DEPARTMENT OF COMPUTER SCIENCE

Conditional Reasoning in Language and Logic: Transfer of Non-logical Heuristics?

R J Vinter
M J Loomes
D E Kornbrot

Technical Report No. 276

March 1997

Conditional Reasoning in Language and Logic: Transfer of Non-logical Heuristics?

Rick Vinter and Martin Loomes
Faculty of Information Sciences

Diana Kornbrot
Faculty of Health and Human Sciences

University of Hertfordshire, Hatfield, UK [1]

Historically, the use of natural language based techniques for the purpose of software specification has tended to result in the production of ambiguous or verbose system descriptions. It is thought that this imprecision is to some extent responsible for the erroneous development decisions which lead to the introduction of defects in software systems. This view is supported by previous psychological studies which suggest that people are prone to error and bias when reasoning with natural language statements containing specific logical operators. In this paper, we are concerned with the implicative, conditional operator. We describe a study which sought to test whether the reasoning of trained computer scientists is any more logical when they are reasoning about conditional arguments expressed in formal logic itself. In addition, it aimed to test the extent to which reasoning performance is affected by: the type of inference to be drawn, the degree of realistic content in problem material and the polarity of logical terms. In general, the results suggest that the same errors and biases which people exhibit on a frequent basis when reasoning with everyday discourse can, under certain circumstances, transfer across into the domain of formal reasoning. The implications of this finding are discussed in relation to the software engineering community where formal logic based notations are gaining increasing acceptance.

Introduction

Judging by the growth in popularity of formal methods during the past decade, the software engineering community in general now appears to be seriously considering a more rigorous approach to the specification of computer systems. This can perhaps be attributed to two factors. Firstly, an increase in the numbers of business- and safety-critical systems projects has led customers and independent regulatory bodies alike to demand increased assurance that computer systems will not fail at the most inopportune moments. Secondly, the use of conventional natural language based techniques for the purpose of software specification has tended to result in the production of verbose, unwieldy documents that are prone to inconsistency and imprecision (Gehani, 1986; Ince, 1992; Liskov and Berzins, 1986). Recent developments in formal technologies, along with several highly successful applications in industry, have helped to dispel many initial doubts surrounding the commercial viability of formal methods (Bowen and Hinchey, 1994; Bowen and Stavridou, 1993). Their proponents point to possible advantages over natural language based methods such as: providing a means for verifying correctness, allowing for precise and concise system descriptions, not to mention the increased levels of insight and confidence gained through the translation of system requirements into "scientifically proven" mathematical notations (Bowen, 1988; Liskov and Berzins, 1986; Potter et al., 1991; Sommerville, 1992).

[1] The authors would like to thank the following reviewers for their help in the production of this paper: Carol Britton, Richard Ralley and Jane Simpson. The authors would also like to thank all of the computing staff, students and professionals who took part in the experiment. The research reported here was supported by Grant No. J00429434043 from the Economic and Social Research Council. Requests for reprints of this paper should be sent to Rick Vinter, Faculty of Information Sciences, Hatfield, Hertfordshire, AL10 9AB, UK.

It is a well known fact that imperfect reasoning can lead to erroneous decisions. In the context of software engineering, erroneous development decisions can lead to the introduction of defects into software systems. Owing to its hazily defined methods and inherently complex nature, the potential for erroneous decisions in the software development process has always been high. If one accepts that people are more likely to endorse what is logically sanctionable when they are reasoning with terms expressed entirely in a system of logical symbolism, and one recognises that all formal technologies are based on such systems, then we begin to realise the tremendous benefits that the adoption of a formal approach could provide. So, although it is rarely propounded in such explicit terms, perhaps one of the key advantages that formal methods might hold over conventional techniques is the possibility that they will lead to fewer erroneous development decisions and, hence, to the inclusion of fewer defects in delivered systems.

The view that natural language contains potentially distracting details and ambiguities, whereas symbolic logic abstracts away extraneous details and allows reasoners to concentrate purely on the underlying form of arguments, derives from Kantian philosophy (Kant, 1781/1993). Kantian theory recognises the importance of human intuition and prior belief in everyday reasoning based on practical inference, but claims that they are seldom influential in formal logic where reasoning is guided predominantly by well defined mathematical axioms and rules of inference. This might be a false view in reality, but it is one which now appears prominent among a sizable proportion of the software community. To many, the possibility that a major source of software defects could be eliminated through the adoption of formal techniques, aside from being overwhelmingly appealing, also appears to be intuitively plausible. The Kantian view thus represents an implicit claim in favour of the formal approach, but one that remains to be tested empirically.

> "It is true that simple logical operations can be performed without the help of symbolic representation; but the structure of complicated relations cannot be seen without the aid of symbolism. The reason is that the symbolism eliminates the specific meanings of words and expresses the general structure which controls these words, allotting to them their places within comprehensive relations."
>
> Reichenbach (1966, p.3).

Over the past three decades in particular, psychology has shown that people are prone to error and bias when reasoning about natural language statements containing logical connectives such as: "if" (Braine and O'Brien, 1991), "and" (Lakoff, 1971), "or" (Newstead et al., 1984), "not" (Johnson-Laird and Tridgell, 1972), "all" and "some" (Johnson-Laird, 1977). But despite obvious syntactic differences, most formal notations contain propositional connectives and predicate quantifiers with roughly equivalent semantical definitions to these same natural language constructs: $\Rightarrow$, $\wedge$, $\vee$, $\neg$, $\forall$ and $\exists$. The present study was conducted as part of a series of experiments aimed at testing whether the same non-logical heuristics which people exhibit when reasoning about natural language transfer across into the domain of formal reasoning. Results from a preliminary study (Loomes and Vinter, 1997; Vinter et al., 1996) suggest that, under certain circumstances, people are liable to abandon logical principles in formal reasoning and are satisfied to rely primarily on prior belief and pure intuition in order to formulate plausible, rather than logically necessary, solutions. In view of the growing acceptance of formal notations, findings such as this have far-reaching implications for the software engineering community, whose historical use of natural language based notations has shown that developers who are unable to interpret or reason clearly about system specifications are more likely to make the type of erroneous development decision which lead to the production of defective systems. But if software developers are liable to employ the

same cognitive procedures when reasoning about formal specifications as they do when reasoning about natural language, then their potential for error is liable to remain the same. The present study aims to test this hypothesis in the case of the logical conditional.

