DIVISION OF COMPUTER SCIENCE

Measuring Human Inferential Complexity in Formal Specifications: A Predictive Model for the Z Notation

R J Vinter, M J Loomes and D E Kornbrot

Technical Report No. 304

September 1997

Measuring Human Inferential Complexity in Formal Specifications: A Predictive Model for the Z Notation

Rick Vinter and Martin Loomes
Faculty of Information Sciences

Diana Kornbrot
Faculty of Health and Human Sciences

University of Hertfordshire, Hatfield, UK [1]

Abstract

The entire history of software engineering informs us that failure to interpret or reason correctly with software specifications causes developers to make incorrect development decisions which can lead to the introduction of faults or anomalies in software systems. Most key development decisions are usually made at the early system specification stage of a software project and developers do not receive feedback on their accuracy until near its completion. Software metrics are generally aimed at the coding or testing stages of development, however, when the repercussions of erroneous work have already been incurred. This paper presents a tentative model for predicting those parts of formal specifications which are most likely to admit erroneous inferences, in order that potential sources of human error may be reduced. The empirical data populating the model was generated during a series of cognitive experiments aimed at identifying linguistic properties of the Z notation which are prone to admit non-logical reasoning errors and biases in trained users.

1 Specification as a Medium for Communication

A software specification is an abstract description which seeks to delineate the software components that a desired system will eventually comprise and is often used as a basis for evaluating the correctness of the final system. A specification is normally employed at various stages in a project life cycle including: contract negotiation, planning, design or code derivation, and program maintenance. Owing to the nature of these activities, it is likely that they will need to be understood by different audiences, each with varying degrees of expertise. A specification is typically produced as a joint venture between developer and customer, between whom it represents a form of contractual agreement (Balzer and Goldman, 1986; Imperato, 1991), but is intended to be read primarily by members of the development team for whom it also serves as a constant source of reference during the system's construction and maintenance. It is generally accepted that the specification process is an essential part of the development

[1]The authors would like to thank David Boniface, Richard Ralley and Ken Ryder for their technical assistance, Carol Britton, Ben Potter and Jane Simpson for reviewing this paper, and all of the computing staff, students and professionals who took part in the experiments. The research reported here was supported by Grant No. J00429434043 from the Economic and Social Research Council. Requests for reprints should be sent to Rick Vinter, Faculty of Engineering and Information Sciences, Hatfield, Hertfordshire, AL10 9AB, UK.

life cycle and the potential benefits that stem from a well written specification are widely documented in the computing literature (Cohen et al., 1986; Potter et al., 1996; Sommerville, 1992).

Although certain forms may be parsed, verified or animated by machines, software specifications are written predominantly for human audiences. For the members of a development team, a system specification represents an account of the software operations that must be implemented in order to meet a customer's requirements. Experience has shown that the errors which developers make when interpreting or reasoning about specifications are liable to manifest themselves in, and propagate throughout, subsequent design and code work, leading to the appearance of faults or anomalies in the system developed (Fenton and Pfleeger, 1996). The most serious defects are often not detected until integration or testing, by which time the costs of backtracking through design, code and specification in order to rectify them will have increased dramatically (Sheppard and Ince, 1993). The costs of failing to discover the defects, however, could be much higher, particularly where the system under development is business or safety critical in nature.

It is clearly important that specifications capture a customer's requirements in as complete and consistent a manner as possible. In view of their central role in project communication (Barroca and McDermid, 1992; Imperato, 1991), it seems equally important that they are written in ways which are clearly and unambiguously accessible to their intended audiences, with minimal potential for admitting erroneous human reasoning. The early stages of software projects are often critical to their success and software specifications have caused some of the most costly and intractable development problems (Cohen, 1989a; Sheppard, 1990). It would therefore seem reasonable to hypothesise that expending additional effort at the early specification stage, in particular, would substantially reduce development effort overall (Hall, 1990; Potter et al., 1996). But perhaps more importantly, it could help to reduce the numbers of latent defects that find their way into "finished" software systems.

2 Claims for Formalisation

One of the main problems historically associated with the specification process has been one of communication. Ambiguous software specifications are liable to be understood in different ways by different people, with the danger that not every member of the development team will work towards achieving the same system solution. The problem has, in general, been attributed to the use of natural language based specifications which are notoriously prone to imprecision (Gehani, 1986; Imperato, 1991; Meyer, 1985). This view is supported in the linguistic literature which reaffirms that natural language is inherently vague and ambiguous (Empson, 1965; Turner, 1986). It is argued that formal methods alleviate the communication problem in this respect by giving rise to precise and unambiguous specifications (Imperato, 1991; Jack, 1992). Based on the assumption that the semantics underlying a formal notation gives to every statement expressed in that notation a precise mathematical meaning, it

is argued that formal specifications are open to only one form of interpretation (Bowen, 1988; Liskov and Berzins, 1986; Thomas, 1993).

Human reasoning can either be formal, and based on well defined mathematical rules with explicitly recorded intermediate steps, or informal, and based on belief or intuition with implicit or undefined steps (Jacky, 1997). In the context of software specification, an instance of formal reasoning might involve checking via mathematical proof that relationships between a specification and its implementation hold, whereas an instance of informal reasoning might involve checking via subjective judgement that the system model corresponds to the customers' requirements. Based on the assumption that formal specifications are generally more concise and less ambiguous than their natural language based counterparts, it is argued that formal methods support and encourage disciplined reasoning (Rushby, 1995) and that it is easier to reason about formal specifications, even at an informal level (Thomas, 1995). The justification for this argument appears to stem from the belief that it is easier to manipulate and reason about problems expressed in mathematical logic than natural language (Ince, 1992; Lemmon, 1993). Reichenbach (1966, p.3) argues "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". Intuitively, there appears to be no reason why the argument should not generalise to specification languages whose grammatical foundations lie in these domains. This view is supported in the cognitive literature which suggests that natural language sentences containing certain common forms of logical connective are prone to evoke human reasoning errors: "if" (Braine and O'Brien, 1991), "and" (Lakoff, 1971), "or" (Newstead et al., 1984), "not" (Johnson-Laird and Tridgell, 1972), "some" and "all" (Erickson, 1978; Johnson-Laird, 1977). Providing the benefits are realised and the formalists' claims proven, the software community stands to make tremendous gains from the adoption of formal methods because they could provide a long awaited key to the development of safer systems.

3 Limits of Formalisation

In the context of software engineering, the term "method" tends to evoke images of an ordered, prescriptive set of procedures which can be followed to guide the development of software, such as stepwise refinement or object oriented design (Cooke, 1992; Hinchey and Bowen, 1995). In the context of software specification, the term "formal method" suggests a precisely defined set of procedures for developing or manipulating specifications. Yet the processes of writing, refining or verifying a formal specification rarely comprise any predefined systematic steps of action whatsoever. It is in fact only the notations in which formal specifications are written that are formally defined, owing to their precisely defined grammatical foundations (Woodcock and Loomes, 1988). The development processes associated with formal methods are guided mainly by the spontaneous judgement and ingenuity of individual developers (Bottaci and Jones, 1994; Bowen and Hinchey, 1995; Jacky, 1989; 1997; Oakley, 1990), and

the lack of support for developmental methodology in this sense has led much of the software community to regard the term "formal method" itself as somewhat misleading (Barden and Stepney, 1993; Bowen and Hinchey, 1994; Jacky, 1997; Macdonald, 1991; Woodcock and Loomes, 1988).

> "So 'Formal Methods' (only) provides a framework in which programs can be developed in a justifiable way. It does not dictate, or even advise, on how manipulations should be applied. There is still a need for the program developer to make decisions and to determine appropriate programming strategies" (Cooke, 1992, p.420).

Formal methods provide software developers with a selection of tools that can be used to produce, and possibly verify, specifications, but, like any other form of tool, it is possible for formal methods to be misused (Bowen and Hinchey, 1995). Software engineering has always been predominantly driven by human judgement and no conceivable developments in the formal methods community are likely to change this. So despite their mathematical foundations and much reputed scientific basis (Cohen et al., 1986; Hall, 1990), errors will continue to arise in, and from, formal specifications simply because of the natural fallibility of their human users (Bowen and Stavridou, 1993; Hall, 1990; Hinchey and Bowen, 1995). Although supporting tools, guidelines and standards might help to reduce the numbers of errors committed, they will not be able to prevent errors from arising completely (Cohen, 1989b). In this light, Hoare's (1984) vision of future software engineering practice, in which mathematical techniques are used to guarantee that specifications can no longer give rise to software defects and conventional testing methods are discarded in favour of formal reasoning, now seems unrealistic and unachievable (Loomes, 1991). The possibility that incorrect development decisions might continue to emanate from specifications, however, is disconcerting in view of the increasingly critical engineering questions being asked of the software community (MacKensie, 1992) and the high degrees of confidence that tend to be placed in systems developed using formal methods (Wing, 1990).

> "Formal methods cannot guarantee correctness; they are applied by humans, who are obviously error prone. Support tools - such as specification editors, type checkers, consistency checkers, and proof checkers - might reduce the likelihood of human error but will not eliminate it. Systems development is a human activity and will always be prone to human whim, indecision, the ambiguity of natural language, and simple carelessness" (Bowen and Hinchey, 1995, p.60).

Formal methods research has historically concentrated on developing new notations and supporting tools. This research has mostly been conducted in purely academic environments and without a full understanding of the types of development problem experienced by industry (Garlan, 1996). Aside from resulting in a proliferation of tools that are often awkward to use, unreliable and unsuited to the types of problem faced by real engineers in industry (Hollo-

way and Butler, 1996), this near exclusive preoccupation with formal methods' supporting technology seems to have distracted computing research from the most important element of the design process; the human users of formal methods themselves. It seems appropriate that computing research should at least pause to consider the cognitive implications of using such technology. Empirical research has, in particular, failed to question the extent to which the human potential for error will remain after formalising the specification process.

4 An Empirical Basis for Assessment

> "One difficulty we encountered in determining the relative advantage of formal methods is the lack of strong scientific evidence that the technology is, in fact, effective. Various surveys have provided reasonably systematic anecdotal evidence of effective industrial use of formal methods. However, none of the formal methods application projects has used strict, scientifically based, measurement data" (Craigen et al., 1995, p.407).

The increasing interest in formal methods shown by the software community (Bowen and Hinchey, 1994; Oakley, 1990) may be attributed to the much publicised claims that formalisation of the specification process will lead to greater benefits than are currently realised with informal methods. Many of these claims, however, are based on subjective belief or isolated case studies from which results can be difficult to generalise (Craigen et al., 1995; Fenton, 1996). That software research has yet to produce substantive evidence which might refute such claims appears, for many, to support the case for formalisation (see for example: Bowen, 1988; Hall, 1990; Sommerville, 1992; Potter et al., 1996). But irrespective of how plausible they might appear at face value, such claims rest on anecdotal, rather than empirical, grounds.

> "In the absence of a suitable measurement system, there is no chance of validating the claims of the formal methods community that their models and theories enhance the quality of software products and improve the cost-effectiveness of software processes" (Fenton and Kaposi, 1989, p.293).

During the course of this research a series of cognitive studies have been conducted in order to determine the linguistic properties of formal specifications which are particularly likely to elicit human errors or biases. Later in this paper, the results of these experiments are recast in terms of a descriptive model designed to yield quantitative measures of complexity in formal specifications. Given that the results were independently and objectively generated, the model also provides an empirical basis for assessing two of the claims associated with formal methods which rest upon psychological assumptions; namely, that formal specifications are open to only one form of interpretation (Bowen, 1988; Liskov and Berzins, 1986; Thomas, 1993), and that it is easier to reason about formal expressions than their informal counterparts (Ince, 1992; Reichenbach, 1966; Thomas, 1995; Wing, 1990).

5 Principles of Software Measurement

The term "measure", in the software measurement literature, refers to a number or symbol assigned to characterise an attribute of an entity (Fenton and Pfleeger, 1996; Kaposi and Myers, 1994; Kitchenham, 1991). Aside from its connotations with decimalisation and mathematical functions, the term "metric" has, in the context of software engineering, become synonymous with "measure". Although, the term "metric" might suggest both a measure and an underlying model or theory, this is not a view that reflects the ways that metrics are normally perceived or applied (Sheppard and Ince, 1993). No such distinction is maintained here and the terms are used interchangeably in this paper.

Numbers may be used to characterise the quality of products or processes, and there exist two corresponding divisions of software metric (Fenton and Pfleeger, 1996; Ince, 1989; Roche, 1994; Sheppard, 1988). "Product metrics" are oriented towards the tangible outputs from development activities such as requirements documents, system design or program code. Typical examples include the number of executable lines of code and the percentage of comments per program module. "Process metrics" are oriented towards the development processes themselves, such as program design, implementation and testing. Typical examples include the predicted costs of a stage of development and the number of defects found during a phase of program testing.

Software metrics can be applied as descriptors or predictors of quality, which reflect the reactive and proactive ways in which they can be used respectively (Ince, 1989; Kaposi, 1991; Kitchenham, 1991). "Descriptor metrics" are used to describe existing products or processes. Typical examples include product measures such as the number of possible routes that may be taken through a program module, and process measures such as the total expenditure on testing. "Predictor metrics" are used to estimate final characteristics of products yet to be developed or to produce estimates of descriptor metrics. These might include product metrics such as the estimated number of lines of source code that will comprise the final system, and process metrics such as the projected costs of an entire project based on the amount spent so far and the work yet to be completed.

Software metrics can again be further categorised according to the type of attribute to be measured; internal or external (Fenton and Pfleeger, 1996). The "internal" attributes of a product or process are those which can be measured purely in terms of the product or process itself. The length of source code, for example, can be measured directly without reference to its behaviour. The "external" attributes" of a product or process are those which are measured in terms of how it interacts with its environment. The reliability of a system, for example, can be measured in terms of the number of times it fails to fulfill a valid service requested by its users. It is a common misconception that the numeric value yielded by an external metric provides a definitive, all encompassing account of the quality of a product or process; the higher the value, the more quality is assumed to exist. External metrics tend to be defined in terms of only a small subset of attributes, however, which are weighted and combined in just one of numerous possible ways.

The metrics to be developed during the course of this paper may be classified as predictive, external and product based. They are predictive because they relate to possible future events; namely, the types of conclusion likely to be drawn by human reasoners in response to formal specifications. They are external because their values are not calculable from formal specifications alone, but also from the ways in which people have reasoned about similar specifications in the past. The metrics are product based because they yield measures oriented towards grammatical properties of specifications such as the presence of specific logical operators, the degree of realistic or believable content, and the types of logical statement that may be inferred from combinations of these.

