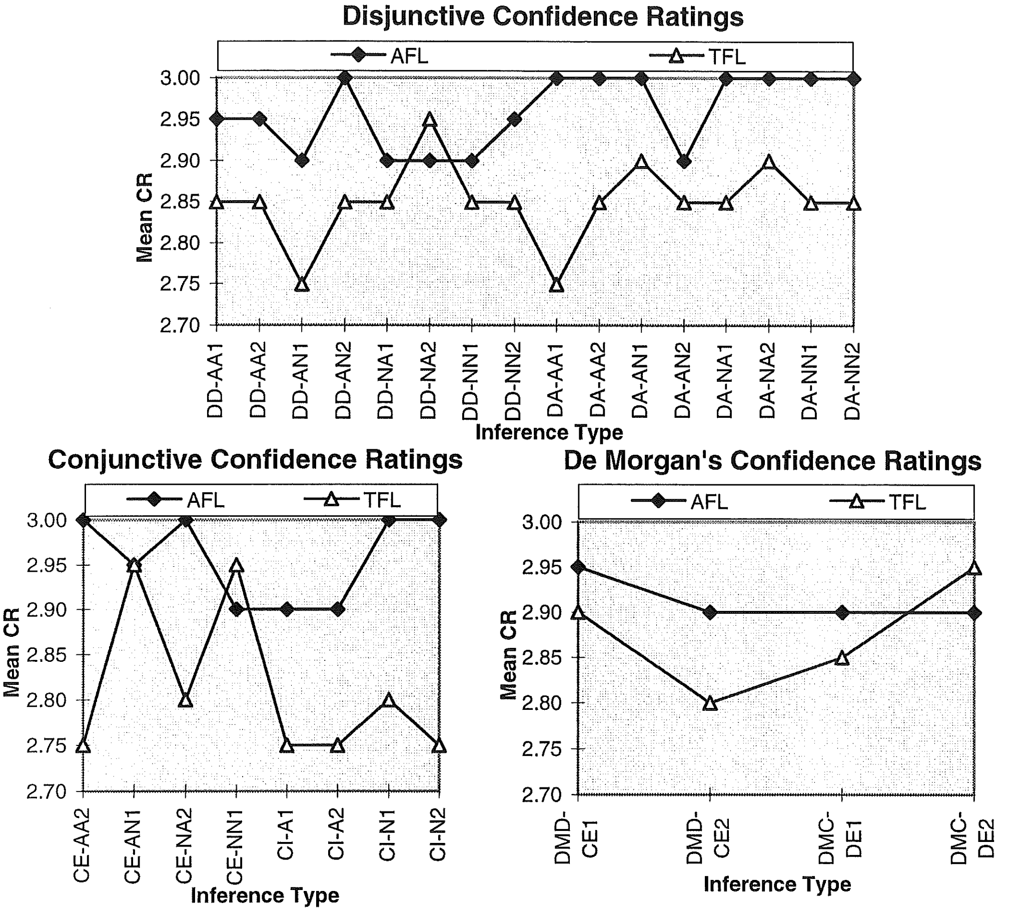

Figure 6.12: Confidence ratings by inference type ($1 \leqslant CR \leqslant 3$)

6.5 Discussion

Disjunctive Inferences

Given a disjunction, "$p \vee q$", and the assertion of one of its disjuncts, "p", an exclusive interpretation allows us to deduce the falsity of the other, "$\neg q$". Under an inclusive interpretation, however, we cannot logically infer anything about the truth value of "q" because both disjuncts might be true. Although most participants responded correctly to the disjunctive affirmation tasks, of the remainder who erred nearly all gave responses which were consistent with exclusive interpretations.

> "The point is that the ordinary language connective *or* does not always correspond neatly to the inclusive and exclusive disjunctions of logic, and there are times when it seems to defy translation into any truth table pattern" (Newstead et al., 1984).

