DIVISION OF COMPUTER SCIENCE

Reasoning About Formal Software Specifications: An Initial Investigation

R J Vinter, M J Loomes and D E Kornbrot

Technical Report No. 249

March 1996

Reasoning About Formal Software Specifications: An Initial Investigation

Rick Vinter, Martin Loomes and Diana Kornbrot
School of Information Sciences
(in collaboration with the School of Health and Human Sciences)
University of Hertfordshire, Hatfield, Hertfordshire, UK

Within the software engineering community, it is widely believed that formal logic based notations could hold the key to overcoming some of the classical problems associated with program specification. Over the past three decades, psychology has investigated the difficulties that people experience when reasoning about logical statements in natural language. The Human Cognition and Formal Methods project aims to test whether these studies' findings carry over into the domain of formal specification by conducting a series of specially designed experiments. The first experiment concentrated on five cognitive activities central to formal specification: reading, writing, understanding, translating and reasoning. It also investigated the ways in which designers' personalities affect their specifications and their audience's interpretations of them. Its results are significant from a psychological perspective because they suggest that many of the erroneous inferences that people make about implicit logic in natural language also occur when reasoning about explicit logic in formal specifications. Its results are also significant from a computer science perspective because they appear to contradict several popular software engineering beliefs. This paper reports these results and points to similar findings obtained from previous psychological studies.[1]

Although formal methods have been available for nearly thirty years, industry in general has been reluctant to adopt them wholeheartedly because of doubts surrounding their commercial viability. The lack of other feasible alternatives might explain designers' near exclusive preoccupation with natural language based techniques. This has tended to result in the production of large, unwieldy documents which are prone to inaccuracy, inconsistency and ambiguity. However, some notable advances in formal methods technology and some highly successful applications in industry have helped to dispel many of the software engineering community's initial doubts and formal methods are now gradually gaining widespread acceptance.

A formal specification is one that is written entirely in a "a language with an explicitly defined syntax and semantics" (Liskov and Berzins, 1979, p. 277). Underlying every formal notation can be found one or more systems of formal logic which provide a basis for describing in precise, mathematical terms the software systems to be developed. Psychology has proposed various theories to account for the apparently systematic errors that people make when reasoning about specific forms of implicit logical statement in natural language. These forms include: conditionals (Braine and O'Brien, 1991), conjunctives (Lakoff, 1971), disjunctives (Newstead et al., 1984) and negatives (Johnson-Laird and Tridgell, 1972). The Human Cognition and Formal Methods research project aims to test whether people experience similar difficulties when reasoning about explicit logic in formal specifications by conducting a series of specially designed experiments. If so, there should be a genuine reason

[1] Requests for reprints of this paper or the task sheet presented to the experiment's participants should be sent to Rick Vinter, School of Information Sciences, Hatfield, Herts., AL10 9AB, UK. Email: R.Vinter@herts.ac.uk.

The authors would like to thank the following reviewers for their many helpful comments regarding this paper: Jean Baillie, Carol Britton, Dennis Hilton, Ben Potter, Richard Potter, Richard Ralley, Jane Simpson, Mark Surman, Paul Taylor and Frédéric Vallée-Tourangeau. The authors would also like to thank all of the staff and students who took part in the first experiment. The research reported here was supported by Grant No. J00429434043 from the Economic and Social Research Council.

for concern, especially in view of the fact that formal methods are commonly used in the development of safety-critical systems. In the present study, four experimental tasks were used as a basis for investigating some of the lesser perceived problems associated with the use of formal notations for specifying computer software.

The Z notation (Spivey, 1992) was used as the grammatical framework in which to present the experiment's tasks. The reasons for specifically choosing Z are fourfold: it is popular in both academia and industry, its underlying logical calculi are representative of those used by many other notations currently in use, it is commonly seen as one of the more easily readable specification languages, and it is a well established and commercially viable technique. Perhaps owing to the current drive for a rigorous deductive proof system and an international standard development methodology for Z, not to mention the considerable interest currently being shown by industry, both Z and formal methods in general represent a highly active area of academic research and one in which much still remains to be explored.

Task 1: The Formalised Wason Selection Task

Wason's (1966) abstract selection task has become one of the most intriguing and well documented problems in the history of deductive reasoning. In the task's abstract form, participants are shown four cards which have a capital letter on one side and a number on the other. They are then shown a conditional statement: "If there is an A on one side of the card then there is a 4 on the other". The facing values of the cards show one letter and one number that match those in the rule and one letter and one number that differ, such as: A, 4, S, and 7. These correspond to the p, q, $\neg p$ and $\neg q$ cases for the conditional rule of the form *if* p *then* q. Participants are asked to decide which of the cards they would need to turn over in order to determine whether the rule is true or false. The task is one of hypothesis testing and deductive reasoning based on conditional logic, whereby participants must project the possible consequences of turning each card in order to deduce the correct combination of responses. If participants wrongly 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 answer is the A (p) and 7 ($\neg q$) cards. Typically, participants select either the p card alone or the p and q cards, and fail entirely to select the $\neg q$ card. The correct p and $\neg q$ combination was selected only around 4% of the time during Wason's early studies (Wason and Johnson-Laird, 1972, p. 182). Wason's findings suggest that people are liable to make errors of judgement when reasoning about conditional statements and that they find it much easier to make affirming *modus ponens* inferences rather than denying *modus tollens* ones.