5.1 Benefits of Early Software Measurement

Reliance upon source code metrics alone, such as those proposed by Halstead (1977) and McCabe (1976), has proved unsatisfactory because they can only be applied at a relatively late in the software development process, once code has been implemented. By this time, effort and expense of producing the system has already been incurred and it can be a costly exercise to backtrack through specification, design, coding and testing to rectify even the slightest mistake or omission (Sheppard and Ince, 1993). It would seem more beneficial to apply metrics earlier in the development process where they are likely to help realise greater savings (Bainbridge et al., 1991; Fenton and Kaposi, 1989; Sheppard, 1988). What is needed, then, to complement existing reactive forms of quality control metric, is a more proactive form of measurement which enables developers to predict where errors are likely to stem from before the system has been built. Earlier measures could, then, provide feedback to software developers and act as a basis for making more informed development decisions before errors have had chance to propagate throughout subsequent development work. It is in view of the low levels of tolerance for error in those areas to which formal methods are applied that Wordsworth (1992, p.68) argues "in a formally developed product the aim is defect prevention rather than defect detection".

5.2 Measuring Specification Complexity

The software measurement literature tells us that complexity metrics are useful indicators of many software quality attributes (Kitchenham, 1991; Sullivan, 1975). For instance, complexity is generally related to reliability because complicated documents are more likely to contain residual errors, and to maintainability because complicated documents require more time and effort to understand before incorrect sections can be located and revised. There is a tendency in the software industry, however, to perceive complexity metrics as accurate indicators of the complete range of attributes that contribute towards the quality of a product or process, such as readability, structural simplicity, maintainability and robustness, when in fact they are based on measures taken from only a narrow subset of such attributes (Fenton, 1992). McCabe's (1976) model, for example, claims to provide indications of cognitive complexity, program defects and maintenance costs. Given that its predictions are based only on the

number of logical decision branches in program code, however, it is debatable as to whether or not these notions are actually within its descriptive power (Evangelist, 1983; Fenton, 1991; Fenton and Kaposi, 1989; Sheppard, 1988).

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

Formal methods research has historically been directed at developing or improving tools and notations, that is, it has been oriented towards the supporting technology rather than its human users. But assuming that the human potential for error will remain after formalisation of the specification process and that inaccurate human reasoning about formal specifications will continue to lead to the introduction of defects in software systems, it seems appropriate that we strive to pre-empt errors in human judgement before these have chance to cause erroneous behaviour. This view is supported by Senders and Moray's (1991, p.66) argument that "we need to know the probability of the mistaken decision even more than the probability of the 'incorrect' actions stemming from it". One way of achieving this is by introducing a model for predicting likely sources of cognitive complexity in formal specifications. Unlike those models proposed by the software community which claim to yield generalisable measures of complexity, our aims are much more modest. We aim only for a tentative model to characterise very specific types of human reasoning, under highly specialised conditions. The intention is not to formulate a general metric with a wide scope for application, but rather to demonstrate the feasibility of the approach.

In view of their central role in project documentation and communication, it is important that the readability of formal specifications is not impaired. To this end, complexity metrics can be used to indicate those parts of a specification which could potentially give rise to human comprehension or reasoning difficulties. The model which we are working towards is concerned with quantifying "inferential complexity", that is, the ability of people to reason or draw conclusions about formal specifications. It is not claimed that the model is indicative of other attributes belonging to formal specifications, such as maintainability or reliability, although the model does incorporate the notions of interpretability and representability. This claim is based on the assumption that, in order to reason about any form of written text, a reader must interpret its meaning and then maintain some internal representation of it whilst the reasoning process is conducted. The proposed model is therefore psychological in nature.

5.2.1 Psychological Versus Computational Complexity

A study of specification complexity could be approached from one of two possibly interrelated directions: computational or psychological. A model of the computational complexity in a formal specification might aim to quantify the efficiency of the algorithms used in its mathematical calculations or the structure of its modules. The main emphasis of such a model would be on achieving maximum system efficiency. A model of psychological complexity, in contrast, might aim to quantify the extent to which attributes of a specification give rise to difficulties in its creation or comprehension. These attributes might include: the symbology of the language, the meanings of its constructs, or the style of writing employed by its designers. The main emphasis of such a model would be on achieving a system that is easier for people to work with. The psychological complexity of software products and processes has largely been overlooked in favour of quantifying the computational attributes of the development process.

> "Assessing the psychological complexity of software appears to require more than a simple count of operators, operands, and basic control paths. If the ability of complexity metrics to predict programmer performance is to be improved, then metrics must also incorporate measures of phenomena related by psychological principles to the memory, information processing, and problem solving capacities of programmers" (Curtis et al., 1979, p.103).

A distinguishing feature of psychological complexity is the interaction between product characteristics and individual differences (Curtis et al., 1979), such as notational constructs and language expertise. Unlike computational complexity, modelling psychological complexity in a software product requires a great deal more than counts of its internal product attributes. Models of psychological complexity must accept as parameters measures relating to both the product and the people who interact with it (Curtis, 1986; Melton et al., 1990; Öry, 1993). All of the specification metrics published to date are calculable purely in terms of grammatical properties of specifications (see for example: Samson et al., 1987), such as dependencies between modules or decision branches, and do not account for the human developers involved in the specification process. Despite claims to the contrary, the predictions of such models are limited from a psychological perspective. It is presumably due to the increasing complexity and criticality of software generally (MacKenzie, 1992) that the emphasis is changing. In addition to the algorithmic efficiency of software, the measurement community is becoming increasingly concerned with the human complexity of evolving systems (Fenton, 1991). This is reflected in recent cognitive research aimed at identifying sources of psychological complexity in formal software specifications (Finney, 1996).

5.3 Statistical Prediction of Human Error

> "Errors result from the normal operation of the human information processing system, along with effects arising from the environment, the various pressures and biases influencing the actor, and the latter's mental, emotional, and attentional states (ignoring the possibility of traumatic events that damage the functioning of the nervous system). In principle, if we knew all these factors, we could predict errors precisely. In practice, since we cannot know all the factors, we will always have to resort to statistical prediction" (Senders and Moray, 1991, p.61).

The task of predicting errors of human judgement with a reasonable degree of accuracy is complex because the reasoning processes which influence human judgement involve numerous psychological issues, few of which are fully understood by professional psychologists. These include the ways in which reasoning is affected by: personality traits, prior beliefs, motivational states, task structure and presentation. Moreover, errors of judgement are frequently ascribable to multiple causes and it is rare indeed that all of these causes are evident, even to their perpetrators, which complicates the task of predicting when and where similar errors might happen in future. Faced with the problem of predicting future events, such as manifestations of human reasoning errors, our predictions will, where possible, be based on full knowledge of the causal factors which lead to the event occurring. "Full knowledge" is unlikely to be available or accessible, however, for predicting the psychological causes of error in human judgement. In the absence of such knowledge, our predictions might be based on the frequencies with which the same, or similar, errors have occurred for similar people in the past. A practical alternative available to us, then, is to make predictions based on probability. There is, after all, always likely be an element of unavoidable uncertainty associated with our predictions given that we can never predict with absolute certainty what will happen in future possible worlds (Springer et al., 1966).

> "The rules which we employ in life-assurance, and in the other statistical applications of the theory of probabilities, are altogether independent of the mental phaenomena of expectation. They are founded upon the assumption that the future will bear a resemblance to the past; that under the same circumstances the same event will tend to recur with a definite numerical frequency; not upon any attempt to submit to calculation the strength of human hopes and fears" (Aristotle, in Ross, 1949, p.244).

The heuristic methods that people use to assess the probabilities of future events are notoriously inadequate for making consistent and reliable predictions (Kahneman et al., 1991). Human predictions, including those made by professionals with the relevant training or domain knowledge, are liable: to focus on irrelevant factors or neglect relevant ones (Evans et al., 1993), to assign incorrect or inconsistent weightings to factors (Nisbett and Ross, 1980), or to be affected by emotional beliefs and motivational states (Dawes, 1971). It is for

these reasons that we cannot rely purely on humans to predict human errors and, more specifically, that we cannot rely upon human software developers alone to pre-empt the sources of their own, and their colleagues', reasoning errors in formal specifications. One possible solution, which will be explored in this paper, would be to augment the predictions of human developers with an independent, mathematical means for assessing the potential of given specifications to admit human reasoning errors.

> "Human judges are not merely worse than optimal regression equations; they are worse than almost any regression equation. Even if the weights in the equation are arbitrary, as long as they are nonzero, positive, and linear, the equation generally will outperform human judges" (Nisbett and Ross, 1980, p.141).

Although the psychological determinants of error are often less than clear, this does not necessarily mean that explicit mathematical models cannot be used to predict their propensity for admitting human error. This view is supported by Meehl's (1973) argument that the complexity of the human mind should not preclude the use of mathematics in decision making. Moreover, it is argued that statistical methods, particularly regression based models, yield more accurate predictions than the naive intuitions of so-called "experts" (Dawes, 1971; 1979; Dawes and Corrigan, 1974; Goldberg, 1970; Slovic and Lichtenstein, 1971). This argument is based on the assumption that statistical methods abstract away from the processes they represent and are not influenced by the kinds of cognitive state or heuristic which affect human prediction such as: aptitude, fatigue, habit or bias. Regression based models consistently attach the same weightings to causal factors, regardless of such extraneous variables, and can generate quantifiably precise estimates of the extent to which each contributes towards an event's occurrence.

6 The Experiments

The fact that people currently err at all in the specification process is a reason for concern, but the possibility that they will continue to do so even after having adopted a formal approach is especially disconcerting given the business and safety critical nature of the projects to which they are applied (Barroca and McDermid, 1992; Bowen and Stavridou, 1993). If software developers assume that the use of formal methods will promote error-free reasoning when in fact this belief is inaccurate, then it is likely to instill a sense of security where none is warranted. Despite obvious syntactic differences, however, most formal notations contain corresponding logical operators with roughly equivalent meanings as those same natural language constructs which have been shown to evoke incorrect decisions in cognitive studies: $\Rightarrow$ (if), $\wedge$ (and), $\vee$ (or), $\neg$ (not), $\exists$ (some) and $\forall$ (all). The main empirical issue addressed during the experiments was to test how far the same non-logical errors and biases that people exhibit when reasoning about natural language also occur when software developers are reasoning about the logically equivalent statements in formal specifications.

6.1 Selecting a Formal Notation

In order to conduct the experiments and formulate metrics it was necessary to use a notation with a single, concrete syntax. The range of formal notations developed for the purpose of software specification has been steadily increasing during the past decade, however, and the range of notations now available vary widely in their mathematical foundations, purpose and popularity. In order to identify a suitable grammatical framework which could meet the research aims, a systematic review of twenty formal notations was conducted against pre-determined criteria (Vinter, 1996). The reasons for specifically choosing the Z notation (Spivey, 1992) were as follows.

1. The fact that Z is one of the most popular notations used in academia and industry (Dean and Hinchey, 1996) eased the task of finding adequate numbers of suitably skilled participants for the experiments.

2. The mathematical calculi underlying the Z notation, predicate logic (Lemmon, 1993) and set theory (Johnstone, 1987), provide the grammatical basis for many other formal notations. It is therefore possible that the research findings may generalise to those notations sharing the same logical foundations: Gypsy (Ambler, 1977), Larch (Guttag et al., 1985), RAISE (RAISE Language Group, 1992) and VDM (Jones, 1989).

3. Z is reputed to be one of the more easily readable formal notations (Bowen, 1988; Jack, 1992). The probabilities yielded by our model, for quantifying the likely levels of correctness for Z users, are therefore likely to be among the highest for any of the notations that the formal methods community has to offer.

4. Having been applied in industry for developing a diverse range of applications (Barden et al., 1992), Z is one of the more well established and commercially viable formal methods. This has favourable implications for the longevity and credibility of the model's theoretical basis.

5. Perhaps owing to the current drive for a rigorous deductive proof system and an international standard for Z (Brien and Nicholls, 1992), not to mention its increasing acceptance in industry (Hall, 1996; Nix and Collins, 1988), the Z notation represents a highly active area of academic research and one in which much still remains to be explored.

6.2 Tasks

In light of relevant findings from cognitive science, the main concerns of our experiments were to determine how far human reasoning in formal contexts is affected by: the meaningful content of problem material, the polarity of terms in logical rules, the types of inference to be drawn, and the believability of problem material. Participants were divided into two groups for each of the experiments, abstract formal logic (AFL) and thematic formal logic (TFL), with twenty different participants completing the tasks under each linguistic

condition. The stimuli for each task comprised a prose description and several Z statements, at least one of which contained a logical rule. These represented the premisses necessary for the inference to be drawn. Participants were also shown four Z statements representing possible conclusions, one of which was indeterminate; "No valid conclusion" or "Nothing".

6.3 Participants

A total of one hundred and twenty people took part in the experiments, all of whom had the relevant knowledge of the Z formal notation and training in logical deduction. These consisted of staff and students from academic institutions, and computing professionals from industrial software companies. Participants comprised 72 staff, 26 students, and 22 professionals. Their mean age was 33.55 years ($s = 9.83$) and 110 had studied at least one system of formal logic beforehand, such as the propositional or predicate calculi, Boolean algebra or Higher Order Logic. Their mean level of Z experience was 4.63 years ($s = 4.04$). According to their personal ratings of expertise, participants comprised 31 novice, 54 proficient and 35 expert users of the Z notation. All were recruited via personal invitation.

6.4 Procedure

Prior to completing the main tasks, participants were asked to provide the following biographical information: occupation, age, organisation, course, number of years' Z experience, a list of other formal notations known, a subjective rating of their Z expertise (novice, proficient or expert), and details of any system of formal logic studied beforehand, such as the propositional or predicate calculus, Boolean algebra or Higher Order Logic. Experimental groups were counter balanced, firstly, according to participants' personal ratings of Z expertise and, secondly, according to their lengths of Z experience. All task sheets were computer generated. These were distributed to participants and completed anonymously then mailed back to the experimenter. All participants were tested on an individual basis. For a more detailed description of the methodology used in each study readers are referred to Vinter et al. (1997a; 1997b; 1997c).

6.4.1 Conditional Inferences

Cognitive science has generated a wealth of evidence to suggest that people are prone to various forms of error and bias when reasoning about conditional statements expressed in natural language. "Matching bias" is said to occur when an individual selects, or evaluates as relevant, only those conclusions which contain one or more of the terms mentioned explicitly in the given premisses (Evans, 1972a). "Negative conclusion bias" is said to occur where an individual is more inclined to endorse an inference whose conclusion is negative rather than affirmative, which often maximises the individual's chances of making statements that are unlikely to be disproved in everyday reasoning (Evans, 1993; Pollard and Evans, 1980). "Affirmative premiss bias" is said to occur where an individual is predisposed only to draw determinate conclusions from premisses that do not

contain any negative components (Evans, 1993). The hypothesis of "facilitation by realism" argues that realistic, as opposed to abstract symbolic, task content can have strong facilitatory effects on conditional reasoning performance (Gilhooly and Falconer, 1974; Griggs and Cox, 1982; Johnson-Laird et al., 1972; van Duyne, 1974). This hypothesis stems from the theory of "belief bias", whereby the individual is claimed to respond primarily according to his or her prior beliefs towards the real world referents of a task's meaningful content, rather than its logical structure (Barston, 1986; Evans et al., 1993).