Error and Bias in 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).

It is claimed that psychological studies of the conditional are 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. Previous studies have shown that conditionals are prone to incite various forms of erroneous reasoning (Braine and O'Brien, 1991; Evans, 1983a; Johnson-Laird and Tridgell, 1972; Lakoff, 1971; Newstead et al., 1984). However, all such studies have been conducted within the confines of natural language based contexts alone. That is, they have sought to examine the ways in which people reason with abstract sentences such as "If the letter is A then the number is 4", or sentences containing more realistic content such as "If the man is drinking beer then he must be over 18 years of age". Previous findings suggest that there are dominant causes for reasoners' departure from logical rules of reasoning, and that erroneous responses to conditional tasks are often attributable to the influence of non-logical biases and heuristics (Braine and O'Brien, 1991; Evans, 1993; Pollard and Evans, 1980).

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 the given premisses (Evans, 1972b; 1983a; 1983b; Evans and Lynch, 1973). For example, the conditional statements "If A then not 4" and "If not A then 4" both appear to concern the same topic: the letter "A" and the number "4". So, when an individual is presented with a response option that fails to contain one or both of these terms, the theory predicts that he or she will most likely judge that option as irrelevant, regardless of its actual logical validity. It is suggested that reasoners might adopt the matching heuristic only as a last resort, when they do not see which logical principles 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). Undoubtedly, linguistic factors play a major role in determining when matching bias is likely to occur because they direct attention to relevant or irrelevant problem information. However, the relatively high rates at which participants' responses tend to coincide with the predictions of the theory suggest that matching may be part of some higher level reasoning process which is exercised whenever the individual is unwilling to expend the mental effort necessary to perform a full logical analysis of the task.

It is theorised that most reasoners possess a general implicit bias towards positive information (Wason, 1959) which presumably derives from that convention of everyday discourse stating that the use of a negative presupposes the reason to believe a positive. After all, negatives are normally used to deny prior positives, but positives are rarely used to deny prior negatives (Evans, 1972a; 1972b; 1983a; 1983b). For example, "not A" is used to deny "A", but "A" is rarely used to deny "not A", so in both cases attention is directed towards "A". 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. Since most people learn to use conventions such as this to great effect in everyday discourse, it should be hardly surprising that they appear so reluctant to violate them under experimental conditions. Thus, in everyday reasoning, the tendency for participants to see "not p" as the converse of "p" without seeing "p" as the converse of "not p" should be quite understandable. Similarly, there is an almost universal tendency in normal 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. This frequent disinclination in normal discourse to convert a doubly negated proposition into an affirmative can perhaps account for a number of the errors made in strictly logical laboratory based studies.

The influence of component polarity on reasoning performance is a well documented phenomenon in the psychological literature. Evans (1993) points to the existence of two seemingly unconscious biases with respect to conditional reasoning. Firstly, the theory of "negative conclusion bias" claims that an individual is more inclined to endorse an inference whose conclusion is negative rather than affirmative. A similar finding by Pollard and Evans (1980) was attributed to the use of an everyday heuristic bias which maximises the individual's chances of making statements that are unlikely to be disproved. Normally, affirmative conclusions 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 do predict, however, that the effects of this bias may be lessened by the presence of familiar problem content which draws the individual towards specific affirmative conclusions. Secondly, Evans' (1993) 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 may be due to the generality of its predictions. After all, determinate responses to affirmative premisses can often be explained in terms of many other, more specific, factors. In fact, findings from a later study (Evans et al., 1995) seem to oppose the predictions of affirmative premiss bias theory, but do lend support to negative conclusion bias theory.

The theory of "facilitation by realism" (Gilhooly and Falconer, 1974; Griggs and Cox, 1982; Johnson-Laird et al., 1972; Van Duyne, 1974) argues that realistic, as opposed to symbolic, task content can have a strong facilitatory effect on reasoning performance. But whether this "effect" is in fact genuine and the extent to which it supposedly improves reasoning performance has been the subject of much theoretical contention, particularly during the past three decades. The debate has yet to be fully resolved mainly because of the difficulty 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 previous convictions are more likely to be endorsed than those running contrary to prior knowledge, although such inferences are often endorsed at the expense of logical necessity (Barston, 1986; Evans et al., 1983; Henle and

Michael, 1956; Janis and Frick, 1956; Morgan and Morton, 1944; Oakhill et al., 1990; Thistlewaite, 1950; Wilkins, 1928).

Paradigms similar to that originally established by Wason (1966) have been used in studies of conditional reasoning comparing performance under abstract and thematic conditions. The "selection task" has been presented in numerous different guises containing various degrees of realistic material. These range from totally symbolic scenarios describing relations between abstract symbols through to much more realistic scenarios describing: locations and transportation methods, letters and postage rates, foods and beverages, bars and drinking clientele. In some cases, it is proposed that significantly improved performance is attributable to the use of realistic material (Gilhooly and Falconer, 1974; Griggs and Cox, 1982; Johnson-Laird et al., 1972; Pollard and Evans, 1981; Van Duyne, 1974; Wason and Shapiro, 1971), whilst other studies report little or no evidence of improved performance whatsoever (Manktelow and Evans, 1979; Reich and Ruth, 1982; Roberge and Antonak, 1979). Sometimes, even repetitions of the original studies fail to replicate the originally observed trends - for example, Griggs and Cox (1982) were unable to reproduce the findings of Wason and Shapiro (1971) and Johnson-Laird et al. (1972). Taken collectively, these studies 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 high and low 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 interpretative and reasoning strategy when evaluating abstract arguments, so that 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). However, the real problem for cognitive science arises when a reasoner responds inconsistently to different forms of abstract problem because, in these situations, it can be hard 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 previous 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, and cognitive properties of the reasoner, such as prior beliefs and reasoning expertise. Moreover, those tasks which elude seemingly inseparable associations with information stored in reasoners' semantic memories often seem to incite responses that accord with prior belief, albeit at the expense of logical necessity. The findings of those studies conducted by 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.