> *Modus Ponens* (MP): Given premisses *if* p *then* q and p, we can conclude q
> *Modus Tollens* (MT): Given premisses *if* p *then* q and $\neg q$, we can conclude $\neg p$

Cognitive scientists have studied conditional reasoning in a variety of guises, from completely abstract through to realistic scenarios, in order to determine whether certain forms facilitate deductive performance (Griggs and Cox, 1982; Griggs and Jackson, 1990; Johnson-Laird and Wason, 1970; Wason and Shapiro, 1971). In general, these studies have shown that the ways in which conditional statements are presented can have a marked effect on how easily people are able to reason about them. The main reason for implementing the Wason selection task in the Z notation was, then, to determine whether presenting the task in a formal, mathematical context would affect the rate at which participants were able to make the correct selections and, simultaneously, to test whether reasoners' difficulties with conditionals carry over into the domain of formal specification.

Task 1: The Formalised Selection Task

InOut

$in? : Letter$

$out! : \mathbb{N}$

$(in? = A) \Rightarrow (out! = 4)$

(A) $in? = A$ (B) $out! = 4$

(C) $in? = S$ (D) $out! = 7$

Which inputs and outputs would enable you to test whether 'InOut' is working correctly?

The specification presented to participants in Task 1 employed a highly abstract scenario and was expressed using standard notational constructs (shown above). It was hoped that this would minimise the chances of eliciting participants' prior knowledge of existing real-world computer systems and would make the specification understandable to even the most novice of Z users. So, in the chosen implementation, an explicit logical implication of the form "$p \Rightarrow q$" was used to correspond to the implicit conditional *if* p *then* q found in the abstract version of the task. Operational pre- and post-conditions expressing simple mathematical equivalences were used to correspond to the p, q, $\neg p$, $\neg q$ cases. Like the abstract version, the correct response is to select the p and $\neg q$ cases which correspond to the input "$in? = A$" and the output "$out! = 7$", respectively.

In Wason's four-card version of the task, the conditional rule *if* p *then* q is implicit. Whereas, in the formalised version, the conditional is shown explicitly in the form of a logical implication statement $p \Rightarrow q$. Intuitively, this suggests that participants (especially those with a background in formal logic) would be more likely to recognise the type of mental inference necessary in order to deduce the correct response. It was hypothesised that their recogition of the explicit implication operator would lead a greater percentage of participants to select only those responses which follow logically than the 4% observed during Wason's early trials.

Task 2: The Specification Translation Exercises

Task 2a: The Modified Library System Specification

Library

$stock : Copy \nrightarrow Book$

$issued : Copy \nrightarrow Reader$

$shelved : \mathbb{F}\, Copy$

$readers : \mathbb{F}\, Reader$

$shelved \cup \operatorname{dom} issued = \operatorname{dom} stock$

$shelved \cap \operatorname{dom} issued = \varnothing$

$\operatorname{ran} issued \subseteq readers$

$\neg \exists\, r : readers \bullet \neg(\#(issued \rhd \{r\}) > maxloans)$

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

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

In Task 2a, participants were presented with the formal specification for a computerised library system, which was originally written by Potter et al. (1991, p. 124). However, for the purposes of this experiment, the meaning of the fourth predicate was modified in order to oppose participants' expectations (described above). Participants were then asked to translate the specification into natural English form. The Z to English translation exercise was designed to test four specific hypotheses. Firstly, whether people's prior knowledge can bias their interpretation of a formal specification. Secondly, whether significant properties of formal specifications be lost during translation to natural language form. Thirdly, whether readers attempt to simplify complex expressions in order to ease their interpretations of a formal specification. Finally, whether even seemingly trivial specifications can be understood in different ways by different people.

Previous studies of quantified syllogistic reasoning have shown how people sometimes attempt to implicitly simplify the premisses of an argument so that they are able to reason about them more clearly (Newstead, 1990; Revlis, 1975). Their relevance to this experimental task becomes apparent when one recognises that predicates of the following abstract logical forms are equivalent.

$$\neg\,\exists\, x \bullet \neg(x > y) \;\equiv\; \neg\,\exists\, x \bullet x \leq y \;\equiv\; \forall\, x \bullet \neg(x \leq y) \;\equiv\; \forall\, x \bullet x > y$$

It was hypothesised that those participants who recognised that the complex fourth predicate could be simplified in the above manner would have been more likely to offer consistent translations because the simplified forms appear to contain clearer and more intuitive mappings to meaningful natural language statements - although, the final one holds possibly the clearest intuitive mapping to a natural language translation. In the context of Task 2a, this would have involved recognising that the following expressions are logically equivalent.

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

It was recognised that participants could conceivably have used one of three different methods of translation for the first part. Firstly, participants could have offered narrative translations containing information that could reasonably be inferred from the specification and their prior knowledge of library systems. Secondly, they could have given literal translations which are basically restatements of the original predicates, but with each Z construct being replaced by its equivalent English name. Thirdly, they could have offered mixtures of both narrative and literal translations. Thus, two English translations were used as models for assessing the correctness of participants' narrative and literal responses. Although some variability in participants' use of the English language was allowed for, responses were judged incorrect if their meanings clearly did not coincide with those of their corresponding model translations.

Correct narrative response model:

(1) The library's stock comprises shelved and issued books.

(2) There are no other books in the library apart from those that are either shelved or issued.

(3) Books may be loaned to readers.

(4) The number of books that any reader borrows must exceed the maximum number allowed.

Correct literal response model:

(1) The union of shelved books and the domain of the issued relation is equal to the domain of the stocked relation.

(2) The intersection of the set of shelved books and the domain of the issued books relation is equal to the empty set.

(3) The range of the issued relation is a subset of the set of readers.

(4) There does not exist an element from the set of readers such that the number produced from the range restriction of issued books to that element is not more than the value maxloans.