In order to test whether trained users of formal methods are liable to succumb to the same non-logical tendencies when reasoning about conditional statements expressed in the Z notation, our tasks required participants to draw the same types of inference shown to elicit errors and biases in the natural language domain: modus ponens (MP), modus tollens (MT), denial of the antecedent (DA) and affirmation of the consequent (AC). It should be noted that the MP and MT inferences lead to determinate conclusions, whereas AC and DA are fallacious inferences in which nothing can be deduced logically. The presence of affirmatives (A) and negatives (N) were systematically varied in the conditional rules in order to test for possible polarity biases. This gave rise to sixteen logical tasks whose forms are illustrated in Table 1.

TABLE 1
Logical forms of the conditional inference tasks

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

Note: Conclusions shown for DA and AC are fallacious.

Two versions of each task were formulated in order to test for possible effects of realistic material; one expressed in abstract terms and one in thematic terms. All abstract tasks described relations between colours and shapes, whereas the thematic tasks described realistic computing applications including: a library database system, a flight reservation system, a missile guidance system, a video lending system, and a vending machine operation. The materials for the conditional reasoning tasks are exemplified in Figures 1 and 2 which show the abstract and thematic versions of the AC-AN inference task respectively. An asterisk indicates the logically correct conclusion in each case.

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

Figure 1: Abstract conditional AC-AN inference

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

Figure 2: Thematic conditional AC-AN inference

6.4.2 Disjunctive Inferences

Many of the errors that reasoners commit when reasoning about disjunctive rules in laboratory based studies stem from the ambiguity of "or" in everyday communication and people's frequent uncertainty about whether to draw inclusive or exclusive interpretations (Braine and Rumain, 1981; Hurford, 1974). In light of Newstead and Griggs' (1983a) hypothesis that disjunctive reasoning should be better in those languages where disjunctives are defined unambiguously, one would not have expected our participants to experience this form of uncertainty in the explicitly formal context of the Z notation, given that the concept of exclusive disjunction is not defined as part of the standard notation[2] (Brien and Nicholls, 1992). Evans et al. (1993) suggest that people's inclination to adopt inclusive interpretations of disjunctive rules, when exclusive interpretations are appropriate, can lead them to draw determinate conclusions when indeterminate ones are logically appropriate; participants thereby exhibit a bias towards propositional conclusions. Once the correct form of interpretation has been adopted, however, the cognitive literature reports that people generally find it easier to reason with exclusive than inclusive disjunctives (Newstead et al., 1984; Newstead and Griggs, 1983a; Roberge, 1977; 1978). As a possible explanation for this finding, it is argued that exclusive disjunctives lead

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

to symmetrical inferences; that is, by knowing the truth value of one disjunct the truth value of the other can be deduced.

In order to test whether our participants succumbed to the same non-logical tendencies, our tasks tested participants' abilities to draw the same forms of "denial" and "affirmation" inference shown to elicit errors and biases in the natural language domain. The disjunctive tasks involved either the denial of a component or the affirmation of a component from the major premiss by the minor premiss. The presence of negatives in the major premiss and the position of the component affirmed or denied in the minor premiss were systematically varied in order to test for possible polarity biases. Table 2 illustrates the logical forms of these tasks. It should be noted that, whilst the denial inferences are logically sanctionable, all of the affirmation inferences shown are fallacious under a logical, inclusive interpretation of the Z "$\vee$" operator.

TABLE 2
Logical forms of the disjunctive denial and affirmation tasks

Polarity	Term Denied or Affirmed	Denial	Affirmation
AA	1	$p \vee q, \neg p \ \therefore q$	$p \vee q, p \ \therefore \neg q$
AA	2	$p \vee q, \neg q \ \therefore p$	$p \vee q, q \ \therefore \neg p$
AN	1	$p \vee \neg q, \neg p \ \therefore \neg q$	$p \vee \neg q, p \ \therefore q$
AN	2	$p \vee \neg q, q \ \therefore p$	$p \vee \neg q, \neg q \ \therefore \neg p$
NA	1	$\neg p \vee q, p \ \therefore q$	$\neg p \vee q, \neg p \ \therefore \neg q$
NA	2	$\neg p \vee q, \neg q \ \therefore \neg p$	$\neg p \vee q, q \ \therefore p$
NN	1	$\neg p \vee \neg q, p \ \therefore \neg q$	$\neg p \vee \neg q, \neg p \ \therefore q$
NN	2	$\neg p \vee \neg q, q \ \therefore \neg p$	$\neg p \vee \neg q, \neg q \ \therefore p$

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

In order to test for possible effects of realistic material, two versions of each task were formulated; one expressed in abstract terms describing relations between colours and shapes, and one expressed in thematic terms describing realistic computing applications. The materials for the disjunctive tasks are exemplified in Figures 3 and 4 which show the abstract denial AN-1 task and thematic affirmation NN-1 task respectively.

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

GetShapeColour

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

$colour! = white \vee \neg(shape! = rectangle)$

(a) $shape! = square$ (c)* $\neg(shape! = rectangle)$
(b) $shape! = circle$ (d) Nothing

Figure 3: Abstract disjunctive denial AN-1 inference

If $\neg(processor! = Pentium)$ after its execution, what can you say about $display!$ in operation *ComputerHardware*?

ComputerHardware

$processor! : Chip$
$display! : Screen$

$\neg(processor! = Pentium) \vee \neg(display! = HighResolution)$

(a) $display! = LowResolution$ (c)* $display! = HighResolution$
(b) $\neg(display! = HighResolution)$ (d) Nothing

Figure 4: Thematic disjunctive affirmation NN-1 inference

6.4.3 Conjunctive Inferences

The linguistic literature argues that the principles governing the use of disjunction and conjunction in English are similar because both require a common topic between two terms, and this may be overtly present or derivable by presupposition and deduction (Lakoff, 1971). Cognitive science has, in general, been slow to address the question of whether people are prone to similar kinds of systematic error and bias as those observed in studies of conditional and disjunctive reasoning. A series of studies conducted by Tversky and Kahneman (1983), however, provide evidence of people's fallibility when reasoning with conjunctive statements. Probability theory states that the likelihood of a conjunction, "p and q", cannot exceed the likelihood of one of its constituent outcomes, p or q. The experimenters suggest, however, that people's use of intuitive heuristics, rather than the conjunctive laws of logic, led them them to violate this principle in a range of realistic contexts.

Those systems of logic defined in terms of the principles underlying Gentzen's (in Szabo, 1969) deductive calculus include two types of inference rule for connecting logical chains of reasoning; those for introducing and those for eliminating propositional connectives. The conjunctive reasoning tasks presented to participants involved either the introduction or the elimination of logical components in a similar fashion. The polarity of components in the tasks' premisses and the order of components introduced and eliminated were systematically varied. This design gave rise to the tasks whose logical forms are shown in Table 3. It should be noted that, whilst the elimination inferences are logically sanctionable, all of the introduction inferences shown are fallacious.

TABLE 3
Logical forms of the conjunctive elimination and introduction tasks

Polarity	Term Eliminated	Elimination	Polarity	Term Introduced	Introduction
AA	2	$p \wedge q \ \therefore p$	A	1	$p \ \therefore p \wedge q$
AN	1	$p \wedge \neg q \ \therefore \neg q$	A	2	$q \ \therefore p \wedge q$
NA	2	$\neg p \wedge q \ \therefore \neg p$	N	1	$\neg p \ \therefore \neg p \wedge q$
NN	1	$\neg p \wedge \neg q \ \therefore \neg q$	N	2	$\neg q \ \therefore p \wedge \neg q$

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

The materials for the conjunctive reasoning tasks are exemplified in Figures 5 and 6 which show the conjunctive elimination NA-2 and introduction A-1 tasks respectively.

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

HireVideo

$\Delta VideoShop$

$\neg(film' \in FilmsOnShelf) \wedge report' = OnLoan$

(a) $film' \in FilmsOnShelf \wedge report' = OnLoan$ (c)* $\neg(film' \in FilmsOnShelf)$
(b) $\neg(report' = OnLoan)$ (d) Nothing

Figure 5: Thematic NA-2 conjunctive elimination inference

What can you say about the effect of operation *GuidedMissileCheck* on its after state variables?

GuidedMissileCheck

$\Delta Bearings$
$target\_loc? : COORDS$

$current\_loc' = target\_loc?$

(a) $\neg(current\_loc' = target\_loc?) \wedge mission' = Failure$ (c) $\neg(current\_loc' = target\_loc?)$
(b)* $\neg(current\_loc' = target\_loc?) \wedge mission' = Failure$ (d) Nothing

Figure 6: Thematic conjunctive introduction A-1 inference

6.4.4 Quantified Inferences

It is argued that cognitive studies of syllogistic reasoning provide important pointers to the cognitive processes which people employ in quantified reasoning (Johnson-Laird and Bara, 1984) and in human reasoning generally (Dickstein, 1978b). A categorical syllogism is an argument consisting of three statements: a major premiss, a minor premiss and a conclusion. Each of these statements

describe relations between the various "terms" of the argument. The major premiss describes the relation that holds between the predicate of the conclusion (P) and a middle term (M). The minor premiss describes the relation that holds between the subject of the conclusion (S) and the middle term. The aim of the syllogistic task is to use the two premisses as the basis for deducing a conclusion which describes a relation that exists between S and P, or, where the premisses cannot lead to such a deduction, to state that no determinate conclusion follows. Four types of quantifier may range over the assertions made in a syllogism: "All", "Some", "Some ... not" and "No". The quantifier which ranges over a syllogistic predicate reflects that predicate's "mood", conventionally abbreviated as shown in Figure 7.

(A)	All A are B	$\forall x : Type \bullet A(x) \Rightarrow B(x)$
(I)	Some A are B	$\exists x : Type \bullet A(x) \wedge B(x)$
(O)	Some A are not B	$\exists x : Type \bullet A(x) \wedge \neg B(x)$
(E)	No A are B	$\neg \exists x : Type \bullet A(x) \wedge B(x)$

Figure 7: Z translations of the four syllogistic moods

The ordering of terms in a syllogism's premisses is significant. As there are two possible orderings for each of the major and minor premisses, this gives rise to four possible arrangements, or "figures", as shown in Figure 8. Although the order in which terms are presented within the two premisses might vary, the ordering of terms in the conclusion always proceeds from S to P.

Figure 1	Figure 2	Figure 3	Figure 4
M-P	P-M	M-P	P-M
S-M	S-M	M-S	M-S
—	—	—	—
S-P	S-P	S-P	S-P

Figure 8: The four figures of a syllogism

The cognitive literature has been keen to propound numerous explanations for the possible causes of error in the syllogistic task. "Atmosphere theory" predicts that, where the relationship between S and P is less than obvious, the reasoner will draw a conclusion which shares the same qualifiers and quantifiers as those contained in the premisses, with little or no regard for the underlying logic of the syllogism (Begg and Denny, 1969; Woodworth and Sells, 1935). The theory of "matching bias" argues that, where the reasoner is unsure of how to reach a valid conclusion via logical deduction, he or she will simply choose a conclusion whose quantitative form matches one of the two premisses (Evans, 1972a; Wetherick and Gilhooly, 1990). The theory of "implicit premiss conversion" argues that reasoners often attempt to convert one or both premisses to simpler forms which are more amenable to mental representation before attempting a logical analysis. Illicitly converted premisses can therefore form a false basis from which erroneous conclusions are drawn (Dickstein, 1981; Newstead and Griggs, 1983b; Revlin and Leirer, 1980; Wilkins, 1928).

It is argued that the ability of reasoners to differentiate between the pragmatic laws of language and the laws of logic is a major determinant of reasoners' performance in deductive tasks (Politzer, 1986). Several studies suggest that reasoners are predisposed to apply Gricean (1975) conventions from everyday discourse to the syllogistic task (Begg and Harris, 1982; Newstead, 1989; 1995). The theory of the "Same M fallacy" predicts that, whenever the subject and predicate of a speculative conclusion are related by a common middle term (i.e. the same M), reasoners will tend to accept this conclusion at face value, according to the Gricean maxim of relation, irrespective of its logical necessity (Chapman and Chapman, 1959; Dickstein, 1975; 1976). The theory of "caution bias" claims that reasoners are predisposed to accept "Some ... are" conclusions more readily than "All ... are" conclusions, and "Some ... are not" conclusions more readily than "None ... are" conclusions, because reasoners are generally conservative estimators (Woodworth and Sells, 1935). It is argued that reasoners generally exhibit a bias towards propositional conclusions which misleads them into interpreting or combining premisses in ways that can only lead to determinate conclusions, or into discounting hypothetical possibilities which lead to indeterminate conclusions (Chapman and Chapman, 1959; Dickstein, 1975; 1976; 1978b; Revlis, 1975; Traub, 1977).

Reasoners succumb to "belief bias" in syllogistic studies by accepting at face value arguments whose conclusions they believe, regardless of their logical validity, and only scrutinising those arguments whose conclusions do not conform with their beliefs (Evans et al., 1983; Henle and Michael, 1956; Janis and Frick, 1943; Morgan and Morton, 1944; Revlis, 1975; Wilkins, 1928). The theory of "figural bias" claims that the figure of a syllogism determines the order in which end terms are related during premiss integration (Johnson-Laird and Steedman, 1978), and that a directional bias in our cognitive processes makes it easier to scan represented premisses in certain directions (Johnson-Laird and Bara, 1984). It is also argued that people reason about syllogisms in ways analogous to those in which Venn Diagrams or Euler Circles are used in mathematics, whereby interpretation of premiss information creates a combined mental representation showing the set relations that may exist between syllogistic terms (Ceraso and Provitera, 1971; Erickson, 1974; 1978; Traub, 1977). Errors then become explainable as a consequence of reasoners' use of inappropriate representations or their failure to consider all possible hypothetical combinations of set relations that follow from a given premiss pair (Ceraso and Provitera, 1971; Dickstein, 1978b; Erickson, 1974).

In order to test whether our participants were prone to the same forms of error and bias when reasoning about quantified statements expressed in the Z notation, we used logically equivalent tasks to those used in natural language based syllogistic studies. The main aims of our study were to determine the extent to which mood, figure, meaningful content and the believability of conclusions affected reasoning performance. In order to test for possible effects of realistic material, meaningful identifiers were used for function names in the thematic versions of the tasks. These names were chosen so as to refer to concepts with which participants would be familiar including: social groups, occupations, animals and foods. In contrast, arbitrary single letter identifiers

were used for function names in the abstract tasks. For the practical purposes of our study, we included only a cross-section of the possible mood and figure combinations, as shown in Table 4. The abstract tasks comprised 30 syllogisms (15 with determinate and 15 with indeterminate conclusions), whilst the thematic tasks comprised 40 syllogisms (15 with determinate believable conclusions, 15 with indeterminate believable conclusions, 5 with determinate unbelievable conclusions, and 5 with indeterminate unbelievable conclusions).