With regard to Newstead and Griggs' (1983a) hypothesis that disjunctive reasoning should be better in those languages where disjunctives are unambiguous, it is hard to imagine a disjunctive operator that could be more precisely defined than one whose grammar is specified purely in terms of formal logic. Inspection of the results suggests that 6% of participants in the abstract group, and 10% in the thematic group, adopted exclusive interpretations of the inclusive Z operator, in spite of its formally defined semantics. Roberge (1976b; 1977; 1978) reports fallacious response rates for the affirmation task of around 20%, but marked signs of improvement on those affirmation tasks where the disjunctives are explicitly marked as being inclusive or exclusive. In a later study (Evans et al., 1993), where an error rate of 36% is reported for affirmation based tasks, the experimenters attribute participants' poor performance to a "propositional bias", that is, a general preference for drawing determinate true or false, as opposed to indeterminate, conclusions.

Given a disjunction, "$p \vee q$", and the denial of one of its disjuncts, "$\neg p$", logic allows us to deduce the truth of the other disjunct, "q". Inspection of the high rates of correctness for the disjunctive denial inferences in both linguistic groups suggests that participants were generally able to draw this inference regardless of the polarity of terms and the position of the disjunct eliminated. That 6% of responses were incorrect in the abstract group and 9% were incorrect in the thematic group, however, suggests that participants still experienced some difficulties. Although the overall error rate is higher for the thematic group, the presence of meaningful material did not appear to influence reasoning performance to the same extent as that observed in natural language based studies involving the denial task. These report error rates ranging from 20% (Evans et al., 1993; Roberge, 1976b; 1977; 1978) up to 88% (Johnson-Laird and Tridgell, 1972).

Conjunctive Inferences

> "The conjunction error demonstrates with exceptional clarity the contrast between the extensional logic that underlies most formal conceptions of probability and the natural assessments that govern many judgements and beliefs" (Tversky and Kahneman, 1983, p.310).

According to the laws of probability, the likelihood of a proposition "*p*" cannot exceed the likelihood of a conjunction "*p and q*". The results suggest that most participants correctly gave indeterminate responses to the conjunctive introduction tasks. The fact that the thematic group committed the fallacy more often might be explained by the possibility that a conjunction of two realistic terms sharing a plausible relation is more likely to be endorsed than a conjunction of two abstract terms sharing an arbitrary relation. The tendency to commit this fallacy was notably strongest on those tasks in which a causal, rather than arbitrary, relationship appeared to exist between the two conjuncts. In "$current\_loc' = target\_loc? \wedge mission' = Success$" and "$applicant? \notin banned \wedge members' = members \cup \{applicant?\}$", for example, presupposition of the truth of the first conjunct appears to be necessary for an adequate understanding of the second. Both of these fallacious conclusions were endorsed by 10% of participants. In "$print\_queue' = \langle\rangle \wedge \neg(printer\_status' = Online)$" and "$\neg(\#register' > MaxStudents) \wedge \neg(student' \in register')$", the conjuncts do not appear as strongly interdependent, which may account for their lower rates of endorsement.

Given a conjunction, "*p and q*", application of the conjunctive elimination rule allows us to conclude either one of the conjuncts "*p*" or "*q*" in isolation. Judging by the high rates of correctness observed for the conjunctive elimination tasks, participants experienced no difficulty in drawing this inference despite variations in premiss polarity and the position of the conjunct eliminated. Like the high rates of correctness observed for several other tasks, this might be attributed to the expression of the tasks in formal logic and participants' prior experience with Gentzen style deductive calculi. That the thematic group was outperformed by the abstract group might be explained by the possibility that the realistic material elicited intuitive heuristics based on guesswork or association, similar to those employed in everyday reasoning, which are often incompatible with logical tasks of this nature.

De Morgan's Inferences

Based on the findings from Wason's (1959) study of human reasoning with positive and negative statements, Wales and Grieve (1969) argue that it was participants' failure to apply De Morgan's laws of logic which led to many of their downfalls. During the study participants were asked to draw inferences from natural language based statements of the form "There is not both p and q". The fact that most gave responses consistent with an interpretation of the form "There is not p and not q", rather than the logically correct "There is not p or not q", appears to have led to the 80% error rate observed during the study's first trial. These errors might be ascribed to participants' assumption that the negative applied to both of the conjuncts, rather than just the first. Given that most had no prior training in formal logic, however, it might be argued that it was unreasonable of the experimenter to presuppose participants' knowledge of the relevant De Morgan's law which would have enabled them to infer the logically correct conclusion.