In Task 2b, participants were asked to translate the following English requirements description into an appropriate form in the Z notation: "*Operation 'ComputeValue' outputs the sum of its two inputs squared.*" Despite its apparent clarity, this operation's description is actually open to multiple interpretations because it does not specify whether the two inputs must be squared before or after their addition. Therefore, responses resembling either of the two forms $a^2 + b^2$ or $(a + b)^2$ should be considered equally valid despite the fact that they would nearly always generate different solutions for the same inputs. The Z schemas (shown below) correspond to each of these two possible interpretations and were used as a guide for assessing the correctness of participants' responses. Although it was anticipated that some participants would realise that more than one that one form of solution was possible, it was hypothesised that most would resolve this dilemma with recourse to their knowledge of elementary mathematical principles. In this case, the rules of arithmetic state that multiplication precedes addition wherever there is an absence of parentheses. It was therefore predicted that most participants would offer solutions resembling the $a^2 + b^2$ form (Solution A).

Task 2b: Correct Response Models for the English to Z Translation Exercise

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

Solution A

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

Solution B

Task 3: The Specification Style Preference Exercise

In Gravell's (1991) paper, "What is a Good Formal Specification?", the author gives three different specifications of the same operation using the Z notation: one concise, one precise and one verbose. Gravell claims that, in order to communicate effectively with the majority of readers, a formal specification should be expressed with the emphasis on clarity (i.e. precision) rather than brevity (i.e. concision). His claim is based on the opinions expressed during a "straw poll" of software engineers. Hence, the main aim of the third task was to test Gravell's claim scientifically. Participants were presented with the following English description of a required software operation, "*The operation 'Toggle' exchanges the current status of a switch,*" and four different Z implementations: one concise, one verbose, one precise and one imprecise. Participants were then asked to indicate which implementation best describes the operation's behaviour and to justify their choices.

Task 3: The Four Styles of Specification

$$SWITCH ::= on \mid off$$

$$\begin{array}{|l}
\hline
\textit{Toggle} \\
s, s' : SWITCH \\
\hline
s' \neq s \\
\hline
\end{array}$$

Concise

$$\begin{array}{|l}
\hline
\textit{Toggle} \\
s, s' : SWITCH \\
\hline
(s = \mathit{off} \land s' = on) \lor \\
(s = on \land s' = \mathit{off}) \\
\hline
\end{array}$$

Verbose

$$\begin{array}{|l}
\hline
\textit{Toggle} \\
s, s' : SWITCH \\
\hline
s = on \Rightarrow s' = \mathit{off} \\
s = \mathit{off} \Rightarrow s' = on \\
\hline
\end{array}$$

Precise

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

Imprecise

The concise version of operation "Toggle" is the most succinct of the four implementations because it uses the fewest notational constructs. However, it does not specify the values of the "before" and "after" variables s and s' explicitly; these must be inferred from the enumerated values specified in the $SWITCH$ data type definition. The second version is more verbose because it contains more notational constructs than are actually needed to express the required functionality. The third alternative is the most precise because it describes the relations between the before and after variables s and s' explicitly and unambiguously. Finally, the fourth specification is imprecise because it does not define the relations between the before and after variables in sufficient detail for readers to predict the operation's post-condition with absolute certainty, given its pre-condition.

Task 4: The Formal Syllogistic Reasoning Exercises

An investigation conducted by Evans (1977) aimed to determine whether the linguistic form in which arguments were presented and the presence or absence of negative components would affect the rates at which people drew various forms of logically valid inferences. His results suggest that the rates at which participants drew successful inferences and succumbed to classical fallacies when reasoning about conditional syllogisms could be lowered or raised significantly by manipulating these two independent variables. All of Evans' materials were expressed in the form of natural language and were presented in the form of conditional syllogisms, as opposed to categorical or linear syllogisms.

In each of Tasks 4a to 4e, participants were presented with two premisses in the form of Z predicate expressions: one conditional and one equivalence. Participants were asked to specify what followed from the two premisses by selecting one from four given conclusions. The main purpose of these tasks was to determine whether presenting conditional syllogisms in a formal context would affect the rate at which people successfully drew valid inferences or the rate at which they succumbed to common reasoning fallacies. In a similar vain to Evans' study, the premisses shown to participants employed a highly abstract scenario so as to minimise the possibility of interference from thematic content. The five forms of inference that participants were required to draw are summarised below.

Task 4, Parts A to E: A Summary of the Correct Inferences

Task 4a (MP):
$(shape = circle) \Rightarrow (colour = blue)$
$shape = circle$
Therefore, $colour = blue$

Task 4b (affirmative MT):
$(shape = circle) \Rightarrow (colour = blue)$
$colour = red$
Therefore, $shape \neq circle$

Task 4c (DA fallacy avoided):
$(shape = triangle) \Rightarrow (colour = red)$
$shape = square$
Therefore, nothing follows

Task 4d (AC fallacy avoided):
$(shape = square) \Rightarrow (colour = green)$
$colour = green$
Therefore, nothing follows

Task 4e (negative MT):
$\neg(shape = circle) \Rightarrow (colour = blue)$
$colour \neq blue$
Therefore, $shape = circle$

Task 4f aimed to test participants' abilities to reason about affirmative and negative conditionals. Participants were presented with the formal specification for an operation *GetColour* and told that it contains certain inconsistencies. Their task was to deduce which parts are ill-defined by carefully following logical rules of reasoning rather than natural intuition. Although it was anticipated that most, if not all, would identify the fourth conditional as giving rise to the operation's inconsistencies, it was predicted that participants might experience some difficulties in deducing precisely which shapes are ill-defined. It was predicted that casual scrutiny of the fourth predicate would lead around one quarter of participants to interpret it as meaning "If the shape is a circle then its colour is not blue", which would consequently have led them to conclude that the circle's colour is defined inconsistently. However, careful scrutiny reveals the fourth predicate's actual meaning: "If the shape is not a circle then its colour is blue". So, the final conditional does not refer to the circle, but to all shapes except the circle; it is the square and triangle which are defined to be two different colours in the same specification simultaneously. Of course, participants would have then needed to incorporate this new information into their existing mental representations of the concepts and their relations. The specification and the task's correct model answer are shown below.