TABLE 4
Logical forms of the quantified inference tasks

Prem.	Conc.	Prem.	Conc.	Prem.	Conc.	Prem.	Conc.	Prem.	Conc.
AA1	A (I)	AA2	N	AA3	N	AA4	N	AI1*	I
AI3*	I	AO2	O	AO4*	N	AE2	E (O)	AE4	E (O)
IA3	I	IA4*	I	II3*	N	II4	N	IE1*	N
IE2*	N	IE4	N	OA1*	N	OA3*	O	OO3	N
OO4	N	EA1	E (O)	EA2	E (O)	EA3	N	EA4	N
EI1	O	EI2	O	EI3	O	EI4*	O	EE4	N

Note: Two versions of those syllogisms marked with an asterisk were presented to the TFL group; one with a believable conclusion, one with an unbelievable conclusion. Weak conclusions are shown in parentheses.

The materials used for the quantified reasoning tasks are exemplified in Figures 9, 10 and 11. These show the abstract AA1, thematic EA1 (believable) and thematic IA4 (unbelievable) tasks respectively. A plus sign indicates a weak conclusion where a stronger one was also possible.

$$\forall x : X \bullet B(x) \Rightarrow C(x)$$
$$\forall x : X \bullet A(x) \Rightarrow B(x)$$

(a) $\exists x : X \bullet A(x) \wedge C(x)$
(b)* $\forall x : X \bullet A(x) \Rightarrow C(x)$
(c) $\neg \exists x : X \bullet A(x) \wedge C(x)$
(d) No valid conclusion

Figure 9: Abstract determinate AA1 syllogism

$$\neg \exists p : Person \bullet millionaire(p) \wedge poor(p)$$
$$\forall p : Person \bullet rich(p) \Rightarrow millionaire(p)$$

(a)* $\neg \exists p : Person \bullet rich(p) \wedge poor(p)$
(b)$^+$ $\exists p : Person \bullet rich(p) \wedge \neg poor(p)$
(c) $\forall p : Person \bullet millionaire(p) \Rightarrow rich(p)$
(d) No valid conclusion

Figure 10: Thematic determinate EA1 syllogism with believable conclusion

$\exists p : Person \bullet capitalist(p) \wedge Russian(p)$
$\forall p : Person \bullet Russian(p) \Rightarrow communist(p)$

(a)* $\exists p : Person \bullet communist(p) \wedge capitalist(p)$
(b) $\neg \exists p : Person \bullet communist(p) \wedge capitalist(p)$
(c) $\exists p : Person \bullet Russian(p) \wedge \neg capitalist(p)$
(d) No valid conclusion

Figure 11: Thematic determinate IA4 syllogism with unbelievable conclusion

7 Formulating Inferential Complexity Metrics for Z

7.1 Results of the Conditional Reasoning Study

The frequencies of valid inferences for each combination of premiss polarity endorsed by the abstract and thematic groups, during the study of conditional reasoning, are shown in Figures 12 and 13. Participants' levels of correctness under each experimental condition revealed overall rank orders as follows: AFL (79%) < TFL (90%) for group type, DA (73%) < MT (83%) < AC (85%) < MP (98%) for inference type, and NA (81%) < NN (85%) = AA (85%) < AN (88%) for polarity type. Analyses of variance revealed a significant main effect of inference type ($F_{(3,114)} = 7.53, p < 0.01$), and a significant interaction between inference and group type ($F_{(3,114)} = 2.64, p = 0.05$). Analyses by linear regression revealed significant correlations between participants' lengths of Z experience and their levels of correctness ($R = 0.41, F_{(1,39)} = 7.82, p < 0.01$), and between their Z expertise ratings and levels of correctness ($R = 0.45, F_{(1,39)} = 9.40, p < 0.01$).

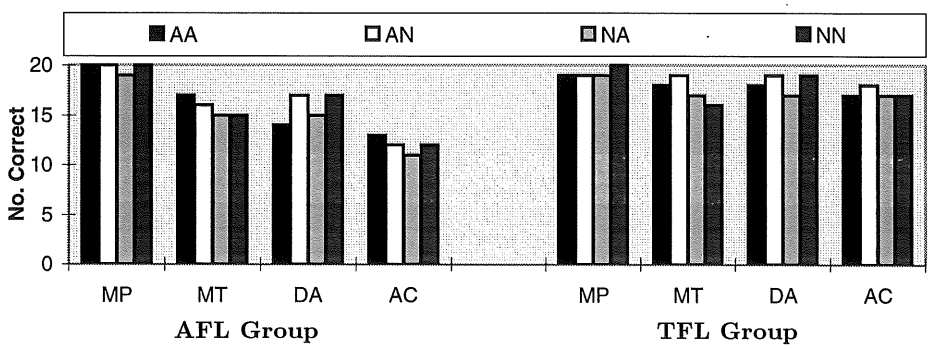

Figure 12: Frequencies of valid conditional inferences endorsed

The frequencies of fallacious DA and AC inferences endorsed by the abstract and thematic groups are shown in Figure 13.

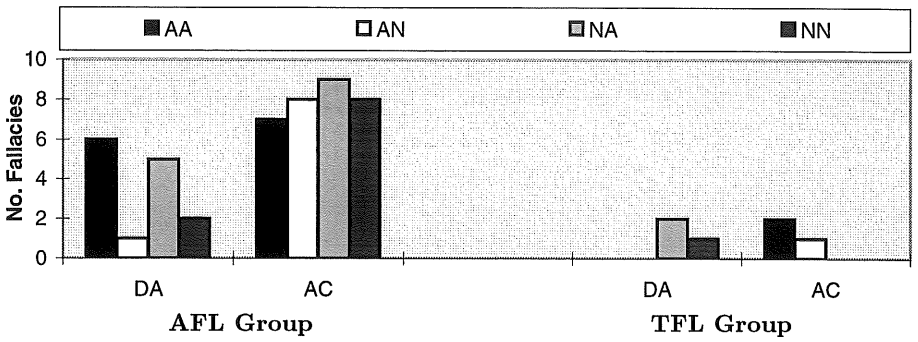

Figure 13: Frequencies of fallacious conditional inferences endorsed

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 both groups. This is supported by natural language based studies in which participants were rarely observed to err when drawing MP inferences, irrespective of the term polarities (Evans, 1972a; 1977; Evans et al., 1995; Taplin, 1971) and the realistic content involved (Griggs and Cox, 1982; Pollard and Evans, 1987). The lower rates of correct MT inferences drawn in comparison to MP is also supported by natural language based studies (Evans, 1977; Kern et al., 1983; Taplin, 1971). The results suggest that participants experienced some difficulty in drawing DA inferences, where up to 30% succumbed to the fallacy, and that participants experienced most difficulty in drawing the AC inferences where up to 45% succumbed to the fallacy. The high rates at which participants succumbed to these fallacies when reasoning about abstract material, in particular, is also supported in the cognitive literature (Evans, 1972b; 1983; Evans et al., 1995; Taplin, 1971; Taplin and Staudenmayer, 1973). A brief summary of the results in relation to cognitive theories of conditional reasoning biases is given below - readers are referred to Vinter et al. (1997a) for a detailed discussion.

- Signs of matching bias (Evans, 1972a) were evident for the fallacious DA and AC inferences in the abstract group.
- Signs of negative conclusion bias (Evans, 1993) were evident for the MT, DA and AC inferences in abstract group.
- Signs of facilitation by realism (Johnson-Laird et al., 1972) and belief bias (Barston, 1986) were evident in the thematic group.
- No signs of affirmative premiss bias (Evans, 1993) were evident.

7.1.1 A Model of Conditional Reasoning

A logistic regression analysis was used to model the data points generated during our study of conditional reasoning. Table 5 shows that the greatest variance in participants' correctness was accounted for, firstly, by the reasoner's level of expertise then, secondly, by the type of inference to be drawn and, lastly, by the degree of meaningful content in task material. The χ^2 values may be

interpreted as the improvements to the accuracy of the model's predictions each time a significant variable was added as a parameter to the model in a forward stepwise manner. Although the accuracy of a logistic regression model's predictions generally increases along with the number of input parameters it allows, there comes a point at which the inclusion of new parameters does not improve the accuracy of the model significantly. This explains why polarity type has been excluded as a parameter from the model and a "fit" to the observed data has been achieved using only three parameters: expertise level, inference type and material type.

TABLE 5
Improvements made to the conditional model by stepwise addition of variables

Step	χ^2 Improvement	DF	Significance	Variable Added
1	53.64	2	0.00	Expertise Level
2	47.37	3	0.00	Inference Type
3	12.55	1	0.00	Material Type

A standard measure of how well a regression based model fits its data is to classify the proportion of predictions given by the model which are consistent with the observed data points from which the model was initially generated (Norušis, 1996). The "Classification-fit" for our model of conditional reasoning is 87.81%. Given that only 12.19% of our data points were misclassified, this suggests that the model provides a reasonable fit to the data. Another standard measure of how well a regression based model fits its observed data is called the "Goodness-of-fit". This statistic compares the observed probabilities with those predicted by the model. The Goodness-of-fit for our model of conditional reasoning is 640.225. This value is calculated as follows (adapted from Norušis, 1996, p.10).

$$\text{Goodness-of-fit} = \sum \frac{Residual_i^2}{P_i(1-P_i)}$$

> ... where Residual is the difference between the observed value and the predicted value P_i

A logistic regression analysis generated the results shown in Table 6. This table shows: how our significant experimental variables became encoded as input parameters to the model, their relative contributions to participants' correctness (β), the standard error (SE), the degrees of freedom (DF), and their significance. β_x is the variable mean, calculated as the summation of the β values for each factor in the variable, divided by the number of factors in the variable. The regression constant, *Const*, refers to the overall mean probability of being correct independent of the influence from other variables.

TABLE 6
Parameters in the model of conditional reasoning

Factor	Parameter	β	SE	DF	Significance	β_x
Material-Abstract	$M1$	-0.8794	0.25	1	0.00	-0.4397
Inference-MP	$I1$	2.9167	0.55	1	0.00	
Inference-MT	$I2$	0.7010	0.30	1	0.02	1.1197
Inference-DA	$I3$	0.8610	0.31	1	0.01	
Expertise-Novice	$E1$	-1.7765	0.40	1	0.00	-0.5991
Expertise-Proficient	$E2$	-0.0207	0.45	1	0.96	
	Const	2.4588	0.20	1	0.00	

The β estimates yielded by a logistic regression show the extent to which each of their corresponding factors influence the dependent variable. In the context of our reasoning studies, as β increases in value so does the participants' likelihood of being correct under the corresponding experimental condition. These values represent the parameters for our conditional model of inferential complexity. According to Kleinbaum (1994), the "odds" of an event occurring are calculated by the probability that it will occur divided by the probability that it will not. The summation of the β estimates give the log of the odds, or "logit" value, as shown in the following general formula.

$$\text{logit}(\textit{Material}, \textit{Inference}, \textit{Expertise}) = \\ \textit{Const} + \beta_{M1} + \beta_{I1} + \beta_{I2} + \beta_{I3} + \beta_{E1} + \beta_{E2}$$

The following examples demonstrate how the formula can be applied to calculate the logit values for conditional inferences under a range of conditions. They also illustrate how the calculations are always performed relative to the regression constant.

$$\text{logit}(\textit{Abstract}, \textit{MP}, \textit{Novice}) = (\textit{Const} + \beta_{M1} + \beta_{I1} + \beta_{E1}) - (\beta_{Mx} + \beta_{Ix} + \beta_{Ex})$$
$$\text{logit}(\textit{Abstract}, \textit{DA}, \textit{Expert}) = (\textit{Const} + \beta_{M1} + \beta_{I3}) - (\beta_{Mx} + \beta_{Ix} + \beta_{Ex})$$
$$\text{logit}(\textit{Thematic}, \textit{MT}, \textit{Expert}) = (\textit{Const} + \beta_{I2}) - (\beta_{Mx} + \beta_{Ix} + \beta_{Ex})$$
$$\text{logit}(\textit{Thematic}, \textit{AC}, \textit{Expert}) = \textit{Const} - (\beta_{Mx} + \beta_{Ix} + \beta_{Ex})$$

7.2 Results of the Disjunctive Reasoning Study

The frequencies of valid and fallacious responses observed during the formalised disjunctive study involving affirmation and denial inferences are shown in Figures 14 and 15. Participants' levels of correctness under each experimental condition revealed overall rank orders as follows: TFL (88%) < AFL (93%) for group type, Affirmation (89%) < Denial (92%) for inference type, First (90%) < Second (91%) for term type, and NA (89%) < AN (90%) < NN (91%) < AA (93%) for polarity type. Analyses of variance failed to reveal any significant effects of the manipulated variables on participants' levels of correctness. Analyses by linear regression revealed significant correlations between participants' years of Z experience and their levels of correctness (R

$= 0.28$, $F_{(1,39)} = 3.11, p = 0.09$), and between participants' ratings of expertise and their levels of correctness ($R = 0.33$, $F_{(1,39)} = 4.55, p = 0.04$).

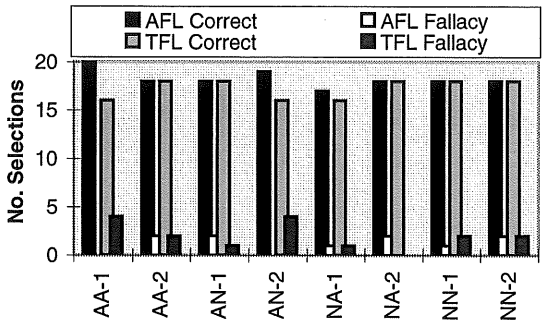

Figure 14: Affirmation inferences

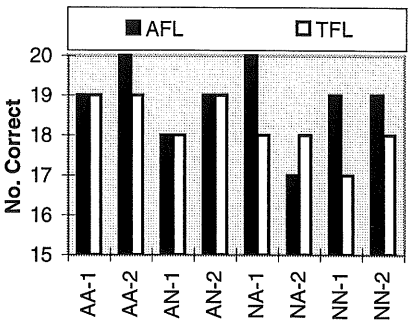

Figure 15: Denial inferences

Given a disjunction, "$p \vee q$", and the assertion of one of its disjuncts, "p", an exclusive interpretation allows us to deduce the falsity of the other, $\neg q$. Under an inclusive interpretation, however, one cannot logically infer anything about the truth value of "q" because both disjuncts might be true. Although most participants responded correctly to the affirmation tasks, around one tenth gave responses consistent with exclusive interpretations of the disjunctive rules, despite the inclusively defined semantics of the Z disjunctive operator. Natural language based studies report fallacious error rates of 36% (Evans et al., 1993) and 20% (Roberge, 1976b; 1977; 1978) for the logically equivalent tasks. The results of the formalised disjunctive study, in comparison, appear to support Newstead and Griggs' (1983a) hypothesis that disjunctive reasoning should be better in those languages where disjunctives are defined unambiguously.