The first aim of the present study was to test whether the same logical errors and biases that people exhibit when reasoning about natural language based conditional statements also occur when trained computer scientists are reasoning about logically equivalent statements expressed in a formal notation. Evans' (1977) study

demonstrates that two logically equivalent statements, "if p then q" and "p only if q", are not psychologically equivalent. That is, people are liable to reason about them in ways which are significantly different. So in a similar manner, it was hypothesised that the formal expression "$p \Rightarrow q$" and the natural language statement "if p then q" are not psychologically equivalent. In view of previous studies which suggest that people experience greater or lesser difficulties in reasoning depending upon the type of inference required to be drawn (Evans, 1977; 1983a; 1983b; Evans et al., 1995; Staudenmayer, 1975; Taplin, 1971; Taplin and Staudenmayer, 1973), the second aim of the experiment was to test the ease with which reasoners are able to draw different types of logically valid inferences and their proneness to classical fallacies. In view of previous 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 is affected following manipulation of the levels of realistic content used in the formal tasks. Finally, in view of previous studies which suggest that reasoning can be impaired by the presence of negative components in premiss information (Evans, 1972c; 1977; 1993; Wason, 1959; Johnson-Laird and Tridgell, 1972), the fourth aim of the present study was to test whether any significant differences are observable in reasoning performance following the systematic variation of affirmative and negative terms in the task information presented to participants.

EXPERIMENT

Method

Participants. A total of sixty computer scientists and computing professionals from various academic institutions and industrial organisations volunteered to participate in the experiment. These were divided into three linguistic groups: Abstract Natural Language (ANL), Abstract Formal Logic (AFL) and Thematic Formal Logic (TFL). These groups were counter balanced, firstly, according to participants' length of Z experience and, secondly, according to their personal ratings of expertise. The ANL group comprised 16 students, 2 academic staff members, 1 software professional and 1 other. Their mean age was 27.00 years ($s = 9.41$) and 13 had studied at least one 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 academic staff and 2 software professionals. Their mean age was 32.75 years ($s = 13.06$) and 13 had studied at least one system of formal logic beforehand. Their mean level of Z experience was 3.82 years ($s = 4.09$). According to participants' personal ratings, the AFL 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 at least one 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.

Design. The study had a three factor design. The first, between groups, factor was the language in which the problem material was presented: ANL, AFL and TFL. The second, repeated measures, factor was the type of inference to be drawn and had four levels: modus ponens (MP), modus tollens (MT), denial of the antecedent (DA) and affirmation of the consequent (AC). It should be noted that, whilst MP and MT are valid forms of deductive inference, DA and AC are fallacious forms in which nothing can be deduced logically. The third, repeated measures, factor was the polarity of the logical premisses shown 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 experimental tasks were devised and presented to the three experimental groups. The underlying logical forms of the tasks corresponded to the sixteen possible types shown in Table 1; that is, four different types of inference each with four different types of premiss polarity. Owing to its popularity amongst the software engineering community and its mathematical foundations in "standard" logic (i.e. propositional logic with first-order predicate extensions), the Z notation (Spivey, 1992) was chosen as a means for expressing the formal logic based tasks. In view of the participants' backgrounds, English was the natural choice for expressing the natural language based tasks.

TABLE 1
Logical forms of the experimental 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 $\neg q$, $\neg q$, $\therefore \neg p$

Task and Materials. Each task required the participant to draw a logical inference from the premisses of a conditional syllogism and then to select the one conclusion that followed logically from four given options. For the ANL group, these premisses and conclusions were expressed in natural English. For the AFL and TFL groups, they were expressed in the form of Z predicate expressions. For both abstract groups, the linguistic content of the tasks was confined to describing relations between colours and shapes so as to minimise the possible interference of real-world content. The AFL group was told that they may assume the existence of the global Z definitions shown in Appendix A. For the thematic group, a series of sixteen different scenarios were designed to elicit associations with participants' prior beliefs and intuitions relating to realistic computing applications such as: a library database system, a flight reservation system, a missile guidance system, a video lending system, and a vending machine operation. In order to minimise any potential conflict between logic and prior beliefs about the realistic problem material presented to the TFL group, all tasks were designed so as to lead to believable conclusions; that is, to plausible conceptions of the corresponding real-world applications. Having said this, the thematic tasks were designed so that the correct answers could not simply be "read off" from memory. This was achieved by inserting more than one plausible conclusion in the available response options. An example of the materials presented to each of the experimental groups is shown in Appendix B of this paper. Following each task, participants were asked to give a subjective rating of the extent to which they believed their response was correct. This was achieved by ticking an appropriate box, as shown below. All task sheets were computer generated.

Confidence rating: ☐ Not confident ☐ Guess ☐ Confident

Procedure. The ANL group was asked to provide brief biographical details including: occupation, age, organisation, course, division, year of study, and details of any system of mathematical logic studied beforehand. The AFL and TFL groups were asked to provide the following additional information: number of years' Z experience, a list of all formal notations studied, and a subjective rating of their Z expertise (novice, proficient or expert). The two formal logic based groups 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."

Precisely the same instructions was shown to the ANL group with the exception of only the first sentence, which was replaced by the following: "In each of the tasks that follow, you will be shown a description of a colours and shapes scenario." Task sheets were distributed to participants and completed anonymously then mailed back to the experimenter. All participants were tested on an individual basis.

Results

TABLE 2
Frequencies of correct and fallacious inferences endorsed in the three groups

	Group ANL ($n = 20$)				Group AFL ($n = 20$)				Group TFL ($n = 20$)			
Type	AA	AN	NA	NN	AA	AN	NA	NN	AA	AN	NA	NN
MP	20	20	20	19	20	20	19	20	19	19	19	20
MT	14	18	11	9	17	16	15	15	18	19	17	16
DA	13(6)	15(4)	6(14)	11(8)	14(6)	17(1)	15(5)	17(2)	18(0)	19(0)	17(2)	19(1)
AC	10(9)	11(9)	7(13)	10(10)	13(7)	12(8)	11(9)	12(8)	17(2)	18(1)	17(0)	17(0)

Note: Numbers in parentheses denote the frequencies of fallacious inferences.

Table 2 shows the frequencies of valid and fallacious inferences endorsed by participants from the three linguistic groups. Analysis in terms of group performance suggested an overall rank order of difficulty as follows: TFL < AFL < ANL. An analysis of variance revealed a significant main effect of group type on correctness ($F_{(2,57)} = 7.19, p < 0.01$). A further analysis of variance revealed a significant main effect of group type on participants' susceptibility to the two fallacious inferences ($F_{(2,57)} = 7.33, p < 0.01$). A Scheffe post-hoc comparison of the three groups 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 ($p = 0.02$). The overall probabilities at which members of each group appeared to draw correct and fallacious inferences were calculated and are shown in Table 3.