Task 4f: Formal Specification of Operation *GetColour*

GetColour

$shape : SHAPE$
$colour : COLOUR$

$(shape = circle) \Rightarrow (colour = blue)$
$(shape = triangle) \Rightarrow (colour = red)$
$(shape = square) \Rightarrow (colour = green)$
$\neg(shape = circle) \Rightarrow (colour = blue)$

Correct model answer:

circle:	Blue (from 1^{st} predicate)
triangle:	Red (from 2^{nd} predicate) and Blue (from 4^{th} predicate)
square:	Green (from 3^{rd} predicate) and Blue (from 4^{th} predicate)

EXPERIMENT

Method

Participants. Twelve computer scientists with at least a basic understanding of the Z notation volunteered to take part in the experiment: six staff and six students. Their mean age was approximately 30 years, ranging from 20 to 51 years, and their mean level of Z experience was approximately 31 months, ranging from 1 to 108 months.

Design. The study had a repeated measures design: all participants completed all experimental tasks. In total, there were ten reasoning problems as some tasks were divided into multiple parts. All task sheets were computer generated.

Materials. Task 1 presented participants with a variation on the Wason abstract selection task in the form of a Z specification and required participants to select the combination of inputs and outputs that would have enabled them to test whether the operation was working correctly. Tasks 2a and 2b required participants to translate a Z specification into English, then to translate an English requirements description into the Z notation. Task 3 presented participants with an English description of a software operation and four possible Z implementations. Participants were asked to indicate which specification best described the operation's behaviour and to justify their responses. Tasks 4a to 4e required participants to draw logical conclusions from the premisses of conditional syllogisms expressed in the form of Z predicates. In Task 4f, participants were presented with an ill-defined Z specification containing affirmative and negative conditionals and they were asked to deduce which of its parts were defined inconsistently.

Procedure. Task sheets were taken away by participants and completed anonymously then returned to the experimenter via internal mail. Participants were provided with general instructions describing how each task was to be performed and told that the experiment should take no longer than thirty minutes to complete. They were asked to provide brief biographical details including: occupation, age, course, length of Z experience and a list of Z courses previously attended.

Results

TABLE 1
Task 1: Frequencies of Response Combinations Selected During the Formal ($N = 12$) and Abstract ($N = 128$) Versions of Wason's Selection Task

Study	$p, \neg q$	p, q	p	$p, q, \neg p$	*Others*
Present study	0	7	3	1	1
Wason (1972)	5	59	42	9	13

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

Table 1 shows the frequency at which participants chose specific combinations of cards during the formal and abstract versions of Wason's selection task. Having estimated a higher rate of correct responses for the formalised version containing an explicit conditional, it was surprising to find that the actual observed success rate of 0% was even lower than Wason's 4% for an implicit conditional. Although every participant seemed able to evaluate the p case as being relevant, none appeared to see the relevance of the $\neg q$ case. Nevertheless, there are clear correlations between the results obtained from the two studies. In both studies, the frequency at which participants chose the p, q combination is highest, with selection of the p case being second highest. Also, in both cases, the rate at which participants' chose the correct $p, \neg q$ combination was very low, but notably lower in the case of the formal version.

TABLE 2
Task 2a: Frequency of Correct Translations ($N = 12$)

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

Table 2 shows the frequency of correct translations for each of the specification's predicates shown to participants in Task 2a: the Z to English translation task. Responses for the first three predicates suggest that participants were generally able to interpret their true meanings and give consistent translations. Notably, none of the participants' natural language translations of the fourth predicate succeeded in preserving the meaning of the original Z expression.

TABLE 3
Task 2a: Z to English Translation Methods ($N = 12$)

Translation Method			*Structure of Translation*	
Narrative	*Literal*	*Mixture*	*Composite*	*Holistic*
9	2	1	11	1

Table 3 shows the methods of translation given by participants in response to Task 2a. It suggests that most preferred to give narrative, rather than literal or mixed, translations of the original formal specification. In addition, it suggests that a majority preferred to give composite translations of each individual predicate in turn as opposed to holistic translations of the entire specification.

TABLE 4
Task 2b: English to Z Translation Methods ($N = 12$)

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

Table 4 summarises the methods of computation specified by participants in response to Task 2b: the English to Z translation task. Although every participant offered a solution resembling one of the two predicted forms, the table suggests that there was an equally balanced difference of opinion regarding which method was the most appropriate.

TABLE 5
Task 3: Frequency of Preferences Expressed ($N = 12$)

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

Table 5 summarises the preferences expressed by participants in Task 3: the specification style preference exercise. It indicates that participants' preferences were equally divided amongst three of the four different styles of specification presented, with one third selecting each of the concise, verbose and precise styles. None appeared to favour the imprecise style.

TABLE 6
Task 4: Frequencies of Conditional Inferences Endorsed During the Formal Logic ($N = 12$) and Natural Language ($n = 16$) Based Studies

Study	*Modus Ponens*	*Modus Tollens (simple)*	*Modus Tollens (negated)*	*Denied Antecedent*	*Affirmed Consequent*
Present study	12	4	4	1	2
Evans (1977)	16	12	2	11	12

Source: Evans (1977).