In a study of deductive reasoning in natural language (van Duyne, 1974), strong correlations between realistic material and conditional reasoning performance are reported, but no such correlations are found for disjunctive reasoning performance. Given that the abstract group consistently outperformed the thematic group, our results suggest that a similar situation might exist in the formal domain, whereby the expression of disjunctive tasks in meaningful material does not facilitate performance in the same ways observed for conditionals (Gilhooly and Falconer, 1974; Griggs and Cox, 1982; Johnson-Laird et al., 1972). This finding is supported by Roberge (1977), who reports suppressive effects of meaningful material on reasoning with inclusive disjunctives.

The results suggest that formalisation can lead to notable improvements in logical reasoning for disjunctive based inferences, but that users are still liable to err on occasion. A brief summary of the results in relation to cognitive theories of disjunctive reasoning biases is given below - readers are referred to Vinter et al. (1997b) for a detailed discussion.

- The exclusive interpretation of inclusive disjunctives (Hurford, 1974) appears to have accounted for around one tenth of participants' errors.

- Rather than facilitate disjunctive reasoning, as has been reported in conditional reasoning studies (Gilhooly and Falconer, 1974; Griggs and Cox,

1982), the use of thematic material appeared to suppress it.

- Participants' experienced more difficulty with disjunctive premisses containing mixed, rather than matching, premiss polarities (Roberge, 1976a).
- No signs of reasoning difficulties caused by double negation (Roberge, 1976b) were evident.

7.2.1 A Model of Disjunctive Reasoning

A logistic regression analysis was used to model the data points generated during our study of disjunctive reasoning. Table 7 shows that the greatest variance in participants' correctness was accounted for, firstly, by the reasoner's level of expertise then, secondly, by the degree of meaningful content in the task material. A fit to the data (Classification-fit = 90.63%, Goodness-of-fit = 641.615) was achieved by excluding the following parameters: the type of inference to be drawn, the polarity of its premisses, and the position of the term denied or affirmed.

TABLE 7
Improvements made to the disjunctive model by stepwise addition of variables

Step	χ^2 Improvement	DF	Significance	Variable Added
1	33.272	2	0.00	Expertise Level
2	4.336	1	0.37	Material Type

The β estimates quantifying the degree of influence exerted by each of these variables on participants' correctness during our study of disjunctive reasoning are shown in Table 8.

TABLE 8
Parameters in the model of disjunctive reasoning

Factor	Parameter	β	SE	DF	Significance	β_x
Material-Abstract	$M1$	0.5888	0.29	1	0.04	0.2944
Expertise-Novice	$E1$	-2.4737	0.55	1	0.00	-1.4719
Expertise-Proficient	$E2$	-1.9421	0.54	1	0.00	
	Const	2.5742	0.20	1	0.00	

The general logit formula for predicting the level of inferential complexity associated with a Z disjunctive expression is as follows.

$$\text{logit}(Material, Expertise) = Const + \beta_{M1} + \beta_{E1} + \beta_{E2}$$

7.3 Results of the Conjunctive Reasoning Study

The frequencies of valid and fallacious responses observed during the formalised study involving the conjunctive elimination and introduction inferences are shown in Figures 16 and 17. Participants' levels of correctness under each experimental condition revealed overall rank orders as follows: TFL (90%) < AFL (94%) for group type, Elimination (91%) < Introduction (93%) for inference type, Second (90%) < First (94%) for term type, AN (85%) < NN (90%) < AA (95%) = NA (95%) for eliminated polarity type, and N (90%) < A (96%) for introduced polarity type. An analysis of variance revealed a main effect of group type on correctness approaching significance ($F_{(1,38)} = 3.85, p = 0.06$). An analysis by linear regression revealed a correlation between participants' ratings of expertise and their levels of correctness approaching significance ($R = 0.3$, $F_{(1,39)} = 3.76, p = 0.06$).

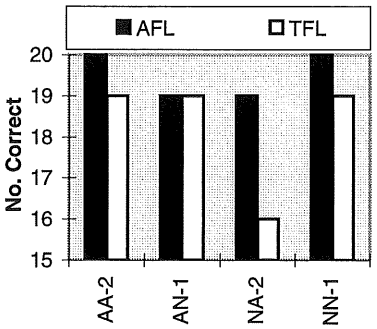

Figure 16: Elimination inferences

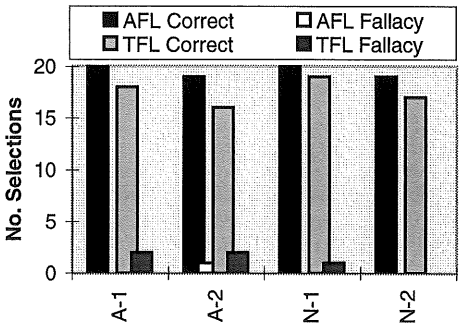

Figure 17: Introduction inferences

According to the laws of probability, the likelihood of a proposition "p" cannot exceed the likelihood of a conjunction "p *and* q" (Tversky and Kahneman, 1983). Inspection of participants' responses to the conjunctive introduction tasks suggests that most avoided committing this "conjunctive fallacy". The fact that a higher rate of participants committed the fallacy in the thematic group, however, could be explained by the fact that a conjunction of realistic terms sharing a plausibly valid relation is more likely to be endorsed than a conjunction of abstract terms sharing an unmeaningful relation. Notably, the tendency to commit this fallacy was strongest where a causal relationship seemed to exist between the two conjuncts. In "$current\_loc' = target\_loc? \wedge mission' = Success$" and "$applicant? \notin banned \wedge members' = members \cup \{applicant?\}$", for example, the presupposition of the truth of the first conjunct appears to be necessary for an adequate understanding of the second. In the fallacious conclusions "$print\_queue' = \langle\rangle \wedge \neg(printer\_status' = Online)$" and "$\neg(\#register' > MaxStudents) \wedge \neg(student' \in register')$", however, there does not appear to be such a degree of dependency which may account for their lower rates of selection.

Given a conjunction, "p *and* q", application of the formal rule of inference for conjunctive elimination allows us to conclude either one of the conjuncts, "p" or "q", in isolation. Judging by the high rates of correctness observed

for the elimination tasks, participants experienced little difficulty in drawing this inference, despite the variation of premiss polarity and the order of terms eliminated. This finding might be attributed to the expression of the tasks in formal logic and participants' prior experience with Gentzen style logical calculi. A brief summary of the results in relation to cognitive theories of conjunctive reasoning biases is given below - readers are referred to Vinter et al. (1997b) for a detailed discussion.

- Meaningful material did not appear to improve conjunctive reasoning performance in the same ways observed in previous studies of conditional reasoning (Gilhooly and Falconer, 1974; Griggs and Cox, 1982).
- Signs of the conjunctive fallacy (Tversky and Kahneman, 1983) became more apparent in thematic material, especially where there existed a "causal" relationship between two conjuncts.

7.3.1 A Model of Conjunctive Reasoning

A logistic regression analysis was used to model the data points generated during our study of conjunctive reasoning. Table 9 shows that the greatest variance in participants' correctness was accounted for, firstly, by the degree of meaningful content in the task material then, secondly, by the reasoner's level of expertise. A fit to the data (Classification-fit = 93.75%, Goodness-of-fit = 495.507) was achieved by excluding the following parameters: the type of inference to be drawn, the polarity of its premisses, and the position of the term denied or affirmed.

TABLE 9
Improvements made to the conjunctive model by stepwise addition of variables

Step	χ^2 Improvement	DF	Significance	Variable Added
1	8.190	1	0.00	Material Type
2	11.261	2	0.00	Expertise Level

The β estimates quantifying the degree of influence exerted by each of these variables on participants' correctness during our study of conjunctive reasoning are shown in Table 10.

TABLE 10
Parameters in the model of conjunctive reasoning

Factor	Parameter	β	SE	DF	Significance	β_x
Material-Abstract	$M1$	1.5017	0.41	1	0.00	0.7508
Expertise-Novice	$E1$	-2.4445	1.08	1	0.02	-1.6309
Expertise-Proficient	$E2$	-2.4482	1.05	1	0.02	
	Const	3.3331	0.41	1	0.00	

The general logit formula for predicting the level of inferential complexity associated with a Z conjunctive expression is as follows.

$$\text{logit}(Material, Expertise) = Const + \beta_{M1} + \beta_{E1} + \beta_{E2}$$

7.4 Results of the Quantified Reasoning Study

The frequencies of syllogisms solved correctly during the study of quantified reasoning are shown in Figure 18. Participants' levels of correctness under each experimental condition revealed overall rank orders as follows: TFL (90%) < AFL (93%) for group type, unmatching (91%) < matching (95%) for mood type, mixed polarity (89%) < two affirmatives (95%) < two negatives (96%) for mood polarity type, second (91%) = third (91%) = fourth (91%) < first (93%) for syllogistic figure, indeterminate (91%) < determinate (92%) for syllogistic determinacy, believable (90%) < unbelievable (92%) for conclusion type. Analyses of variance revealed that participants' levels of correctness were significantly affected by: determinate syllogisms ($F_{(14)} = 2.79, p < 0.01$), indeterminate syllogisms ($F_{(14)} = 3.75, p < 0.01$), matching moods ($F_{(8)} = 3.27, p < 0.01$), unmatching moods ($F_{(20)} = 2.55, p < 0.01$), two affirmative moods ($F_{(9)} = 2.86, p = 0.03$), first figure ($F_{(5)} = 1.96, p = 0.09$), second figure ($F_{(5)} = 3.73, p < 0.01$), third figure ($F_{(7)} = 4.09, p < 0.01$), fourth figure ($F_{(9)} = 1.89, p = 0.05$), and unbelievable conclusions ($F_{(9)} = 1.97, p = 0.05$). Analyses by linear regression failed to reveal any significant correlations between participants' levels of correctness and either their ratings of expertise or lengths of Z experience.

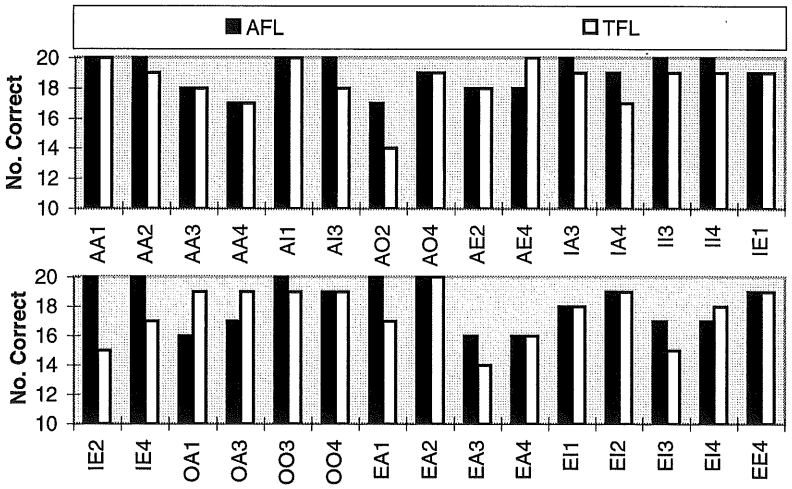

Figure 18: Syllogisms solved correctly by the two groups

The frequencies of thematic syllogisms with believable and unbelievable conclusions solved correctly are shown in Figure 19. The frequencies of strong and weak conclusions endorsed by participants in those tasks where both options were available are shown in Figure 20.

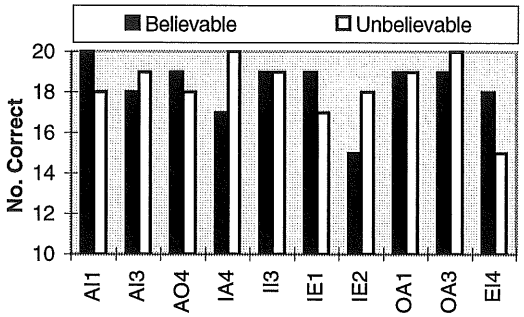

Figure 19: Believable versus unbelievable

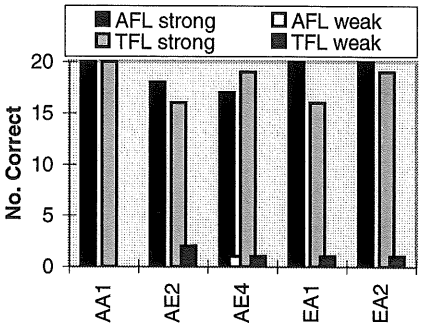

Figure 20: Strong versus weak

Although overall differences in group performance were not significant, evidence of improved reasoning performance under the abstract condition is evident in its higher overall mean rate of correctness and the fact that the abstract group achieved three times as many perfect scores for individual syllogisms. Wilkins (1928, p.77) attributes improved performance under abstract conditions to the "bad habits of everyday reasoning which are much in force in the familiar situation, but are not so influential when the material is symbolic or unfamiliar". The fact that more sporadic rates of correctness were observed in the thematic group suggests that meaningful material affected performance in certain tasks but not in others. This finding is supported by cognitive studies which suggest that the effects of thematic material are normally specific to the content of the individual task and the extent to which it relates to reasoners' prior beliefs (Barston, 1986; Traub, 1977). This is evident in the marked differences in correctness observed for the abstract and thematic versions of the following tasks: AO2, IE2, OA3 and EA1.

The theory of belief bias claims that people will tend to accept conclusions which they believe, and reject conclusions which they disbelieve, with little regard for their logical necessity (Barston, 1986; Begg and Harris, 1982; Janis and Frick, 1943; Revlin and Leirer, 1980). Evans et al., (1983) report overall rates of correctness as high as 97% when logic accords with belief and as low as 27% when logic conflicts with belief, while Revlin et al. (1980) report respective rates of 83% and 67%. Inspection of Figure 19 suggests that there were no discernible differences in performance for the ten syllogisms with abstract (93.5%), believable (91.5%) and unbelievable (91.5%) conclusions. This finding is supported by cognitive studies which report no significant difference in group performance under abstract and thematic conditions (Henle and Michael, 1956; Newstead, 1995). Analysis at the individual task level, however, suggests individual cases of performance facilitation or suppression caused by participants' beliefs towards the real world referents of the syllogistic terms. Although perfect scores were observed for the IA4 and OA3 tasks with unbelievable conclusions, the fact that similarly high rates of correctness were not observed for the other eight tasks leading to unbelievable conclusions lends some support to the belief bias hypothesis. Further support is gained from the marked differences in participants' correctness for the following thematic tasks leading to believable and unbelievable conclusions: IA4, IE2 and EI4.