TABLE 3
Probabilities of drawing correct and fallacious inferences ($0 \leq p \leq 1$)

Probability	Group ANL	Group AFL	Group TFL
Correct Inference	0.67 (0.41)	0.79 (0.47)	0.90 (0.30)
Fallacious Inference	0.46 (0.50)	0.29 (0.45)	0.08 (0.27)

Note: Standard deviations are shown in parentheses.

Analysis in terms of inference type suggested a rank order of difficulty as follows: MP < MT < DA < AC. With regard to correctness, an analysis of variance revealed a significant main effect of inference type ($F_{(3,171)} = 31.68, p < 0.01$), and a significant interaction between inference type and group type ($F_{(6,171)} = 4.92, p < 0.01$). A further analysis of variance revealed significant main effects of inference type on participants' susceptibility to fallacious inferences ($F_{(1,57)} = 11.79, p < 0.01$).

Analysis in terms of premiss polarity suggested a rank order of difficulty as follows: AN < AA < NN < NA. An analysis of variance revealed a significant main effect of polarity type on correctness ($F_{(3,171)} = 5.41, p < 0.01$). A further analysis of variance revealed a significant main effect of premiss polarity on participants' susceptibility to the two fallacious inferences ($F_{(3,171)} = 7.73, p < 0.01$).

TABLE 4
Mean confidence ratings for the three groups

	Group ANL ($n = 20$)				Group AFL ($n = 20$)				Group TFL ($n = 20$)			
Type	AA	AN	NA	NN	AA	AN	NA	NN	AA	AN	NA	NN
MP	3.00	2.85	3.00	2.90	2.90	2.85	2.85	2.85	2.85	2.80	2.80	2.80
MT	2.90	3.00	2.60	2.50	2.85	2.75	2.80	2.85	2.70	2.80	2.80	2.85
DA	2.80	2.85	2.95	2.75	2.90	2.75	2.80	2.80	2.85	2.80	2.80	2.80
AC	2.95	2.75	2.90	2.75	2.80	2.80	2.85	2.80	2.80	2.85	2.85	2.85

Note: All confidence ratings range from 1.00 (not confident) to 3.00 (confident).

Table 4 reveals that participants generally declared high levels of confidence in the correctness of their responses for all inference types and premiss polarities across all three linguistic groups. Although few clear trends are evident in the observed confidence ratings, perhaps owing to participants' overconfidence, it is interesting to note isolated correlations between participants' levels of confidence and correctness. For example, all participants expressed a maximum confidence rating for the ANL group's MP-AA inference which correlates with the universal success rate observed for this task. Whereas, a relatively low mean confidence rating was observed for the ANL group's MT-NN inference which correlates with the fact that less than half of the participants gave correct responses for this task. An analysis of variance revealed a significant main effect of inference type on participants' confidence ($F_{(3,171)} = 2.90, p = 0.04$), and a main effect of polarity type on confidence approaching significance ($F_{(3,171)} = 2.55, p = 0.06$). A Scheffe post-hoc comparison of the three groups revealed no significant effects of any of the three linguistic group types on participants' confidence.

Numerous cross-cultural studies attribute improved performance in specific cognitive tasks to increased language familiarity (Brown et al., 1980; Kiyak, 1982; Okonji, 1971). Similarly, it was hypothesised that those participants in the present study with relatively high levels of experience and expertise with the Z notation would reason more accurately than those without. This hypothesis was confirmed in several analyses by linear regression, which revealed significant correlations between the formal participants' length of Z experience and their level of correctness ($R = 0.41, F_{(1,39)} = 7.82, p < 0.01$), and between their Z expertise rating and level of correctness ($R = 0.46, F_{(1,39)} = 9.40, p = 0.04$).

Discussion

In general, the results suggest that participants were frequently prone to error in both abstract groups, but signs of significantly improved reasoning performance were noticeable in the formal thematic group. With regard to the two valid inference types, the results suggest that few participants experienced any difficulty whatsoever in drawing the MP inference, with near ceiling levels observed for all combinations of premiss polarity across all three groups. This finding is supported by numerous studies of conditional reasoning in natural language contexts which also suggest that reasoners rarely err when drawing MP inferences irrespective of the polarities of terms (Evans, 1972b; 1977; Evans et al., 1995; Taplin, 1971) and the realistic content of the problem material involved (Griggs and Cox, 1982; Manktelow and Evans, 1979; Mason et al., 1975; Pollard and Evans, 1987). This trend might be explained by the possibility that reasoners are accustomed to drawing MP inferences on a frequent basis in everyday life and that this additional experience accounts for their improved performances in laboratory based studies. The slightly lower rates of correct MT inferences is supported by previous studies reporting significantly lower rates of correct MT rather than MP inferences (Evans, 1977; Taplin, 1971; Taplin and Staudenmayer, 1973). A post-hoc means comparison for the MP and MT inferences revealed this difference in performance to be significant ($F_{(1)} = 19.98, p < 0.01$). The variability of correct MT inferences was particularly noticeable in the ANL group where up to 55% failed to draw the correct response. With regard to the two fallacious inference types, the results suggest that most participants were able to deduce that nothing follows in response to the premisses of DA arguments, however, the two abstract groups in particular seemed prone to deny the antecedent where up to 70% succumbed to the fallacy. The results suggest that participants experienced most difficulties in drawing the AC form of inference where up to 65% succumbed to the fallacy. These high rates of observed fallacy are supported by previous studies which suggest that reasoners are especially prone to DA and AC fallacies when reasoning with abstract task material in natural language (Evans, 1972c; 1983b; Evans et al., 1995; Taplin, 1971; Taplin and Staudenmayer, 1973).

If it is true that 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 presented in its logical premisses. However, analyses of the rates of correct MT, DA and AC inferences revealed no such linear relation in the results observed in the present study. Instead, the results for all three groups suggest much more complex, non-linear relations between reasoning performance, inference type and premiss polarity. In particular, the rates of correct MT inferences endorsed by the ANL group seem to support a previous finding that negating the antecedent of the conditional premiss renders MT inferences much more difficult, whereas negating the consequent alone has little effect, for natural language based reasoners (Evans, 1972b). In addition, the results suggest that participants were most likely to succumb to both DA and AC fallacies in the natural language group where the antecedent of the conditional

premiss was negative and the consequent was affirmative. This finding appears to contradict the predictions of Evans' (1993) theory of affirmative premiss bias because, in this study, it appears that participants were more inclined to draw determinate conclusions from negative premisses.