Table 6 summarises participants' responses obtained from the present formal logic and Evans' natural language based conditional syllogistic reasoning studies. Firstly, it suggests that 100% of participants successfully made the MP inference in both studies. Secondly, it suggests that there was a significant difference in the rates at which people were able to draw the simple MT inference during the two studies. Thirdly, the table suggests that both groups of participants experienced difficulties in drawing the negated MT inference. Fourthly, it suggests that an equal number of participants succeeded in drawing both the simple and more complicated forms of MT inference during the formal logic based study. Finally, the table indicates that there was a wide difference in the rates at which participants succumbed to the two reasoning fallacies in the two studies. Task 4f tested participants' abilities to reason about mixtures of affirmative and negative conditionals. Of the twelve solutions offered by participants in response to this task, ten were consistent with the correct model solution.

Discussion

The clear correlations between the results obtained from the formalised and abstract selection tasks suggest that Wason's findings do indeed carry over into the domain of formal specification and that, contrary to intuition, people do not necessarily find it easier to reason about conditional statements expressed in formal logic. Wason and Johnson-Laird's "insight theory" (1972, p. 183-188) attempts to account for the apparently systematic patterns in participants' selection rates. It proposes that people can exhibit different levels of insight into the task. Those with no insight attempt only to verify the rule and select the p, q combination. Those with insight (a) test those cards which could verify the rule to see whether they could also falsify it. If this latter criterion is not met then they are rejected. Hence, these participants select the p card alone. Participants with insight (b) examine those cards that could either verify or falsify the rule. Owing to the fact that the $\neg q$ card could falsify the rule, these participants select the p, q and $\neg q$ cards. Wason and Johnson-Laird argue that participants who possess both levels of insight are more likely to select the correct combination, that is, p and $\neg q$. Alternatively, one possible explanation for the high rate at which participants specifically chose the p, q combination in both the abstract and formal versions of the task is that many had succumbed to a form of "matching bias" (Evans, 1983). That is, participants tended to focus on those terms explicitly mentioned in the conditional rule; their selections were therefore based mainly on probablistic guesswork, rather than logical deduction.

The results from Task 2a suggest that participants were generally able to give consistent translations of the specification's first three predicates, although the fact that some 25-33% erred in each case may be a cause for some concern. However, it is participants' responses for the fourth predicate which potentially have the most far-reaching implications. Firstly, the fact that all participants' translations failed to preserve the logical meaning of the original expression illustrates how significant properties of formal specifications can be lost during the process of interpretation. Secondly, the fact that no two participants gave exactly the same translation suggests that people do indeed comprehend even seemingly trivial specifications in numerous subtly different ways. Thirdly, the form of participants' responses suggests that none were willing to expend the necessary mental effort to simplify the complex expression in order to ease their interpretations of the original text. Instead, all appear to have relied upon guesswork based on probabilities implied by the surrounding context in order to arrive at a plausible, but incorrect, meaning. Specifically, it is thought that all participants obtained the gist of the predicate's meaning by relating its key linguistic components (i.e. variable identifiers) to their own, misleading preconceptions of real world library systems. This is supported by the fact that all participants gave incorrect responses of the form "No reader may borrow more books than the maximum number of loans allowed".

Although the two forms of computation specified by participants in response to Task 2b could yield quite different results for the same input values, they might both be considered valid interpretations of the operation's ambiguous English requirements description. The equal division in participants' response types supports the commonly held view that natural language based specifications are prone to ambiguity. Furthermore, the fact that a majority of participants did not give responses resembling the predicted $a^2 + b^2$ form suggests that software designers should be careful about which aspects of their audience's prior knowledge and mathematical experience are taken for granted. But aside from the method of computation, participants' varied use of the Z notation illustrate several further issues of relevance. It is commonly thought that formal notations constrain the way in which their users write because the number of syntactical symbols and semantical rules that govern them are severely restricted in comparison with, say, those of natural languages. However, participants' responses to Task 2b showed no evidence of this. Despite the apparently limited scope of the English requirements description presented, participants nevertheless offered a wide variety of consistent solutions. In fact, no two responses were exactly the same. This illustrates an important, but often overlooked, issue with regard to the production of formal specifications: much is implied by requirements descriptions without being explicitly stated in them. Normally, these implicit requirements are implemented by designers according to their own discretions and personal styles of writing. In the case of Task 2b, participants appeared to make implicit but conscious decisions involving at least the following issues: the use of valid and invalid Z notation, the choice of meaningful identifier names, the data types assigned to each variable, the use of parentheses to clarify operator precedence, the ordering of expressions and the use of variables for storing intermediate results.

The results from Task 3 appear to contradict Gravell's claim that a majority of readers find precise specifications in particular easiest to understand. The fact that participants expressed equal preferences for the concise, verbose and precise styles implies that, whilst precision might be an extremely important quality, designers should not necessarily aim for precision alone when writing formal specifications. In fact, the responses obtained suggest strong correlations between participants' ages, levels of experience and their style preferences. Whilst the youngest and least experienced tended to choose the concise style, the oldest and most experienced appeared to prefer the precise style. But, irrespective of the preferences expressed by participants in completing a survey of this kind, one cannot necessarily generalise that these would represent the views of all readers in all circumstances. Obviously, the phrase "best describes", used in the task's prompt for responses, might have different meanings for different people. It is therefore highly probable that the criteria used to discriminate between the four specification styles will have differed between participants.