The results of our study of quantified reasoning suggest that most of parti-

cipants' errors were not attributable to single, independent causes, but rather to the combination of several non-logical reasoning heuristics or biases. It is postulated that the factors which evoked errors differed between participants, and that the errors which participants committed on one task often did not generalise to others. Most of the observed non-logical responses are, however, consistent with cognitive theories of error in both syllogistic reasoning and everyday communicative experience. A brief summary of the results in relation to cognitive theories of syllogistic reasoning biases is given below - readers are referred to Vinter et al. (1997c) for a detailed discussion.

- Participants showed signs of improved reasoning under the abstract condition (Wilkins, 1928).
- Individual cases of reasoning facilitation and suppression under the thematic condition may have been caused by participants' beliefs towards the real world referents of the syllogistic terms (Evans et al., 1983; Revlin and Leirer, 1980).
- Some errors could be attributed to a failure to consider all possible representations of the given premisses (Erickson, 1974; Evans et al., 1993), especially since this may have involved more mental effort than many participants were willing to exert (Johnson-Laird and Bara, 1984) and some thematic premisses seemed to invite only one form of representation.
- Many errors could be attributed to participants' failure to consider A, E, I or O interpretations of considered representations, especially where this involved searching through numerous possible premiss combinations (Erickson, 1974; Johnson-Laird and Bara, 1984).
- Signs of atmosphere bias were limited in comparison to natural language based studies (Sells and Koob, 1937; Woodworth and Sells, 1935).
- No signs of matching bias (Evans, 1972a; Wetherick and Gilhooly, 1990) were evident, especially since most participants drew conclusions which did not match the forms of the given premisses.
- No signs of figural bias (Johnson-Laird and Steedman, 1978) were evident, although the finding that reasoning was most logical for first figure syllogisms is well supported (Dickstein, 1978a).
- Signs of illicit premiss conversion (Begg and Denny, 1969; Chapman and Chapman, 1959) were frequently evident but most prominent, perhaps, where participants appeared to convert the indeterminate premisses to forms from which determinate conclusions could be logically drawn.
- Adherence or non-adherence to Gricean conventions may have accounted for reasoning performance in the following ways:
  - Signs of the "Same M" fallacy (Chapman and Chapman, 1959; Dickstein, 1975; 1976) were evident in the large numbers of determinate conclusions drawn in response to indeterminate tasks.

- No evidence of caution bias (Woodworth and Sells, 1935) was found since nearly all participants gave strong categorical responses where weaker particular responses were also possible.
- Signs of non-logical pragmatic interpretations of the I and O syllogistic premisses (Begg and Harris, 1982; Newstead, 1995; Politzer, 1986) were evident in a series of background tasks given to participants before completing the main tasks.

7.4.1 A Model of Quantified Reasoning

A logistic regression analysis was used to model the data points generated during our study of quantified reasoning. Table 11 shows that the greatest variance in participants' correctness was accounted for, firstly, by the reasoner's level of expertise then, secondly, by the first premiss mood type then, thirdly, by the degree of meaningful content in the task material. A fit to the data (Classification-fit = 91.57%, Goodness-of-fit = 1362.8) was achieved by excluding the following parameters: figure type, second premiss mood type, and believability of the tasks' logical conclusions.

TABLE 11
Improvements made to the quantified model by stepwise addition of variables

Step	χ^2 Improvement	DF	Significance	Variable Added
1	24.476	2	0.00	Expertise Level
2	10.305	3	0.02	First Mood Type
3	6.897	1	0.00	Material Type

The β estimates quantifying the degree of influence exerted by each of these variables on participants' correctness during our study of quantified reasoning are shown in Table 12.

TABLE 12
Parameters in the model of quantified reasoning

Factor	Parameter	β	SE	DF	Significance	β_x
Material-Abstract	$M1$	0.5363	0.21	1	0.01	0.2682
Expertise-Novice	$E1$	0.6624	0.42	1	0.19	-0.0476
Expertise-Proficient	$E2$	-0.805	0.24	1	0.00	
First Mood-A	$F1$	-1.87	0.34	1	0.58	
First Mood-E	$F2$	-0.7561	0.33	1	0.02	-0.23
First Mood-I	$F3$	0.0233	0.36	1	0.95	
	Const	2.8894	0.16	1	0.00	

The general logit formula for predicting the level of inferential complexity associated with a Z quantified expression is as follows.

$$\text{logit}(Material, Expertise, First\ Mood) = \\ Const + \beta_{M1} + \beta_{E1} + \beta_{E2} + \beta_{F1} + \beta_{F2} + \beta_{F3}$$

7.5 Conversion to Absolute Probabilities

The model developed thus far provides a means by which the users of formal methods can predict the likelihood that the reasoner of a given expertise will draw an inference of a given type about a given type of logical statement expressed in a given degree of thematic material. At present the model yields logit values which appear to have little meaning in isolation. What we are lacking is a means for translating these values into absolute probabilities ($0 \leq p \leq 1$). The following formula performs the necessary translation (Norušis, 1996).

$$p = \frac{e^z}{1+e^z}$$

... where z is the logit value, and e is the exponential function.

8 A Brief Demonstration

We can envisage Will Wise, a senior software developer working on a defence based project, having been presented with the operational specification for a guided missile system, as shown in Figure 21. Supposing Will is asked by his team leader to determine the implications of the inclusion of schema *MissileStatus* within *MissileCheck*.

$$[COORDS]$$
$$MESSAGE ::= Hit \mid Miss$$

MissileStatus

$$current, target : COORDS$$
$$report : MESSAGE$$
$$report' = Hit$$

MissileCheck

$$\Xi MissileStatus$$
$$current \neq target \Rightarrow report = Miss$$

Figure 21: Thematic Z specification for a missile system

Given that the specification is expressed in realistic material, whose variable identifiers refer to rather fast moving animate objects, we can safely classify its material as being thematic in nature. Supposing Will had acquired a fair amount of Z experience by formally verifying part of a previous project, had studied several systems of logic at university and had even gone on expensive Z training courses run by the company, we might be inclined to regard Will as an

expert Z user. If we were to analyse the logic of the terms involved we would see that the consequent of a conditional rule is being denied, which suggests that Will is being invited to draw a modus tollens inference. We now have the three parameters that we need to apply our model of conditional reasoning: the material type (Thematic), the type of inference to be drawn (MT), and the Z expertise of the reasoner (Expert). The question that we must ask is: how likely is Will to infer the logically correct conclusion, *current* = *target*, under these conditions? Application of the model predicts that Will is 95.6% likely to draw this conclusion, which is calculated as follows.

To calculate logit($Thematic, AC, Expert$):
$z = Const - (\beta_{Mx} + \beta_{Ix} + \beta_{Ex})$

To translate into an absolute probability:
$p = e^z/(1 + e^z)$

Now suppose that the same specification and instructions had been given to Sam Slow, a new recruit and self-professed "novice" Z user. Suppose also that the specification given to Sam was not expressed in thematic material at all, but used single letters for variable names seemingly bearing little relation to real world objects, as shown in Figure 22.

$[C]$
$M ::= m1 \mid m2$

MS

$p, q : C$
$r : M$

$r' = m1$

MC

ΞMS

$p \neq q \Rightarrow r = m2$

Figure 22: Abstract Z specification for the missile system

How would these changes affect Sam's ability to infer the logical conclusion, $p = q$? In the absence of a suitable statistical method, most software engineers would make an subjective, educated guess based on their intuitive feelings towards the situation. The scope of the model is fortunately sufficient to account for such combinations of factor and can provide us with a much more quantifiably precise estimate. The question arises, however, of whether Sam's team leader would be prepared to risk the 35% differential in probability that Sam would not reach the same logical conclusion as Will, given the criticality of the inference. Now consider the revised version of our missile system's formal specification shown in Figure 23.

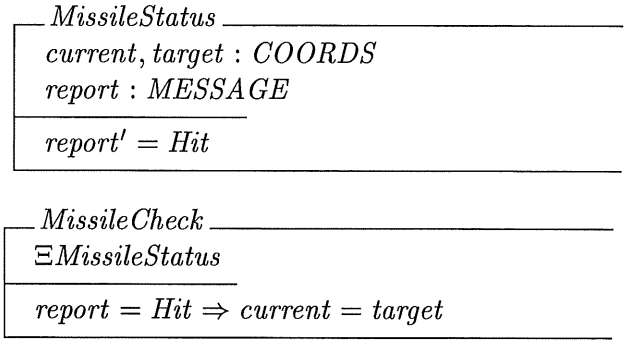

Figure 23: Revised Z specification for the missile system

If we were now to analyse the logic of the terms following the schema inclusion we would see that the antecedent of the conditional rule is being affirmed, which suggests that a much simpler modus ponens inference is required. Supposing Will and Sam are now asked to determine the implications of the schema inclusion, the model predicts that their potential for failing to draw the logical conclusion, *current* = *target*, has decreased to just 0.5% and 2.9% respectively. It is becoming evident that application of the model has strong implications for the ways in which formal specifications are written and the for the levels of expertise acquired by those people who work with them.

9 Model Evaluation

A popular methodology advocated for the procurement of software metrics begins by identifying those attributes which influence the quality of a product or process, formulating these in terms of a model, and then conducting empirical research to validate the model (Curtis, 1979; Fenton and Pfleeger, 1996). It is sometimes the case that the theoretical or empirical foundations for software metrics, however, are improperly considered prior to their formulation or are checked only as an afterthought. Roche (1994, p.80) claims that the "usual method involves developing a metric and then searching for some data for a validation study that often involves correlations between the metric values and some attribute that can be found to be correlated with the data!" The methodology used to develop our model differs from conventional approaches in that an initial empirical study gave rise to our theories about which attributes of a formal specification influence the development process (Vinter et al., 1996). It was also empirical research that generated the data which populates our metrics (Vinter et al., 1997a; 1997b; 1997c). So rather than construct a formal model and then subject it to empirical validation, our methodology proceeded in the converse direction by feeding data from empirical studies into a formal model.

It is argued in the software measurement literature that the evaluation of software metrics must be performed at both a theoretical and an empirical level (Sheppard and Ince, 1993). In simple terms, the former asks whether the correct model has been built, whereas the latter asks whether the model has been built correctly.

9.1 Theoretical Validation

The criteria specified by Sheppard and Ince against which a theoretical validation of software metrics may be performed, along with the extent to which our model satisfies these criteria, are described as follows.

1. *The model must conform to widely accepted theories of software development and cognitive science.* This criterion is satisfied insofar as the model rests upon the well supported theory from software engineering that errors in reasoning with software specifications are a potential source of software defects or anomalies (Fenton and Pfleeger, 1996; Potter et al., 1996), and the well supported theory from cognitive science that people are prone to error and bias when reasoning about specific types of logical statement in natural language (Braine, 1978; Evans et al., 1993).

2. *The model must be as formal as possible. In other words, the relationship between the input measurements and the output predictions must be precise in all situations. Furthermore, the mapping from the real world to the model must be made as formal as possible.* The model meets this criterion insofar as: it always generates the same output for a given combination of inputs, every valid combination of input parameter yields a deterministic output, and its predictions are always given in quantifiably precise numeric form.

3. *The model must use measurable inputs rather than estimates or subjective judgements. Failure to do so leads to inconsistencies between different users of the metric and potentially anomalous results.* The model meets this criterion insofar as the task of determining which values to input to the model is as intuitive as one could reasonably expect for a model of psychological complexity. For example, the "material type" parameter lends itself to whether the identifiers used in the logical terms involved in the inference have real world referents, the "inference type" parameter to the type of logical reasoning to be performed, and the "user expertise" parameter to the length and type of Z experience acquired by the reasoner. There is some room for inconsistency in users' assessment of which values to use as input parameters. Different users might not, for example, classify a given individual at the same expertise level. This kind of inconsistency might be overcome through adherence to simple guidelines.

4. *The ordering of model evaluations is intentional, since meaningful empirical work is of questionable significance when based upon meaningless models of software. Therefore, theoretical analysis of the properties of a model ought to precede validation.* The model meets this criterion insofar as its central underlying hypothesis is well-founded. The possibility that users of formal methods are liable to err in ways similar to those observed for the users of natural language is a reasonable one in light of recent cognitive findings. This hypothesis underlies the model whose worth is evident from its capacity to highlight potential sources of erroneous de-

velopment decisions, and its potential to provide empirical support for several of the claims associated with formal methods.

9.2 Empirical Validation

In order to justify the way in which a software metric is defined it is often necessary to seek independent and objective evidence which supports the credibility of its calculations. A common criticism of many systems is that the proposed measures are either totally unsupported by empirical evidence or that the methods used to validate them are flawed or inadequate (Card and Glass, 1990; Öry, 1993; Kitchenham, 1991). It is argued that several of the measures proposed by Halstead (1977), for example, are unreliable because: they are based on subjective personal belief or discredited psychological theories, there are flaws in the mathematical derivation of the formulas, the metrics do not scale up to larger programs, and design of the experiments used to validate the metrics was flawed (Coulter, 1983; Ince, 1989).

The method used to formulate the model has been advantageous in the sense that the research necessary for its empirical validation was performed during the model's formulation. In order to see this one only has to ask the question: how might one approach the empirical validation of the model or its underlying hypotheses? The answer is that one would run empirical experiments designed to test the extent to which the trained users of formal methods are prone to succumb to error and bias when reasoning about specific combinations of formal operator. But this is clearly something which has already been done, indeed, it is something we needed to do in order to generate the model. This is not to suggest, however, that a replication of our empirical studies would not be of value. Further empirical studies would help to refine the probability data which populates the model because the greater and more representative the samples which underly the model, the more accurate its predictions are likely to be. A discussion of how far the model meets Sheppard and Ince's criteria for an empirical validation now follows.

1. *The hypothesis under investigation.* When the aim of a model has not been clearly defined it can be unclear as to what is being validated, which can lead to statistically significant results being derived from an unusable model. The central hypothesis underlying our model is that the trained users of formal methods are liable to reason in similar ways to those observed for novice reasoners with logically equivalent expressions in natural language. The first criterion is met insofar as, during the course of the model's validation, we sought to answer a series of well-founded questions stemming from this hypothesis, based on existing knowledge from cognitive science.

2. *The artificiality of the data used.* Given that the data which populates the model is based on actual, rather than theoretical, instances of human reasoning by relatively large numbers of staff, students and professional users each with varying levels of expertise, it is representative of the complete range of formal methods' users. This provides for a degree of flexibility

in the model's predictions. The null hypothesis which we sought to test during validation was whether the trained users of formal methods are liable to succumb to similar non-logical errors as those observed for natural language based reasoners. Given that the results of our empirical validation could have shown users not to reason in these ways by failing to err, or by erring in different ways, the null hypothesis gave rise to a fair test of the model.