In order to derive the correct responses for the De Morgan's based tasks in the present study, participants were required to apply an appropriate De Morgan's law followed by application of an appropriate elimination rule. The logical structures of these tasks are more complex than the other forms under scrutiny because they involve multiple stage inferences. If deductive competence increases with age, as developmental studies have shown (Inhelder and Piaget, 1958; Neimark and Slotnick, 1970; Paris, 1973), then one might expect experienced reasoners who have developed proficiency in logical deduction to perform well on these tasks. Linear regression analyses revealed significant correlations between participants' correctness with the De Morgan's based inferences and, first, their years of Z experience ($R = 0.33$, $F_{(1,39)} = 4.76, p = 0.04$) and, second, their levels of Z expertise ($R = 0.32$, $F_{(1,39)} = 4.47, p = 0.04$). These results do indeed suggest that the likelihood of drawing formalised De Morgan's based inferences correctly increases along with the experience and expertise of the reasoner.

Content Effects

The possible facilitatory effects of thematic content have been a source of contention in the cognitive science community for many years. The debate has come to the fore during the past three decades following the publication of results which suggest that conditional reasoning performance can be facilitated by embedding problem content in realistic material (see for example: Gilhooly and Falconer, 1974; Griggs and Cox, 1982; Johnson-Laird et al., 1972; Wason and Shapiro, 1971). It is a well supported finding that putative conclusions conforming with prior beliefs are more likely to be endorsed than those running contrary, and that the former type is often endorsed at the expense of logical necessity (Barston, 1986; Evans et al., 1983; Henle and Michael, 1956; Janis and Frick, 1943; Morgan and Morton, 1944; Wilkins, 1928). Van Duyne (1974) reports strong correlations between degree of realistic material and reasoning performance for conditional inference tasks, but no such correlations for the logically equivalent tasks expressed in terms of disjunctions and conjunctions. The results of this research suggest that a similar situation might exist in the formal domain; the expression of disjunctive and conjunctive rules in thematic material does not appear to facilitate reasoning performance in the same ways observed for conditionals. The fact that higher rates of correctness were observed for many of the logically equivalent abstract inferences suggests that meaningful material actually had an inhibitory effect on reasoning performance, despite the fact that all of the thematic inferences led to intuitively plausible conclusions. This finding is replicated in Roberge's (1977) natural language based study of reasoning with inclusive disjunctives.

Polarity Effects

It is argued that term polarity significantly affects disjunctive reasoning (Evans and Newstead, 1980; Johnson-Laird and Tridgell, 1972; Roberge, 1974; 1976a; 1976b; 1978; Wason and Johnson-Laird, 1969; 1972). The theory of "negative conclusion bias" (Evans, 1972c, 1977a; 1993a) argues that people are more inclined to endorse inferences whose conclusions are negative rather than affirmative because the former maximises the reasoner's chances of making statements which are unlikely to be disproved. Affirmative conclusions usually have particular referents, whereas negative conclusions usually have multiple referents. So cautious reasoners are liable to favour statements that make non-specific negative predictions over specific affirmative predictions, which are more likely to be refuted. The results of Roberge's (1976a) study involving exclusive disjunctives support this hypothesis insofar as significantly more inferences with negative conclusions were drawn correctly. Supposing participants had succumbed to this bias in the present study, we would expect to see significantly more negative conclusions drawn. That no such trend is born out in the results suggests that participants did not succumb to this bias. The theory of "affirmative premiss bias" (Evans, 1993a) argues that people are more inclined to endorse determinate conclusions from premisses that contain only affirmative components. The results, however, suggests that participants were equally likely to endorse determinate and indeterminate conclusions, which suggests that they avoided succumbing to this bias. The relative lack of support for these polarity biases in the present study suggest that they are primarily conditional reasoning biases which are specific to the grammatical form of the implicative rule.