The arguments offered in justification of their choices give an indication of the factors which influenced participants' responses. Comments resembling the following forms were received from participants who chose the concise style: "It is the simplest to understand because there is less to read", "The $\neq$ relation captures the intuition that s and s' must always have opposite states" and "If there are only two members of *SWITCH* then the concise specification correctly exchanges the switch's current status." Clearly, in the case of the concise specification, most people would be able to deduce that, if the switch's setting is *on* before the operation then it must be *off* after its execution. They are able to infer this immediately and without recourse to the *SWITCH* type declaration because their prior knowledge of electronic devices tells them that switches normally have two opposite states, *on* and *off*. However, it might be argued that the degree of concision shown in the first specification would not be suitable for use in certain applications, nor indeed suitable for certain kinds of audience. It can therefore be concluded that it is possible for concision to be used with great effect in specifications, however, its effectiveness will depend largely upon readers' prior knowledge of what is being specified; what one person may consider trivial and take for granted, another may require further explicit elucidation. Comments resembling the following forms were received from participants who chose the verbose style: "It is easier to read than the concise style and less procedural than the precise style" and "It has the closest intuitive mapping to natural language." This latter comment in particular is noteworthy because it suggests that the participant implicitly attempted to convert the specification into natural language form during interpretation. Comments resembling the following forms were received from participants who chose the precise style: "Though longer than the concise style I find the explicit presentation more intuitive" and "The implication captures the notion of a toggle."

The results of the style preference task appear to have implications for the styles of writing employed by software designers. The extent to which a particular writing style coincides with a reader's natural form of interpretation might depend upon numerous independent variables which add further implicit meaning to what is said explicitly. These include: what is being specified, the surrounding context, the reader's prior knowledge and their language expertise. In theory, it might be argued that the ideal level of abstraction that one could use in a specification would take into account its audience's prior knowledge and language expertise. However, in reality, specifications are typically aimed at different readers with differing backgrounds. It is obviously impractical for designers to write several versions, each one aimed at a particular group with a certain level of expertise - for example, designers, programmers, managers and customers. This might explain why precision is rarely compromised in practice and designers employ a large degree of explicit detail so as to leave nothing to chance. Whether this principle should be applied in all cases is debatable. Considerate designers writing for novice readers[2] might aim to specify the maximum amount of detail clearly so as to leave nothing to chance, using only the simplest of a notation's constructs. In contrast, considerate designers writing

[2] The terms "expert" and "novice" are used here to distinguish between those readers who do and do not possess full knowledge of a notation's grammatical rules and constructs, respectively.

for an expert audience 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 the former case, this might enable all of a document's potential audience to comprehend, without relying upon readers' knowledge of the notation's more complex features, but at the expense of expert readers finding the document more laborious to read than others. In the latter case, designers rely entirely upon their audience's expert knowledge of the notation. In this case, there is always the danger that novice readers will not be able to comprehend parts of the specification and will accept the first plausible meaning that appeals to their intuitions, as exemplified by participants' responses for Task 2a. This may be dangerous because readers might use their inaccurate interpretations as a false basis from which to make incorrect judgements.

The results from the fourth task provide some insight into the difficulties that people experience when attempting to draw different types of inference and their proneness to classical fallacies when reasoning about conditional syllogisms expressed in formal logic. The fact that every participant appeared to make the MP inference in both the formal logic and natural language studies suggests that people are generally adept at drawing this kind of inference, irrespective of the argument's linguistic form. However, only one third of participants appeared to draw the formal logic based MT inference, whereas three quarters drew the corresponding natural language based inference. Again, this constitutes evidence that people do not necessarily find it easier to reason logically about explicit conditionals in formal logic than implicit conditionals in natural language. The fact that such a large proportion of participants failed to deduce the correct response for this part suggests that many were unaccustomed to the MT form of reasoning. This might begin to explain the same participants' poor performance on the formalised Wason selection task, where it was necessary to make an MT inference in order to evaluate the $\neg q$ case as being relevant. An equal rate of participants succeeded in drawing both the simple and negated forms of MT inference during the formal logic based study. Although these two rates are both relatively low, the fact that they are equal suggests that the participants who reasoned logically were not at all distracted by the presence or absence of the negational operator. One might theorise that those participants who deduced the correct response for Task 4b were capable of MT reasoning and would therefore have been more likely to derive the correct solution for Task 4e. The fact that three quarters of those same participants who made the simple MT inference also made the more complex, negated MT inference appeared to confirm this theory. The results from Tasks 4c and 4d suggest that only around 8% of participants denied the antecedent and 17% affirmed the consequent, as compared with 69% and 75% in Evans' natural language based study, respectively. These differences point to a possibility that people may be less prone to commit these two reasoning fallacies when reasoning about abstract, formal logic.

The fact that ten of the twelve participants responses for Task 4f matched the correct model solution suggests that participants were able to reason clearly about the specification's mixture of affirmative and negative conditionals and that they were able to adjust their mental representations of the scenario in order to accommodate the conflicting information. Most participants appeared to deduce correctly that it is the square and triangle which are defined to be two different colours in the same specification simultaneously. Only two participants erred by giving responses that did not resemble the model solution - a much lower rate than was originally predicted. Of these erroneous responses, one stated that the circle is defined as *blue* and $\neg blue$, which can perhaps be attributed to the participant's casual scrutinisation of the fourth predicate.

CONCLUSIONS

(1) *It is not necessarily easier to reason about explicit logic in a formal specification than the equivalent, implicit logic in natural language.*

Based on the fact that formal logic abstracts away unnecessary details and allows reasoners to concentrate purely upon the underlying form of arguments, it would seem plausible that it must be easier to reason logically about statements expressed in a formal system of logic with a precisely defined syntax and semantics rather than in a language littered with vague, distracting and ambiguous constructs. However, the results from two separate tasks in this experiment suggest that people do not find it easier to reason about conditional statements expressed in abstract logic than in natural language. This is supported firstly by the observation that the rate at which participants deduced the correct response for the formalised version of the Wason task was much lower than Wason's observed rate for the abstract version. Secondly, it is supported by the fact that far fewer participants deduced the correct response for the formal syllogistic reasoning task involving a simple MT inference, than the natural language based version originally presented by Evans.