3. *The validity of the statistics employed.* The statistical tests used in the empirical validation of a model must be capable of refuting the hypothesis under investigation. The decision to use analyses of variance was dictated by the need to know which factors played a significant role in determining participants' reasoning performance. The decision to use regression based techniques was dictated by the need for quantifiably precise estimates of how far each factor contributed towards participants' reasoning performance. The statistics employed were therefore valid and appropriate for their purpose. Given that they could, and sometimes did, lead to the refutation of hypotheses relating to the precise combinations of formal operator prone to admit non-logical errors, these statistics were also applied in an objective manner.

Having evaluated the model against Sheppard and Ince's criteria at an empirical level, we now scrutinise the relation between between its predictions and the data generated during our empirical studies. We calculated earlier that the probability of drawing an TFL-AC conditional inference for an expert reasoner is 92% ($p = 0.9151$). If we were to calculate the inferential complexity for the same inference type and the same material but for lesser experienced Z users, we would expect to obtain lower probability values. These calculations yield p values of 0.9135 for a proficient user and 0.6460 for a novice user. This shows that there is an incremental effect on the model's predictions for users with increasing levels of expertise, and that the increment in p caused by an increment in one of the model's input parameters is far from being a uniform one, as one would expect. The same incremental effect for expertise level is observable in the conditional reasoning model's predictions across all four inference types, and both material types. This trend is consistent with the results of our conditional reasoning study which revealed significant correlations between participants' expertise levels and their levels of correctness. Intuitively, we would expect to see increasing relations by maintaining the same material and expertise type and changing inference type from MP to AC to MT to DA, or by maintaining the same inference and expertise type and changing material type from TFL to AFL. So according to the model, a user's chances of drawing a logically correct inference diminishes along with their level of expertise, the ease of the inference, or the amount of realistic material. The real worth of the model, however, lies in the fact that its predictions frequently do not suggest such rigidly defined rank orders of difficulty. Not only are the differences in the model's predictions intuitively linear and relative but they are entirely consistent with the results of our empirical studies.

10 Conclusions

> "Logic does not really contain the rules in accordance with which man actually thinks but the rules for how man ought to think. For man often uses his understanding and thinks otherwise than he ought to think and use his understanding. Logic thus contains the objective laws of the understanding and of reason" (Kant, in Young, 1992, p.13).

The experimental results suggest that, although the trained users of formal methods are often logical in their reasoning about formal expressions, they are liable, under certain conditions, to err in ways similar to those observed for untrained reasoners with the logically equivalent statements expressed in natural language. That the results point to the possible existence of non-logical encoding, processing and response biases suggests that the psychological causes of human error in formal contexts are deep-rooted. On the assumption that the specification process will always involve a certain degree of human input, the results suggest that the application of formal methods will always be vulnerable to the fallibility of human reasoning, and "if it is impossible to guarantee the elimination of errors, then we must discover more effective ways of mitigating their consequences in unforgiving situations" (Reason, 1990, p.148).

It is disconcerting to think that software developers will exhibit similar, or even increased, potentials for error in critical industrial projects, where alternative conclusions are rarely offered explicitly in the form of multiple-choice options, where formal expressions might contain more complicated combinations of logical operators, and where the repercussions of erroneous reasoning are much more serious than in laboratory based studies. The software community has been keen to emphasize the role of specification as a medium for communication (Barroca and McDermid, 1992; Imperato, 1991). In light of our research results, we would like to add the proviso that specifications are given minimal potential for admitting erroneous development decisions. It is only when we appreciate the negative repercussions that erroneous development decisions can have on software projects and the quality of their delivered products that we can begin to understand the need for capturing and verifying the reasoning processes of software developers.

> "A formal model of a system must be able to be represented in a manner which both elucidates the inferences which may be drawn from it and, where possible, captures the designers' intended interpretation. We note that such representations can provide invaluable support both to an expert, by making aspects of the model more immediate, and also to a non-expert, by providing a more tractable visualisation of it" (Gurr, 1995, p.395).

Although the cognitive processes involved in the creative process of writing formal specifications were not a direct focus of concern for this research, examination of the ways in which people interpret and reason about existing specifications has yielded implications for the ways in which specifications are written because poorly written specifications are more likely to admit errors of human

judgement. Owing to the logical nature of their underlying grammars, for each of the statements expressed in a formal notation it is nearly always possible to find an alternative expression which conveys the same meaning. The question is whether each of these alternatives have the same propensity for admitting human reasoning errors. It should be clear from the earlier demonstration that the model can be used as means for choosing between the alternatives. That is, the measures of inferential complexity yielded by it may be used as an independent form of justification for deciding which of the possible alternatives would be the "safest" to use in a given situation. Application of the model is likely to prove most beneficial, then, at the initial creative stage of the specification process when a designer frequently makes implicit discriminations of this kind and "there exists a multiplicity of potential designs for even the most trivial problem" (Sheppard and Ince, 1989, p.91).

By accepting as parameters the measures of attributes belonging to human users of formal methods, the model recognises that the grammatical properties of a formal specification are not the only determinants of a specification's "complexity". The validity of our decision to build a model of psychological, rather than computational, complexity is supported by Curtis' (1986, p.155) argument that "two different programmers can experience completely different levels of complexity in working with the same piece of software." Application of the model, then, has potential implications for the staff selected to work with formal methods because it helps to identify those whose development decisions are most likely to be influenced by intuitive heuristics and non-logical biases. Where the model consistently predicts high probabilities of error for particular staff, this might provide the justification for a management decision to select more appropriate staff or to encourage staff to undergo training.

The increasing interest in formal methods being shown by the software community (Bowen and Hinchey, 1994; Oakley, 1990) may be partly attributable to the belief that it is easier to reason about formal software specifications than conventional natural language based specifications (Thomas, 1995). The software community, however, has been slow to support this, and other claims pertaining to the use of formal methods, with empirical evidence (Craigen et al., 1995; Fenton, 1996). Although the model of inferential complexity developed in this paper is a tentative one and no claims are made regarding its suitability for direct industrial application, the methodology used during its formulation demonstrates an approach via which these claims can be subjected to independent and objective examination. We have used this approach as an empirical means for quantifying the extent to which the human potential for error is liable to remain after formalisation of the software specification process. Rather than being based on subjective personal belief, which might not accurately reflect reality, the lines of inquiry pursued in the present research stem from the well supported empirical findings of cognitive science. Rather than using isolated case studies from which it can be difficult to extrapolate results, we have borrowed standard experimental procedures from cognitive science in order to subject our theories to empirical scrutiny. It is argued that software measurement pursuits stand to benefit by taking on board correctly interpreted findings from psychological studies in this manner (Coulter, 1983; Ott, 1996).

This approach is advantageous over conventional approaches to software engineering research in that it generates theories "grounded" in the observed data (Glaser and Strauss, 1968) which can subsequently be used to refine initial hypotheses and to generate new theories in a fashion analogous to the Popperian "underlying pattern of continuous development" (Magee, 1985). The finalised theories and data that emerge from such a line of investigation, then, provide an empirical basis via which the psychological claims of the software community can be assessed.

> "When you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind: it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science" (Thomson, 1891, p.80).

Although our model is oriented towards identifying properties of formal Z specifications which are likely to cause errors or biases in human reasoning, there is no reason why similar models should be restricted to the Z notation, to formal methods, to the specification process, nor even to predicting sources of human reasoning difficulty. It is conceivable that similar models might be formulated to predict other potential sources of cognitive difficulty in the other products created by software engineering technologies, such as program code or design. Besides helping to quantify the levels of quality associated with such products, the formulation of similar psychological models might provide us with a better understanding of the factors leading to human error in the software engineering process, and provide a basis for taking corrective actions or refining the technologies around their human users.

The overall aim of the research reported here was to identify combinations of grammatical construct which are particularly susceptible to elicit errors and biases when people are reasoning with formal specifications. As compensatory measures are introduced, it is believed that this will help to reduce the potential for human error in the software development process. After all, if we know when and where errors are most likely to occur then erroneous development decisions can be pre-empted and the numbers of defects introduced into "finished" software systems reduced. In order to help us achieve our aim, we have borrowed from cognitive science the relevant theoretical knowledge and experimental methodology in order to determine the precise conditions under which trained users are particularly susceptible to error and bias when reasoning about formal Z specifications containing logical conditionals, disjunctives, conjunctives and quantifiers. In so doing we have demonstrated the feasibility of a cognitive approach to evaluating formal specifications which, we are convinced, is at least as important as the results themselves.

References

Ambler, A.L. (1977). Gypsy: A language for specification and implementation of verifiable programs. *ACM SIGPLAN Notices, 12*, 3, March 1977.

Bainbridge, J., Whitty, R.W. and Wordsworth, J.B. (1991). Obtaining structural metrics of Z specifications for systems development. In J.E. Nicholls (Ed.), *Proceedings of the Fifth Annual Z User Meeting, Oxford 1990*, 269-281, London: Springer-Verlag.

Balzer, R. and Goldman, N. (1986). Principles of good software specification and their implications for specification languages. In N. Gehani and A.D. McGettrick (Eds.), *Software Specification Techniques*, 25-39, Wokingham: Addison-Wesley.

Barden, R. and Stepney, S. (1993). Support for Using Z. In J.P. Bowen and J.E. Nicholls (Ed.), *Proceedings of the Seventh Annual Z User Meeting, London 1992*, 255-280, London: Springer-Verlag.

Barden, R., Stepney, S., Cooper, D. (1992). The Use of Z. In J.E. Nicholls (Ed.), *Proceedings of the Sixth Annual Z User Meeting, York 1991*, 99-124, London: Springer-Verlag.

Barroca, L.M. and McDermid, J.A. (1992). Formal methods: Use and relevance for the development of safety-critical systems. *The Computer Journal, 35*, 6, 579-599.

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

Begg, I. and Denny, J. (1969). Empirical reconciliation of atmosphere and conversion interpretations of syllogistic reasoning errors. *Journal of Experimental Psychology, 81*, 351-354.

Begg, I. and Harris, G. (1982). On the interpretation of syllogisms. *Journal of Verbal Learning and Verbal Behaviour, 21*, 595-620.

Bottaci, L. and Jones, J. (1994). *Formal Specification Using Z.* London: Thomson.

Bowen, J.P. (1988). Formal specification in Z as a design and documentation tool. *Second IEE/BCS Conference, Software Engineering 88*, 164-168, Conference Publication No. 290, July 1988.

Bowen, J.P. and Hinchey, M.G. (1994). Seven more myths of formal methods: Dispelling industrial prejudice. In T. Denvir, M. Naftalin and M. Bertran (Eds.), *FME'94: Industrial Benefit of Formal Methods*, 105-117, LNCS 873, London: Springer-Verlag.

Bowen, J.P. and Hinchey, M.G. (1995). Ten commandments of formal methods. *IEEE Computer, 28*, 4, 56-63.

Bowen, J.P. and Stavridou, V. (1993). Safety-critical systems, formal methods and standards. *Software Engineering Journal*, July 1993.

Braine, M.D.S. (1978). On the relation between the natural logic of reasoning and standard logic. *Psychological Review, 85*, 1-21.

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.

Braine, M.D.S. and Rumain, B. (1981). Development of comprehension of "or": Evidence for a sequence of competencies. *Journal of Experimental Child Psychology, 31*, 46-70.

Brien, S.M. and Nicholls, J.E. (Eds., 1992). *Z Base Standard. Version 1.0*, ZIP Project Technical Report ZIP/PRG/92/121, Oxford University.

Card, D.N. and Glass, R.L. (1990). *Measuring Software Design Quality.* London: Prentice-Hall.

Ceraso, J. and Provitera, A. (1971). Sources of error in syllogistic reasoning. *Cognitive Psychology, 2*, 400-410.

Chapman, L.J. and Chapman, J.F. (1959). Atmosphere effect re-examined. *Journal of Experimental Psychology, 58*, 220-226.

Cohen, B. (1989a). Justification of formal methods for system specification. *Software Engineering Journal*, 26-35, January 1989.

Cohen, B. (1989b). Rejustification of formal methods for system specification. *Software Engineering Journal*, 36-38, January 1989.

Cohen, B., Harwood, W.T. and Jackson, M.I. (1986). *The Specification of Complex Systems.* Wokingham: Addison-Wesley.

Cooke, J. (1992). Formal methods - mathematics, theory, recipes or what? *The Computer Journal, 35*, 5, 419-423.

Coulter, N. (1983). Software science and cognitive psychology. *IEEE Transactions on Software Engineering, SE-9*, 2, 166-171.

Craigen, D., Gerhart, S. and Ralston, T. (1995). Formal methods technology: Impediments and innovation. In M.G. Hinchey and J.P. Bowen (Eds.), *Applications of Formal Methods*, 399-419, Hemel Hempstead: Prentice-Hall.

Curtis, B. (1979). In search of software complexity. *Proceedings of the IEEE Workshop on Quantitative Software Models*, 95-106, October 1979.

Curtis, B. (1986). Conceptual issues on software metrics. In *Proceedings of the Nineteenth Hawai International Conference on Systems Sciences, 2*, 154-157, January 1986.

Curtis, B., Sheppard, S.B., Milliman, P., Borst, M.A. and Love, T. (1979). Measuring the psychological complexity of software maintenance tasks with the Halstead and McCabe metrics. *IEEE Transactions on Software Engineering, SE-5*, 2, March 1979.

Dawes, R.M. (1971). A case study of graduate admissions: Application of three principles of human decision making. *American Psychologist, 26*, 180-188.

Dawes, R.M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34*, 571-582.

Dawes, R.M. and Corrigan, B. (1974). Linear models in decision-making. *Psychological Bulletin, 81*, 95-106.

Dean, N. and Hinchey, M.G. (1996). Formal methods and modeling in context. In C.N. Dean and M.G. Hinchey (Eds.), *Teaching and Learning Formal Methods*, 99-116, London: Academic Press.

Dickstein, L.S. (1975). Effects of instructions and premiss order on errors in syllogistic reasoning. *Journal of Experimental Psychology: Human Learning and Memory, 104*, 376-384.

Dickstein, L.S. (1976). Differential difficulty of categorical syllogisms. *Bulletin of the Psychonomic Society, 8*, 330-332.

Dickstein, L.S. (1978a). Error processes in syllogistic reasoning. *Memory and Cognition, 5*, 537-543.

Dickstein, L.S. (1978b). Error processes in syllogistic reasoning. *Memory and Cognition, 5*, 537-543.

Dickstein, L.S. (1981). The meaning of conversion in syllogistic reasoning. *Bulletin of the Psychonomic Society, 18*, 3, 135-138.

Diller, A. (1994). *Z. An Introduction to Formal Methods.* Second edition, Chichester: Wiley and Sons.

Empson, W. (1965). *Seven Types of Ambiguity.* Harmondsworth: Penguin.