Studies of deductive reasoning suggest that the complexity of a deductive task increases along with the presence of negative components (Evans, 1972c; Johnson-Laird and Tridgell, 1972; Wason, 1959). Some studies, however, suggest that disjunctive premisses containing only one negative term are more difficult than those containing two (Evans and Newstead, 1980; Roberge, 1976b; 1978). As a possible explanation for this phenomenon, it is proposed that people implicitly drop the negatives when they are present in both disjuncts because they find it easier to reason with affirmatives only. People's responses therefore give the impression that inferences involving two negatives are simpler than those containing just one because the latter cannot easily be converted into an affirmative form. The conversion of premisses in this manner, however, does not always lead to logically valid conclusions. A comparison of the rates of correctness for disjunctive premisses containing two affirmatives and those containing two negatives suggests that participants were equally adept at reasoning with premisses containing terms of the same polarity, and did not succumb to this form of illicit conversion during the present study. Inspection of the scores for the disjunctive inferences are, however, consistent with the hypothesis that reasoners experienced more difficulty when the major premiss contained mixed polarities. The rank orders of difficulty for these inferences are directly comparable with those observed by Roberge (1976a), that is, AA < NN < NA = AN. Furthermore, when the rank orders for the disjunctive inference polarities are analysed together with those for the conjunctive inferences, it seems clear that participants found it easier to reason with affirmative components.

People are notoriously prone not to interpret doubly negated statements as affirmatives in everyday communication. Given a negated description of an object, "not blue", it can be difficult for an individual to see how a further negation, "not not blue", could result in the object becoming any less blue than it already is. This frequent disinclination to convert a doubly negated proposition into an affirmative can perhaps account for many of the errors made under logical experimental conditions. Roberge (1976b) reports that reasoners experience difficulties with denial inferences wherever a negative disjunctive term in the major premiss is denied by an affirmative term in the minor premiss. The fact that the rates of correctness for this form of denial inference were not significantly different to those for the other denial inferences in the present study, however, suggests that the users of formal methods do not experience the same difficulties when reasoning with disjunctive statements. Even the low rates of correctness observed for the De Morgan's based tasks in the thematic group cannot be attributed to the presence of extra negatives because high rates of correctness were observed for the logically equivalent tasks in the abstract group. It seems more likely that the performance differential observed for these tasks is attributable to the degree of realistic content used.

People are generally more inclined to recognise that negatives deny prior positives, rather than the converse, because the function of negation in everyday language is to deny plausible conceptions (Evans, 1972a; 1972b; 1983a; 1983b; Roberge, 1978). Johnson-Laird and Tridgell (1972) report that reasoners find it easier to draw denial inferences where the minor premiss comprises an explicit denial of a term from the major premiss, as opposed to an implicit denial. Figure 6.13 illustrates the two forms of explicit and implicit denial inference presented during the study.

p or q		*not p or q*	
not p		*p*	
	Explicit denial		Implicit denial
q		*not q*	

Figure 6.13: Explicit and implicit denial inferences

In the explicit denial example, a negative is used to deny a statement that might plausibly be true. The experimenters attribute participants' difficulties with the implicit denial example to the fact that, in everyday communication, the disjunct "*p*" would not have been negated in the first place if there had been reason to suspect that it might be true. Had participants exhibited a similar tendency to draw the denial inference only from explicit denials of disjuncts in the present study, one would expect to see significantly more correct inferences for the four forms of explicit disjunctive denial than the four forms of implicit disjunctive denial. Although no such trend is evident in the observed results, the sporadic rates of correctness observed in the thematic group could be attributed to the possibility that the negation operator makes more or less plausible denials depending upon the specific meaningful term which it denies.

Matching Bias

"Matching bias" theory claims that reasoners are liable to select those conclusions which contain one or more of the terms mentioned explicitly in the given premisses (Evans, 1983a; 1983b). The disjunctive statements "I will travel by car or by train" and "I will not travel by car or by train" both appear to concern the same objects: "car" and "train". So when a reasoner is presented with a response option that fails to contain one or both of these terms, matching bias theory predicts that he or she will judge that option as irrelevant, regardless of its logical validity. Although strong signs of the bias have been reported in studies of conditional reasoning (Evans, 1972b; 1983a; 1983b; Van Duyne, 1974), none have been reported in studies of disjunctive or conjunctive reasoning (Evans and Newstead, 1980; Van Duyne, 1974). This trend is attributed to the possibility that matching bias is a special kind of associational bias which is highly sensitive to the form of the logical rule. More specifically, Pollard and Evans (1981) argue that the conditional form "if p then q" suggests a positive relationship between the two terms because one expects "q" to occur with "p", whereas the disjunctive form "p or q" suggests a negative relationship because one expects "p" to occur without "q", and vice versa. The experimenters argue that matching bias might therefore stem from people's frequent reference to positively associated events in everyday language. This theory does not, however, account for the lack of support for matching bias in studies of conjunctive reasoning where there appears to exist a clear, positive relationship between the two logical terms, "p and q".