It is argued that reasoners who correctly evaluate MT arguments but do not rate them as more difficult than MP arguments are less likely to share an awareness of the asymmetry of the logical conditional, whereas participants who evaluate MT inferences correctly and rate them as more difficult than MP inferences are more likely to appreciate the uni-directional nature of the conditional rule (O'Brien and Overton, 1982). An analysis of participants' confidence ratings for the MP and MT inferences in the ANL group revealed a clear correlation between confidence and correctness; the higher the confidence rating, the higher the score. However, the strength of this correlation deteriorates as we inspect participants' mean confidence ratings for the same inferences in the AFL group, and disappears altogether when we inspect the TFL group. On this basis, it might be argued 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. However, inspection of the frequencies of correct responses for the DA and AC inferences prevents us from making any such claim. One would expect reasoners who do not appreciate the asymmetry of the conditional to give responses that conform with a biconditional interpretation when they are presented with DA and AC inferences. This hypothesis is well supported by the results of the present study, where a biconditional interpretation of the conditional appears to have led many participants into logical fallacies, particularly in the two abstract groups. So, perhaps the most that can be argued in this respect is that participants gained an appreciation of the relative complexity of the different types of inference, but they failed to recognise that the asymmetrical nature of the conditional was partially responsible for this.

The relatively high rates of fallacious DA and AC inferences endorsed by the ANL group suggest that many of these participants adopted a more symmetrical, biconditional interpretation of the conditional rules than was adopted by the formal logic groups. This might be attributed to two factors. Firstly, it is postulated that the clearly uni-directional physical appearance of the arrow in the formal operator "$\Rightarrow$" leads reasoners away from biconditional interpretations of conditional rules and into an increased appreciation of the asymmetrical nature of conditional expressions. Secondly, biconditional interpretations of "if ... then" statements are often adopted in everyday discourse where they lead reasoners to conclusions which are both pragmatically sanctionable and sufficient for their purpose (Evans et al., 1993). For example, from the statements "If the switch is up then the light is on" and "The light is on", one might be inclined to infer that "The switch is up", despite the fact that this inference does not follow logically. However, attempts to apply the same principles that govern everyday discourse to tasks governed purely by rules of logic inevitably lead to error. Although the DA and AC inferences endorsed by the ANL and AFL groups were unlikely to have been influenced largely by association with pragmatic beliefs, owing to the abstract nature of the material involved, it is postulated that this still did not prevent participants from applying pragmatic heuristics, including biconditional interpretations of the conditional rules.

That participants were frequently observed to err, despite the fact that their confidence ratings approached ceiling levels throughout, suggests that participants were generally overconfident in the correctness of their responses. This might be attributed to two factors. Firstly, the seemingly trivial nature of some tasks might have instilled a false sense of security which participants retained even when completing the more complex tasks. Secondly, that so many participants seemed reluctant to admit to any form of doubt can perhaps be attributed 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 natural language 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 real engineering problems in industry. So, possibly for political reasons, one might have expected members of the two formal groups in particular to have proclaimed high levels of confidence throughout irrespective of the difficulty of the tasks presented.

Signs of Classical Error and Bias

Matching bias theory 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 the given premisses (Evans, 1972b; 1983a; 1983b). In order to determine whether participants succumbed to matching bias in the present study, we focus on those inferences in which the correct conclusion requires making explicit one or more terms that are not explicitly stated in the given premisses: MT, DA and AC. Firstly, for a MT inference, the reasoner is required to conclude the negation of a conditional rule's antecedent in response to the negation of its consequent. In order for there to be evidence of matching bias having occurred, reasoners would have had to select responses containing the same antecedent mentioned in the conditional premiss. However, no such trend was visible in the observed results. In fact, of those participants who responded incorrectly to MT inferences, most stated that nothing followed logically from the given premisses. Secondly, when a reasoner succumbs to a DA fallacy, his or her conclusion necessarily brings out the negation of the second term explicitly stated in the conditional premiss. Similarly, when a reasoner succumbs to an AC fallacy, his or her conclusion necessarily brings out one of the terms explicitly mentioned in the conditional premiss, as is consistent with a biconditional interpretation of the conditional. In this sense, based on the high rates of DA and AC fallacies observed in the present study, it might be argued that matching bias was evident in both abstract groups.

The theory of affirmative premiss bias (Evans, 1993) claims that reasoners are more inclined to endorse determinate conclusions from premisses that do not contain any negative components. In order to determine whether participants exhibited any signs of this bias in the present study, we focus our attention on the scores for those tasks which required reasoners to deduce conclusions from premisses which were both affirmative. In the case of MP-AA, where the endorsement of a determinate conclusion leads to a logically valid inference, comparison between the scores for MP-AA and those for the other MP tasks involving negatives suggests that participants were no more likely to draw the correct inference simply because both premisses were affirmative. Similarly, in the case of AC-AA, where the endorsement of a determinate conclusion results in a logically invalid inference, comparison between the scores for AC-AA and those for the other AC tasks revealed that participants were no more likely to succumb to this fallacy simply because both premisses were affirmative. Taken together, these results suggest that participants exhibited little or no real evidence of affirmative premiss bias in the present study.

The theory of negative conclusion bias claims that an individual is more inclined to endorse an inference whose conclusion is negative rather than affirmative (Evans, 1972c; 1977; 1993). Had participants succumbed to this bias in the present study we would expect to see four trends in the observed results. Firstly, we would expect to see more correct MP inferences on conditionals with negative rather than affirmative consequents. However, as the correct responses for MP inferences approach ceiling levels in all three groups, the presence or absence of negatives in the conclusions of MP inferences seem to have had little effect on reasoning perform-

ance. Secondly, we would expect to see more correct MT inferences on 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 this trend is born out in the results, particularly in the ANL group. Although this suggests a general preference for drawing negative conclusions from MT inferences, the extent of this preference seems to depend upon the language in which the problem is presented. This result replicates the findings of an earlier study (Evans et al., 1995) which suggests that reasoners are particularly susceptible to negative conclusion bias when drawing MT inferences about problems expressed in abstract natural language. Thirdly, we would expect to see more fallacious DA inferences on 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 occur in the formal domain for DA inferences but only where the problem is devoid of realistic content. That this is also supported by an earlier natural language based study (Evans et al., 1995) suggests that DA inferences are liable to incite negative conclusion bias in both linguistic domains. Finally, we would expect to see more fallacious AC inferences on 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. Again, this suggests that non-logical bias is strongest on those tasks couched in abstract terms.