Erickson, J.R. (1974). A set analysis theory of behaviour in syllogistic reasoning tasks. In R.L. Solso (Ed.), *Theories of Cognitive Psychology: The Loyola Symposium.* Hillsdale: Erlbaum.

Erickson, J.R. (1978). Research on syllogistic reasoning. In R. Revlin and R.E. Mayer (Eds.), *Human Reasoning*, Washington DC: Winston.

Evangelist, W.M. (1983). Software complexity metric sensitivity to program structuring rules. *Journal of Systems and Software, 3*, 231-243.

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

Evans, J.St.B.T. (1972b). 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. (1983). 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., 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.

Fenton, N.E. (1991). The mathematics of complexity and measurement in computer science and software engineering. In J. Johnson and M. Loomes (Eds.), *The Mathematical Revolution Inspired by Computing*, 243-256, Oxford: Oxford University Press.

Fenton, N. (1992). When a software measure is not a measure. *Software Engineering Journal*, 357-362, September 1992.

Fenton, N. (1996). The empirical basis for software engineering. In A. Melton (Ed.), *Software Measurement*, 197-217, London: Thomson.

Fenton, N.E. and Kaposi, A.A. (1989). An engineering theory of structure and measurement. In B.A. Kitchenham and B. Littlewood (Eds.), *Software Metrics. Measurement for Software Control and Assurance*, 27-62, London: Elsevier.

Fenton, N.E. and Pfleeger, S.L. (1996). *Software Metrics. A Practical and Rigorous Approach.* London: Thomson.

Finney, K. (1996). Mathematical notation in formal specification: Too difficult for the masses? *IEEE Transactions on Software Engineering, 22*, 2, 158-159.

Garlan, D. (1996). Effective formal methods education for professional software engineers. In C.N. Dean and M.G. Hinchey (Eds.), *Teaching and Learning Formal Methods*, 11-29, London: Academic Press.

Gehani, N. (1986). Specifications: Formal and informal - a case study. In N. Gehani and A.D. McGettrick, *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.

Glaser, B.G. and Strauss, A.L. (1968). *The Discovery of Grounded Theory. Strategies for Qualitative Research.* London: Weidenfeld and Nicolson.

Goldberg, L.R. (1970). Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inferences. *Psychological Bulletin, 73*, 6, 422-432.

Grice, H.P. (1975). Logic and conversation. In P. Cole and J.L. Morgan (Eds.), *Syntax and Semantics, 3: Speech Acts*, 41-58, New York: Academic Press.

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.

Gurr, C. (1995). Supporting formal reasoning for safety-critical systems. *High Integrity Systems, 1*, 4, 385-396.

Guttag, J.V., Horning, J.J. and Wing, J.M. (1985). The Larch family of specification languages. *IEEE Software, 2*, 5, 24-36.

Halstead, M.H. (1977). *Elements of Software Science.* London: Elsevier.

Hall, A. (1990). Seven myths of formal methods. *IEEE Software*, 11-19, September 1990.

Hall, A. (1996). Using formal methods to develop an ATC information system. *IEEE Software*, 66-76, March 1996.

Hellinger, Z. (1995). *Frequently Asked Questions About RAISE.* Croydon: Lloyd's Register House.

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

Hinchey, M.G. and Bowen, J.P. (1995). Applications of formal methods FAQ. In M.G. Hinchey and J.P. Bowen (Eds.), *Applications of Formal Methods*, 1-15, Hemel Hempstead: Prentice-Hall.

Holloway, C.M. and Butler, R. (1996). Impediments to industrial use of formal methods. *IEEE Computer, 29*, 4, 25-26.

Hurford, J.R. (1974). Exclusive or inclusive disjunction. *Foundations of Language, 11*, 409-411.

Imperato, M. (1991). *An Introduction to Z.* Bromley: Chartwell-Bratt.

Ince, D. (1989). *Software Metrics.* In B.A. Kitchenham and B. Littlewood (Eds.), *Measurement for Software Control and Assurance*, 27-62, London: Elsevier.

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

Jack, A. (1992). It's hard to explain, but Z is much clearer than English. *Financial Times, 22*, 21st April 1992.

Jacky, J. (1989). Programmed for disaster: Software errors that imperil lives. *The Sciences*, 22-27, September/October 1989.

Jacky, J. (1997). *The Way of Z. Practical Programming with Formal Methods.* Cambridge: Cambridge University 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 Bara, B. (1984). Syllogistic inference. *Cognition, 16*, 1-61.

Johnson-Laird, P.N. and Steedman, M.J. (1978). The psychology of syllogisms. *Cognitive Psychology, 10*, 64-99.

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.

Jones, C.B. (1989). *Systematic Software Development Using VDM.* Second edition, Hemel Hempstead: Prentice-Hall.

Kahneman, D., Slovic, P. and Tversky, A. (1991). *Judgement Under Uncertainty: Heuristics and Biases.* Cambridge: Cambridge University Press.

Kaposi, A. (1991). Measurement theory. In J.A. McDermid (Ed.), *Software Engineer's Reference Book*, London: Butterworth-Heinemann.

Kaposi, A. and Myers, M. (1994). *Systems, Models and Measures.* London: Springer-Verlag.

Kern, L.H., Mirels, H.L. and Hinshaw, V.G. (1983). Scientists' understanding of propositional logic: An experimental investigation. *Social Studies of Science, 13*, 131-146.

Kitchenham, B.A. (1991). Metrics and measurement. In J.A. McDermid (Ed.), *Software Engineer's Reference Book*, London: Butterworth-Heinemann.

Kleinbaum, D.G. (1994). *Logistic Regression. A Self-learning Text.* New York: Springer.

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.

Lemmon, E.J. (1993). *Logic.* Second edition, London: Chapman and Hall.

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

Loomes, M. (1991). *Software Engineering Curriculum Design.* PhD thesis, University of Hertfordshire.

Loomes, M.J. and Vinter, R.J. (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*, 67-78, London: Springer-Verlag.

Loomes, M., Ridley, D. and Kornbrot, D. (1994). Cognitive and organisational aspects of design. In F. Redmill and T. Anderson (Eds.), *Proceedings of the Second Safety-Critical Systems Symposium, Birmingham, 1994*, 186-193, London: Springer-Verlag.

Macdonald, R. (1991). *Z usage and abusage*, Report No. 91003, RSRE, Ministry of Defence, Malvern, Worcestershire, UK, February 1991.

MacKensie, D. (1992). Computers, formal proof, and the law courts. *Notices of the American Mathematical Society, 39*, 9, 1066-1069, November 1992.

Magee, B. (1985). *Popper*, London: Fontana Press.

McCabe, T.J. (1976). A complexity measure. *IEEE Transactions on Software Engineering, SE-2*, 4, 308-320, December 1976.

Meehl, P.E. (1973). Psychodiagnosis. Selected Papers. London: Oxford University Press.

Melton, A.C., Gustafson, D.A., Bieman, J.M. and Baker, A.L. (1990). A mathematical perspective for software measures research. *Software Engineering Journal*, 246-254, September 1990.

Meyer, B. (1985). On formalism in specifications. *IEEE Software, 2*, 1, 6-26.

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. (1989). Interpretational errors in syllogistic reasoning. *Journal of Memory and Language, 28*, 78-91.

Newstead, S.E. (1995). Gricean implicatures and syllogistic reasoning. *Journal of Memory and Language, 34*, 5, 644-664.

Newstead, S.E. and Griggs, R.A. (1983a). The language and thought of disjunction. In J.St.B.T. Evans (Ed.), *Thinking and Reasoning. Psychological Approaches*, 76-106, London: Routledge.

Newstead, S.E. and Griggs, R.A. (1983b). Drawing inferences from quantified statements: A study of the square of opposition. *Journal of Verbal Learning and Verbal Behaviour, 22*, 535-543.

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

Nisbett, R.E. and Ross, L. (1980). *Human Inference: Strategies and Shortcomings of Social Judgement.* Englewood Cliffs: Prentice-Hall.

Nix, C.J. and Collins, B.P. (1988). The use of software engineering, including the Z notation, in the development of CICS. *Quality Assurance*, 103-110, September 1988.

Norušis, M.J. (1996). *SPSS Advanced Statistics 6.1.* Chicago: SPSS Incorporated.

Oakley, B. (1990). Opening address: The state of use of formal methods. In J.E. Nicholls (Ed.), *Proceedings of the Fourth Annual Z User Meeting, Oxford 1989*, 1-5, London: Springer-Verlag.

Öry, Z. (1993). An integrating common framework for measuring cognitive software complexity. *Software Engineering Journal*, 263-272, September 1993.

Ott, L.M. (1996). The early days of software metrics: Looking back after 20 years. In A. Melton (Ed.), *Software Measurement*, 7-25, London: Thomson.

Politzer, G. (1986). Laws of language use and formal logic. *Journal of Psycholinguistic Research, 15*, 47-92.

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. (1987). Content and context effects in reasoning. *American Journal of Psychology, 100*, 1, 41-60.

Potter, B., Sinclair, J. and Till, D. (1996). *An Introduction to Formal Specification and Z.* Second edition, Hemel Hempstead: Prentice-Hall.

Reason, J. (1990). *Human Error.* Cambridge: Cambridge University Press.

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

Revlin, R. and Leirer, V.O. (1980). Understanding quantified categorical expressions. *Memory and Cognition, 8*, 447-458.

Revlin, R., Leirer, V., Yopp, H. and Yopp, R. (1980). The belief-bias effect in syllogistic reasoning: The influence of knowledge on logic. *Memory and Cognition, 8*, 584-592.

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

Roberge, J.J. (1976a). Reasoning with exclusive disjunctive arguments. *Quarterly Journal of Experimental Psychology, 28*, 419-427.

Roberge, J.J. (1976b). Effects of negation on adults' disjunctive reasoning abilities. *Journal of General Psychology, 24*, 87-91.

Roberge, J.J. (1977). Effects of content on inclusive disjunction reasoning. *Quarterly Journal of Experimental Psychology, 29*, 669-676.

Roberge, J.J. (1978). Linguistic and psychometric factors in propositional reasoning. *Quarterly Journal of Experimental Psychology, 30*, 705-716.

Roche, J.M. (1994). Software metrics and measurement principles. *ACM SIGSOFT Software Engineering Notes, 19*, 1, January 1994.

Ross, W.D. (Ed. and Trans, 1949). Aristotle, *Aristotle's Prior and Posterior Analytics.* Oxford: Clarendon Press.

Rushby, J. (1995). Mechanising formal methods: Opportunities and challenges. In Bowen, J.P. and Hinchey, M.G. (Eds.), *Proceedings of the 9th International Conference of Z Users, Limerick, Ireland, September 1995, LNCS 967*, 105-113. London: Springer-Verlag.

Samson, W.B., Nevill, D.G. and Dugard, P.I. (1987). Predictive software metrics based on a formal specification. *Journal of Information and Software Technology, 29*, 5, 242-248.

Senders, J.W. and Moray, N.P. (1991). *Human Error: Cause, Prediction and Reduction.* Hillsdale: Erlbaum.

Sells, S.G. and Koob, H.F.A. (1937). A classroom demonstration of "atmosphere effect" in reasoning. *Journal of Educational Psychology, 28*, 514-518.

Sheppard, M. (1988). An evaluation of software product metrics. *Journal of Information and Software Technology, 30*, 3, 177-188.

Sheppard, M. (1990). Early life-cycle metrics and software quality models. *Information and Technology, 32*, 4, 311-316.

Sheppard, M. and Ince, D. (1989). Metrics, outlier analysis and the software design process. *Journal of Information and Software Technology, 31*, 2, 91-98.

Sheppard, M. and Ince, D. (1993). *Derivation and Validation of Software Metrics.* Oxford: Oxford University Press.

Slovic, P. and Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgement. *Organizational Behaviour and Human Performance, 6*, 649-744.

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

Spivey, J.M. (1992). *The Z Notation: A Reference Manual.* Second edition, London: Prentice-Hall.

Springer, C.H., Herlihy, R.E., Mall, R.T. and Beggs, R.I. (1966). *Statistical Inference.* Illinois: Richard Irwin.

Sullivan, J.E. (1975). Measuring the complexity of computer software. Technical Report MTR-2648, Vol. V, MITRE.

Szabo, M.E. (Ed. and Trans., 1969). G. Gentzen, *The Collected Papers of Gerhard Gentzen.* Amsterdam: North-Holland.

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.

Thomas, M.C. (1993). The industrial use of formal methods. *Microprocessors and Microsystems, 17*, 1, 31-36.

Thomas, M. (1995). Formal methods and their role in developing safe systems. *High Integrity Systems Journal, 1*, 5, 447-451.

Thomson, W. (1891). *Popular Lectures and Addresses.* Second edition, London: Macmillan.

Traub, B.H. (1977). A set theory approach to deduction with meaningful syllogisms. Unpublished PhD thesis, Ohio State University.

Turner, G.W. (1986). *Stylistics.* Harmondsworth: Penguin.

Tversky, A. and Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgement. *Psychological Review, 90*, 4, 293-315.

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

Vinter, R.J. (1996). *A Review of Twenty Formal Specification Notations.* University of Hertfordshire, Division of Computer Science, Technical Report No. 240.

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

Vinter, R.J., Loomes, M.J. and Kornbrot, D.E. (1997a). *Conditional Reasoning in Language and Logic: Transfer of Non-logical Heuristics?* University of Hertfordshire, Division of Computer Science, Technical Report No. 276.

Vinter, R.J., Loomes, M.J. and Kornbrot, D.E. (1997b). *A Study of Disjunctive and Conjunctive Reasoning in Formal Logic.* University of Hertfordshire, Division of Computer Science, Technical Report No. 298.

Vinter, R.J., Loomes, M.J. and Kornbrot, D.E. (1997c). *Quantified Reasoning in Formal Logic: Transfer of Everyday Errors and Biases?* University of Hertfordshire, Division of Computer Science, Technical Report No. 299.

Wetherick, N.E. and Gilhooly, K.J. (1990). Syllogistic reasoning: Effects of premise order. In K.J. Gilhooly, M.T.G. Keane, R.H. Logie, and G. Erdos (Eds.), *Lines of Thinking: Reflections on the Psychology of Thought. Volume 1. Representation, Reasoning, Analogy and Decision Making*, 99-108, Chichester: Wiley.

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

Woodcock, J. and Loomes, M. (1988). *Software Engineering Mathematics.* London: Pitman.

Woodworth, R.S. and Sells, S.B. (1935). An atmosphere effect in syllogistic reasoning. *Journal of Experimental Psychology, 18*,

Wordsworth, J.B. (1992). Formal methods and product documentation. In *Proceedings of FM'91*, Berlin: Springer-Verlag.

Young, M. (Ed. and Trans., 1992). I. Kant, *The Cambridge Edition of the Works of Immanuel Kant. Lectures on Logic.* Cambridge: Cambridge University Press.