Many of the logical responses to tasks in the present study are consistent with the predictions of matching bias because they involve bringing out one or more of the terms explicitly mentioned in the given premisses. In order to determine whether participants succumbed to the bias, however, we must focus our analysis on those tasks where significant proportions of the responses contained terms which were only implied in the given premisses. This is the case for the De Morgan's based inferences only. Supposing matching had occurred, one would expect to see a significant proportion of responses containing the negation of the correct conclusion, owing to the form of these tasks. Inspection of the results suggests that, of those participants who responded incorrectly to these tasks, nearly all endorsed indeterminate conclusions. This suggests that few, if any, participants succumbed to the bias. Our study of conditional reasoning, however, revealed marked signs of matching bias, especially under abstract conditions. Taken together, these findings suggest that a similar situation in formal logic exists for that observed in natural language, albeit to a lesser extent, whereby matching is strongest on conditional inferences but relatively non-existent for disjunctive and conjunctive based inferences. These findings are therefore consistent with Evans and Newstead's (1980) claim that matching bias is a highly sensitive associational bias which is dependent on specific linguistic properties of the logical rule.

Confidence Ratings

Although no significant effects of the main variables on participants' confidence ratings were evident for any of the types of inference under scrutiny, several noteworthy correspondences between confidence and correctness are evident. Nearly 90% of the AFL group's confidence ratings are higher than or equal to the ratings for the corresponding inferences in the TFL group, which seems to correspond with the finding that the abstract group outperformed the thematic group overall. Unlike the relatively sporadic confidence ratings of the TFL group, the AFL group's ratings are consistently high across all types of inference. Assuming a correspondence between participants' confidence and correctness, this pattern in the observed results could be explained by the possibility that participants employed a fixed interpretative strategy across all of the abstract tasks, but contemplated different interpretations of the terms involved in the thematic tasks, which led to more inconsistent responses. Similar trends are reported by Staudenmayer (1975) in a comparison of conditional reasoning performance using abstract and thematic material.

6.6 Conclusions

Assuming the observed rates of correctness are a fair reflection on participants' abilities to reason about disjunctives and conjunctives in formal specifications, the present study suggests that the users of formal methods are susceptible, albeit to a lesser extent, to the same non-logical heuristics and biases exhibited in natural language based contexts. It is worthy of note that the expertise and experience of the participants played a significant role in determining whether the correct inferences were drawn. Whilst reasoners responded largely in accordance with the dictates of logic, however, there were still many occasions on which they succumbed to error. This might be considered surprising in view of participants' logical training. As a possible explanation for many of the errors committed, especially in the thematic group, it might be argued that participants analysed the formal expressions only at a syntactic level and assumed an informal semantics cued by the realistic connotations of the material, similar to those used for the corresponding propositional connectives in everyday reasoning. It is also noteworthy that manipulation of the experimental variables had less of an impact on reasoning performance than expected, particularly in light of findings from our study of conditional reasoning. This might be attributed to the relative complexities of the inferences involved or the possibility that any facilitation of non-logical heuristics is dependent upon the specific form of the logical rule and specific contextual conditions. The findings should, nevertheless, be of concern to the software engineering community because they point to possible sources of erroneous reasoning in formalised development contexts.

6.7 Summary

This chapter has reported the second main experiment which focuses on disjunctive and conjunctive reasoning in formalised contexts. It has described how design of the experiment stemmed from cognitive theories of reasoning in natural language contexts. The results suggest that the users of formal methods are often logical, but occasionally fallible, in drawing disjunctive and conjunctive inferences. Participants appeared to commit some of the same errors and biases demonstrated in natural language based versions of logically equivalent tasks. The results suggest that those factors which elicited errors, however, were specific to the form of the statements under scrutiny and the linguistic contexts in which they occurred. Having explored users' propensities for error when reasoning about formal expressions containing several common types of propositional connective, we now turn to examine users' abilities to reason about Z statements predicated by the logical quantifiers.