Given a proposition "not p" in normal, everyday discourse, it can sometimes be difficult for people to see how an extra negation of the proposition "not p" could result in the proposition becoming any less p than it already is. It is thus argued that people rarely recognise a double negative as an affirmative in everyday discourse (Wason, 1959; Evans, 1972a; 1972b; 1983a; 1983b). If this phenomenon had transferred into the present study, one would have expected members of the TFL group, for example, to fail to see how an additional negation of the proposition $\neg(status! = Success)$ could result in $status!$ becoming any less successful than it already was. According to the theory, participants would have been liable to disregard the extra negation in these circumstances. In order to determine whether participants experienced any difficulties with those tasks involving double negations, we confine our attention to those inferences in which it leads to a valid conclusion: MT-AN, MT-NA, MT-NN. Inspection of the results reveals that participants from all three experimental groups experienced little difficulty in reaching the correct solutions for the MT-AN inferences, but that some did experience difficulties in drawing the MT-NA and MT-NN inferences, particularly in the AFL group. Prior to the study, it had been predicted that high rates of erroneous responses to the MT-NN task might be attributable to the fact that it requires the reasoner to perform 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 relative difference in scores for the MT-NN and MT-AN inferences, it does not explain the difference in scores for the MT-NA and MT-AN inferences where only one double negation is required. However, 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 observed scores for these inferences appears to corroborate more with the theory of negative conclusion bias. Of course, aside from these isolated cases, it should be noted that participants generally showed their logical backgrounds by successfully converting double negatives into affirmatives, even in the case of DA and AC inferences where it frequently led them to endorse fallacious conclusions.

The theory of "facilitation by realism" claims that thematic, as opposed to abstract, problem content can lead to strong improvements in reasoning performance

(Gilhooly and Falconer, 1974; Griggs and Cox, 1982; Johnson-Laird et al., 1972; Van Duyne, 1974). In order to determine whether participants exhibited any signs of this bias in the present study, we conduct our analysis at the group level by comparing performance of the TFL group with that of the two abstract groups. Inspection of Table 3 suggests that the TFL group in general reasoned more logically than both abstract groups, which lends some credence to the theory that thematic problem content can facilitate reasoning performance. Previous studies suggest that reasoning tasks which elude associations with information stored in reasoners' semantic memories are more likely to incite responses that accord with prior belief, albeit sometimes at the expense of logical necessity (Barston, 1986; Evans et al., 1983; Henle and Michael, 1956; Janis and Frick, 1956; Morgan and Morton, 1944; Oakhill et al., 1990; Thistlewaite, 1950; Wilkins, 1928). This kind of "belief bias" seems most prominent when a postulated conclusion does not run contrary to existing convictions, but instead conforms with the reasoner's intuitive beliefs about the world (Dominowski, 1995). Of course, in the present study, all thematic tasks were designed so as to lead to believable conclusions and it is thought that this was a critical factor in leading to the apparently strong improvements in participants' reasoning performance.

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 for a rule of the form "if p then q" (Griggs and Cox, 1982; Manktelow and Evans, 1979; Pollard and Evans, 1981). In abstract variants of the task, most reasoners tend to succeed in selecting the p case, but nearly all fail in selecting 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 specific realistic variants of the task (Gilhooly and Falconer, 1974; Griggs and Cox, 1982; Johnson-Laird et al., 1972; Pollard and Evans, 1981; Van Duyne, 1974; Wason and Shapiro, 1971), the rates of correct MT inference are generally much higher in the present study. This difference might be explained by the fact 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 that of 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 time and mental effort that they are willing to expend on the task, this would seem to account for the differences 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.

CONCLUSIONS

In general, erroneous reasoning tends to lead to erroneous decisions, and the entire history of software development has shown us that incorrect development decisions can lead to the introduction of defects in delivered systems (Potter et al., 1991; Sommerville, 1992). Software engineering has always been a human centered activity and it is likely to remain so in the foreseeable future. Thus, the potential for human error in the software development process will remain despite the use of formal methods. Several of the specifications presented to participants in the present study were for safety-critical systems, where malfunction could lead to disastrous consequences such as loss of human life or injury - see, for example, the specification of a nuclear reactor cooling system shown in Appendix B of this paper. Although the increasing application of formal methods in business- and safety-critical systems development projects may increase overall confidence in the integrity of delivered

systems, the results of this study suggest that even trained formalists are liable to make erroneous decisions about formal specifications. We believe that the kind of question that the software community should be asking itself is: "Can we afford to risk even one developer failing to recognise that a dangerous reactor status may not necessarily be the result of a high cooling system temperature?" It is only when one appreciates the potentially catastrophic consequences that erroneous reasoning can have in those situations where formal methods are now starting to be employed that we begin to understand the importance of seeking to capture and verify the reasoning processes of software developers.

One of the main reasons for the high rates of error and bias exhibited in cognitive studies in general stems from participants' failure to recognise that the pragmatic rules and conventions which govern human communication in everyday life are different to those which govern laboratory based studies of human reasoning. Attempts to apply inappropriate rules and conventions can often lead to error. The present study suggests that people are still prone to employ pragmatic, non-logical conventions when reasoning with formal specifications containing abstract content. These conventions include: matching bias, negative conclusion bias, illicit conversion of double negatives, and biconditional interpretations of conditional rules. That the rates of correct inference in the formal thematic group were significantly higher for nearly all inference types suggests that these participants exhibited few, if any, signs of the same errors and biases. This may be a somewhat surprising finding given that one might reasonably expect realistic material to cue the reasoner into non-logical lines of thought which they have employed beforehand in everyday life. On the other hand, the results suggest that people are more likely to recognise that strictly logical rules, rather than everyday pragmatic ones, are more applicable when formal specifications are couched in meaningful, real-world terms.