(2) *Different people rarely understand a formal specification in exactly the same way.*

It is widely, though not universally, theorised that humans have an internal medium of representation in which they think and that, when they are presented with a problem expressed in a different language, they implicitly convert this into an appropriate form in their own internal "language of thought" or "inner speech" (Fodor, 1975; Huttenlocher, 1973; Vygotsky, 1986). Once a solution has been derived, this is implicitly converted back into an appropriate form in the original language. It is theorised that this internal medium resembles an abbreviated form of the person's native natural language. If this is true, then the combination and choice of natural language constructs that a person uses to represent a problem or its solution should be a fair reflection of his or her understanding of it. Based on these assumptions, the fact that no two participants gave exactly the same translation for the experiment's Z to English translation task indicates that people might in fact comprehend even simple specifications in a vast number of subtly different ways. The fact that none of the translations given for the operation's fourth predicate were consistent with the original statement suggests that people's prior knowledge can bias their interpretations of a specification and that significant properties of specifications can actually be lost during this implicit conversion process, if, indeed, this process was attempted. Furthermore, results from the experiment's two translation exercises suggest that there rarely exists a clear, unique and intuitive mapping between formal and natural language statements.

(3) *People do not always adhere to logical rules when reasoning about formal specifications, even when such rules are well known to them.*

The question of whether the human mind contains inference rules similar to those found in formal systems of logic has been vehemently debated in the past. Proponents of the theory that logic itself is a normative theory of deductive competence argue that reasoning is guided by inference rules similar to those found in formal systems of logic. These include rules for propositional connectives and quantifiers such as: *and*, *or*, *if*, *not*, *all* and *every*. In general, these theorists assume that people who have acquired a high degree of deductive competence would, under ideal conditions, always employ the correct rule at the correct time to enable them to derive the correct conclusion (Inhelder and Piaget, 1958; Rips, 1994). Opponents of the theory argue that there exists no reason to suppose that the semantics of the logical components belonging to a person's internal medium of thought reliably map onto those of the logical connectives and quantifiers in formal logic systems. Furthermore, they point to a possibility that people do not necessarily reason via the kinds of logic prescribed in standard text books (Cohen, 1981; Johnson-Laird and Byrne, 1991; Evans, 1993). But, irrespective of whether logic does fully coincide with deductive competence, Henle (1962) states that we can be certain of at least one fact: "logic unquestionably provides criteria by which the validity of reasoning may be evaluated." She suggests that logic is concerned with the "ideal"; with how people ought to think, but not necessarily with how they actually think.

In the past, psychology has pointed to convincing evidence which suggests that people frequently stray from what follows logically when reasoning about arguments expressed in natural language (Byrne, 1989; Evans, 1993). However, this begs a question which is the main concern of this paper: does people's reasoning conform more closely to the rules of formal logic when they are reasoning about formal logic itself? Intuitively, one might think so, but evidence from this experiment suggests otherwise. Firstly, none of the choices made by participants in response to the Wason selection task in Z followed logically. That is, their selected combinations of inputs and outputs would not have enabled them to deduce for absolute certainty whether the rule was true or false. Although every participant appeared to correctly evaluate the p case as being relevant, none appeared to see the relevance of the $\neg q$ case. One might postulate that, if participants had known how to perform the MT form of reasoning needed to identify the $\neg q$ case as being relevant, then they would have deduced the correct response. However, results from the formal syllogistic reasoning tasks indicate that at least one third did know how to perform both the simple and complicated forms of MT reasoning. The question of why the same number did not manage to derive the correct response for the Wason task in Z can perhaps be answered by the fact that, in practice, people's deductive performance rarely equals their deductive competence and the possibility that performance can be impaired or facilitated merely by changing the way in which a problem is presented. So, although many participants may have possessed the necessary MT rule in their mental repertoires of inference rules, none actually identified it as being applicable in this particular case.

(4) *Writing a formal specification is not an automatic process: implicit requirements are frequently implemented according to a designer's own discretion and personal style of writing.*

There is a common misconception that formal specifications are produced via some systematic means whereby designers exert little control over what is eventually written. This misconception might have sprung from the possible misnomer "formal methods", in which it is only the notation used to express a specification that is formally defined; not the process of writing it. As such, the personal characteristics of designers are assumed to have little influence during the specification construction process. Yet, in reality, rarely do two designers arrive at exactly the same specification even when this is based on the same set of trivial requirements. Differences arise because much is implied by a requirements description without being explicitly stated within it. Take, for example, the seemingly innocuous set of requirements presented during this experiment from which every participant still managed to derive a different, but nevertheless consistent, specification of the same problem. Insofar as it is prone to imprecision and ambiguity, this exercise demonstrated the inadequacy of natural language for expressing requirements specifications. Its results suggest that formal specification is far from being a completely automated process and that, despite having much more restricted grammars than natural languages, formal notations are still sufficiently powerful to allow designers to exercise their own discretions, creativity and freedom of expression. Perhaps more importantly, this task showed that the production of a formal specification is frequently guided by subjective human judgement and informal human actions, and is therefore frequently prone to human error.

(5) *Precise specifications are not necessarily the easiest to understand.*

Gravell (1991) claims that most software engineers prefer a precise, rather than concise, style of formal specification. His claim is based on the opinions expressed during a "straw poll" of engineers' preferences and therefore appears to be well-founded. Intuitively, it seems reasonable that placing the emphasis on explicit detail, rather than conciseness and abstraction, would provide for a more effective means of communication. However, participants' responses to the style preference exercise indicate an equal preference for each of the concise, verbose and precise styles. So, whilst precision might indeed be universally desirable, the findings from this experiment run contrary to Gravell's claim. The observed correlations between participants ages, experience and preferred writing styles suggest that, in practice, designers must carefully consider the type of audience that they are writing for.

(6) *Formal specifications are rarely understood solely according to what is explicitly stated within them.*

Owing to the fact that formal notations have an explicitly defined syntax and semantics, one might be forgiven for thinking that people always interpret formal specifications according to these alone. The reality is that people read interpret formal specifications according to their own methods which can give rise to interpretations that do not coincide with those prescribed by the notation's underlying formal semantics. The number of incorrect translations of the fourth predicate in the Z to English translation task constitutes evidence that people are sometimes unwilling to expend the necessary mental effort in order to derive logically valid meaning from complex expressions; this is precisely when logical reasoning gives way to intuitive guesswork. Participants appeared to postulate possible meanings and accept the most plausible based on the surrounding context and their prior knowledge of similar situations. Whilst Task 2b has demonstrated that natural language specifications can be prone to ambiguity, Task 2a has shown that, in a sense, so can formal specifications!

(7) *A quality specification is not merely a verifiably correct specification.*

Current research into formal methods appears to be progressing in two main directions: improving automated verification procedures and improving the readability of formal specifications. Although it might be a tremendous advantage for designers to be able to machine verify independently that selected properties of their specifications are both complete and consistent with regard to a customer's requirements, it is debatable whether verified correctness should be the primary aim of a software designer when one considers the role of a system specification in the overall development process. Since a specification typically forms the basis from which future design or implementation work progresses, it is important that its readers are able to understand and reason about it clearly. In the past, imprecise or unintelligible specifications have led to developers making false assumptions or incorrect decisions which have had repercussions throughout the latter stages of software projects, causing the appearance of faults or anomalies in the system design or code produced.

Although it might help to eliminate the number of previously undetected logical flaws and improve the verifier's understanding of a system, verification does not by itself increase the likelihood that a specification will communicate effectively with its intended audience. This is because a logically correct specification might still be expressed in an unreadable or unnecessarily complex manner which could potentially incite erroneous human reason. In order to communicate effectively with its audience, a specification must be readable and, in order to be readable, its designer must employ a style of writing that takes into account what is being specified and the document's intended audience. The findings from the first experiment are evidence of this. They illustrated how formal specifications are prone to misinterpretation where the message conveyed does not conform to readers' intuitive theories about the real world and how, sometimes, certain aspects of even a learned audience's prior knowledge or language expertise cannot be taken for granted. It is therefore the authors' conclusion that a quality specification must primarily be a readable one and that Gravell (1991) was nearest to the truth when he said "clarity and comprehensibility are your main aims in writing a formal specification."

REFERENCES

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.

Byrne, R.M.J. (1989). Suppressing valid inferences with conditionals. *Cognition*, *31*, 61-83.

Cohen, L.J. (1981). Can human irrationality be experimentally demonstrated? *The Behavioural and Brain Sciences*, *4*, 327-383.

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

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

Evans, J.St.B.T. (1993). Bias and Rationality. K.I. Manktelow and D.E. Over (Eds.), *Rationality: Psychological and Philosophical Perspectives.* London: Routledge.

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

Fodor, J.A., (1975). *The Language of Thought.* Cambridge, Mass: MIT Press.

Gravell, A. (1991). What is a good formal specification? In J.E. Nicholls (Ed.), *Z User Workshop, Oxford 1990. Proceedings of the Fifth Annual Z User Meeting, Oxford, 17-18 December 1990*, Workshops in Computing, Springer-Verlag.

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.

Griggs, R.A. and Jackson, S.L. (1990). Instructional effects in Wason's selection task. *British Journal of Psychology*, *81*, 197-204.

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

Huttenlocher, J. (1973). Language and Thought. In G.A. Miller (Ed.), *Communication, Language and Meaning: Psychological Perspectives*, New York: Basic Books.

Inhelder, B. and Piaget, J. (1958). *The Growth of Logical Thinking From Childhood to Adolescence: An Essay on the Construction of Formal Operational Structures.* New York: Basic Books.

Johnson-Laird, P.N. and Wason, P.C. (1970). A theoretical analysis of insight into a reasoning task. *Cognitive Psychology*, *1*, 134-138.

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. and Byrne, R.M.J. (1991). *Deduction.* Hove: Lawrence Erlbaum Associates Ltd.

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. (1979). An appraisal of program specifications. In P. Wegner (Ed.), *Research Directions in Software Technology*, Cambridge, Mass: MIT Press.

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

Newstead, S.E. (1990). Conversion in syllogistic reasoning. In K. Gilhooly, M.T.G. Keane, R. Logie and G. Erdos (Eds.), *Lines of Thinking: Reflections on the Psychology of Thought, 1*, Chichester: John Wiley.

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

Revlis, R. (1975). Two models of syllogistic inference: feature selection and conversion. *Journal of Verbal Learning and Verbal Behaviour*, *14*, 180-195.

Rips, L.J. (1994). The Psychology of Proof: Deductive Reasoning in Human Thinking. Cambridge, Mass: MIT Press.

Spivey, J.M. (1992). *The Z Notation: A Reference Manual.* Second Edition. Prentice Hall International.

Vygotsky, L.S. (1986). *Thought and Language.* Cambridge, Mass: MIT Press.

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

Wason, P.C. and Johnson-Laird, P.N. (1972). *Psychology of Reasoning: Structure and Content.* London: Batsford.

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