That the performances of both formal logic groups were generally better than that of the natural language group appears to lend support to the formalists' claim that people are more logical when reasoning about formal logic itself than natural language. As a consequence, it might be argued that Z expressions containing the formal operator "$\Rightarrow$" are less prone to incite reasoning errors and biases than logically equivalent English statements containing the terms "if ... then". Despite their logical equivalence, there appear to be clear psychological differences in the ways that people reason about statements containing the two terms. Although the observed results suggest that formal methods might have distinct benefits over natural language based techniques for conditional based reasoning in general, there were still specific circumstances in which even the formal reasoners seemed prone to error and bias. Before the results of the present study are used as a basis for making any general claims in favour of formalisation, we should remember that formal notations contain other types of logical construct, besides conditionals, whose natural language counterparts have also been shown to incite reasoning errors and biases. In this respect, ongoing experiments at the University of Hertfordshire are investigating the abilities of people to reason with some of the other major logical properties of the Z notation including: disjunctives, conjunctives, existential and universal quantifiers. These studies aim to shed further light on the cognitive operations underlying the process of formal specification.

REFERENCES

Barston, J.L. (1986). *An investigation into belief biases in reasoning.* Unpublished PhD thesis, University of Plymouth.

Brown, C., Keats, J.A., Keats, D.M. and Seggie, I. (1980). Reasoning about implication: A comparison of Malaysian and Australian subjects. *Journal of Cross-Cultural Psychology, 11*, 4, 395-410.

Bowen, J.P. (1988). Formal specification in Z as a design and documentation tool. *Second IEE/BCS Conference, Software Engineering 88*, 164-168.

Bowen, J.P. and Hinchey, M.G. (1994). *Seven more myths of formal methods: Dispelling industrial prejudices.* Computing Laboratory Technical Report PRG-TR-7-94, Programming Research Group, Oxford University.

Bowen, J. and Stavridou, V. (1993). The industrial take-up of formal methods in safety-critical and other areas: A perspective. In J.C.P. Woodcock and P.G. Larsen (Eds.), *FME'93: Industrial Strength Formal Methods, First International Symposium of Formal Methods Europe, Odense, Denmark, 19-23 April 1993*, 183-195, London: Springer-Verlag.

Braine, M.D.S. and O'Brien, D.P. (1991). A theory of If: A lexical entry, reasoning program, and pragmatic principles. *Psychological Review, 98*, 182-203.

Cheng, P.W. and Holyoak, K.J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology, 17*, 391-416.

Dominowski, R.L. (1995). Content effects in Wason's selection task. In S.E. Newstead and J.St.B.T. Evans (Eds.), *Perspectives on Thinking and Reasoning. Essays in Honour of Peter Wason*, Hove: Erlbaum.

Evans, J.St.B.T. (1972a). On the problems of interpreting reasoning data: Logical and psychological approaches. *Cognition, 1*, 373-384.

Evans, J.St.B.T. (1972b). Interpretation and matching bias in a reasoning task. *Quarterly Journal of Experimental Psychology, 24*, 193-199.

Evans, J.St.B.T. (1972c). Reasoning with negatives. *British Journal of Psychology, 63*, 213-219.

Evans, J.St.B.T. (1977). Linguistic factors in reasoning. *Quarterly Journal of Experimental Psychology, 29*, 297-306.

Evans, J.St.B.T. (1983a). Linguistic determinants of bias in conditional reasoning. *Quarterly Journal of Experimental Psychology, 35A*, 635-644.

Evans, J.St.B.T. (1983b). Selective processes in reasoning. In J.St.B.T. Evans (Ed.), *Thinking and Reasoning. Psychological Approaches*, 135-163, London: Routledge.

Evans, J.St.B.T. (1993). The mental model theory of conditional reasoning: Critical appraisal and revision. *Cognition, 48*, 1-20.

Evans, J.St.B.T. and Lynch, J.S. (1973). Matching bias in the selection task. *British Journal of Psychology, 64*, 391-397.

Evans, J.St.B.T., Barston, J.L. and Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory and Cognition, 11*, 3, 295-306.

Evans, J.St.B.T., Clibbens, J. and Rood, B. (1995). Bias in conditional inference: Implications for mental models and mental logic. *Quarterly Journal of Experimental Psychology, 48A*, 3, 644-670.

Evans, J.St.B.T., Newstead, S.E. and Byrne, R.M.J. (1993). *Human Reasoning. The Psychology of Deduction*, Hove: Erlbaum.

Gehani, N. (1986). Specifications: Formal and informal - a case study. In N. Gehani and A.D. McGettrick (Eds.), *Software Specification Techniques*, Wokingham: Addison-Wesley.

Gilhooly, K.J. and Falconer, W.A. (1974). Concrete and abstract terms and relations in testing a rule. *Quarterly Journal of Experimental Psychology, 26*, 355-359.

Griggs, R.A. and Cox, J.R. (1982). The elusive thematic materials effect in the Wason selection task. *British Journal of Psychology, 73*, 407-420.

Henle, M. and Michael, M. (1956). The influence of attitudes on syllogistic reasoning. *Journal of Social Psychology, 44*, 115-127.

Ince, D. (1992). *An Introduction to Discrete Mathematics, Formal System Specification, and Z*, Oxford: Clarendon Press.

Janis, L. and Frick, F. (1943). The relationship between attitudes toward conclusions and errors in judging logical validity of syllogisms. *Journal of Experimental Psychology, 33*, 73-77.

Johnson-Laird, P.N. (1977). Reasoning with quantifiers. In P.N. Johnson-Laird and P.C. Wason (Eds.), *Thinking. Readings in Cognitive Science*, 129-142, Cambridge: Cambridge University Press.

Johnson-Laird, P.N. and Tridgell, J.M. (1972). When negation is easier than affirmation. *Quarterly Journal of Experimental Psychology, 24*, 87-91.

Johnson-Laird, P.N., Legrenzi, P. and Legrenzi, M.S. (1972). Reasoning and a sense of reality. *British Journal of Psychology, 63*, 395-400.

Kant, I. (1781/1993). *Immanuel Kant's Critique of Pure Reason.* N.K. Smith (Trans.), London: Macmillan.

Kiyak, H.A. (1982). Interlingual interference in naming colour words. *Journal of Cross-Cultural Psychology, 13*, 1, 125-135.

Lakoff, R. (1971). If's, and's, and but's about conjunction. In C.J. Fillmore and D.T. Langendoen (Eds.), *Studies in Linguistic Semantics*, New York: Holt, Rinehart and Winston.

Liskov, B. and Berzins, V. (1986). An appraisal of program specifications. In N. Gehani and A. McGettrick (Eds.), Wokingham: Addison-Wesley.

Loomes, M. and Vinter, R. (1997). Formal methods: No cure for faulty reasoning, In F. Redmill and T. Anderson (Eds.), *Safer Systems. Proceedings of the Fifth Safety-critical Systems Symposium, Brighton, 1997*, London: Springer-Verlag.

Manktelow, K.I. and Evans, J.St.B.T. (1979). Facilitation of reasoning by realism: Effect or non-effect? *British Journal of Psychology, 70*, 477-488.

Mason, E.J., Bramble, W.J. and Mast, T.A. (1975). Familiarity with content and conditional reasoning. *Journal of Educational Psychology, 67*, 238-242.

Morgan, J.J.B. and Morton, J.T. (1944). The distortion of syllogistic reasoning produced by personal convictions. *Journal of Social Psychology, 20*, 39-59.

Newstead, S.E., Griggs, R.A. and Chrostowski, J.J. (1984). Reasoning with realistic disjunctives. *Quarterly Journal of Experimental Psychology, 36A*, 611-627.

Oakhill, J., Garnham, A. and Johnson-Laird, P.N. (1990). Belief bias effects in syllogistic reasoning. In K.J.Gilhooly, M.T.G. Keane, R.H. Logie and G. Erdos, *Lines of Thinking: Reflections on the Psychology of Thought. Volume 1. Representation, Reasoning, Analogy and Decision Making*, 125-138, Chichester: Wiley.

O'Brien, D.P. (1993). Mental logic and human irrationality. We can put a man on the moon, so why can't we solve these logical-reasoning problems? In K.I. Manktelow and D.E. Over (Eds.), *Rationality: Psychological and Philosophical Perspectives*, 110-135, London: Routledge.

O'Brien, D.P. (1995). Finding logic in human reasoning requires looking in the right places. In S.E. Newstead and J.St.B. Evans (Eds.), *Perspectives on Thinking and Reasoning. Essays in Honour of Peter Wason*, Hove: Erlbaum.

O'Brien, D.P. and Overton, W.F. (1982). Conditional reasoning and the competence-performance issue: A developmental analysis of a training task. *Journal of Experimental Child Psychology, 34*, 274-290.

Okonji, O.M. (1971). A cross-cultural study of the effects of familiarity on classificatory behaviour. *Journal of Cross-Cultural Psychology, 2*, 1, 39-48.

Pollard, P. and Evans, J.St.B.T. (1980). The influence of logic on conditional reasoning performance. *Quarterly Journal of Experimental Psychology, 32*, 605-624.

Pollard, P. and Evans, J.St.B.T. (1981). The effect of prior beliefs in reasoning: An associational interpretation. *British Journal of Psychology, 72*, 73-82.

Pollard, P. and Evans, J.St.B.T. (1987). Content and context effects in reasoning. *American Journal of Psychology, 100*, 1, 41-60.

Popper, K.R. (1992). *The Logic of Scientific Discovery*, London: Routledge.

Potter, B., Sinclair, J. and Till, D. (1991). *An Introduction to Formal Specification and Z*, Hemel Hempstead: Prentice-Hall.

Reich, S.S. and Ruth, P. (1982). Reasoning: Verification, falsification and matching. *British Journal of Psychology, 73*, 395-405.

Reichenbach, H. (1966). *Elements of Symbolic Logic*, London: Collier-Macmillan.

Roberge, J.J. and Antonak, R.F. (1979). Effects of familiarity with content on propositional reasoning. *Journal of General Psychology, 100*, 35-41.

Sommerville, I. (1992). *Software Engineering*, Wokingham: Addison-Wesley.

Spivey, J.M. (1992). *The Z Notation: A Reference Manual*, Hemel Hempstead: Prentice-Hall.

Staudenmayer, H. (1975). Understanding conditional reasoning with meaningful propositions. In R.J. Falmagne (Ed.), *Reasoning: Representation and Process in Children and Adults*, 55-79, Hillsdale, NJ: Erlbaum.

Taplin, J.E. (1971). Reasoning with conditional sentences. *Journal of Verbal and Learning Behaviour, 10*, 219-225.

Taplin, J.E. and Staudenmayer, H. (1973). Interpretation of abstract conditional sentences in deductive reasoning. *Journal of Verbal Learning and Verbal Behaviour, 12*, 530-542.

Thistlewaite, D. (1950). Attitude and structure as factors in the distortion of reasoning. *Journal of Abnormal and Social Psychology, 45*, 442-458.

Van Duyne, P.C. (1974). Realism and linguistic complexity in reasoning. *British Journal of Psychology, 65*, 59-67.

Vinter, R.J., Loomes, M.J. and Kornbrot, D.E. (1996). *Reasoning About Formal Software Specifications: An Initial Investigation.* Division of Computer Science Technical Report No. 249, University of Hertfordshire.

Wason, P.C. (1959). The processing of positive and negative information. *Quarterly Journal of Experimental Psychology, 11*, 92-107.

Wason, P.C. (1966). Reasoning. In B.M. Foss (Ed.), *New Horizons in Psychology. Volume 1*, Reading: Penguin.

Wason, P.C. and Shapiro, D. (1971). Natural and contrived experience in a reasoning problem. *Quarterly Journal of Experimental Psychology, 23*, 63-71.

Wilkins, M.C. (1928). The effect of changed material on ability to do formal syllogistic reasoning. *Journal of Social Psychology, 24*, 149-175.

APPENDIX A

Additional Instructions

This appendix contains the additional Z definitions shown to the AFL group.

$$SHAPE ::= square \mid circle \mid triangle \mid rectangle$$
$$COLOUR ::= red \mid green \mid blue \mid white$$

ShapeAndColour

$$shape : SHAPE$$
$$colour : COLOUR$$

APPENDIX B

Examples of the Materials Used

This appendix shows the three logically equivalent versions of the AC-AN task presented to members of the three experimental groups.

Group ANL

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
(B) The shape is a circle
(C) The shape is not a circle
(D) Nothing

Group AFL

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) $shape \neq rectangle$
(B) $shape = circle$
(C) $shape \neq circle$
(D) Nothing

Group TFL

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) $coolertemp \leqslant Maxtemp$
(B) $coolertemp > Maxtemp$
(C) $coolertemp > Mintemp$
(D) Nothing