Chapter 7

Quantified Reasoning

"*Barbara, Darii, Celarent, Ferio*

Camestres, Baroco, Cesare, Festino

Darapti, Datisi, Felapton, Ferison, Disamis, Bocardo

Bramantip, Camenes, Dimaris, Fesapo, Fresison"

(in Evans et al., 1993, p.211).[4]

The first two studies suggest that many of the erroneous heuristics and biases demonstrated during natural language based studies of reasoning with conditionals, disjunctives and conjunctives are also liable to occur in certain formalised contexts. This chapter reports the third main experiment which uses the framework of the syllogistic task to explore quantified reasoning in formalised contexts. Cognitive studies involving categorical syllogisms have shown people without logical training to exhibit a wide range of errors and biases when reasoning about natural language statements predicated by the "some" and "all" quantifiers. The main aim of the present study is to test whether users of formal methods are liable to err in simi-

[4]The mnemonic comprises nineteen names on four separate lines. Each name contains three vowels representing the "moods" for determinate syllogisms with "strong" conclusions. The line on which the name occurs represents the "figure" in which that syllogism is determinate. This terminology is explained in the following section.

lar ways when reasoning with logically equivalent Z expressions predicated by the formal existential ($\exists$) and universal ($\forall$) quantifiers.

7.1 Principles of Syllogistic Reasoning

The syllogistic task, developed by Aristotle (384-322 BC), is of special interest to cognitive science because it appears to encompass several core cognitive processes that pervade many aspects of human reasoning. The interpretation of premisses, the integration and representation of terms, the hypothetical postulation and evaluation of speculative conclusions, and the generation of responses are not cognitive processes which are specific to syllogistic reasoning (Evans et al., 1993; Dickstein, 1978b). Indeed, if psychology proves unable to account for the cognitive determinants of performance in the syllogistic task then it is difficult to see how it will ever be able to explain more complex cognitive functions (Johnson-Laird and Bara, 1984). Studies of syllogistic reasoning therefore provide important pointers to the cognitive processes involved in human reasoning generally and, in particular, the ways in which people reason with quantified statements.

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. Convention states that the major premiss must always precede the minor premiss. The aim of the syllogistic task is to use the two premisses as the basis for deducing a conclusion which describes a relation 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. These comprise the universal quantifiers "All" and "None", and the particular quantifiers, "Some" and "Some ... not". The quantifier which ranges over a syllogistic predicate reflects that predicate's "mood", conventionally abbreviated as shown in Figure 7.1. The determinate conclusion of a syllogism is said to be "strong" if it is quantified by one of the universal quantifiers, or "weak" if it is quantified by one of the particular quantifiers and a strong conclusion is also permissible.

Universal affirmative	All M are P	(A)
Universal negative	No M are P	(E)
Particular affirmative	Some M are P	(I)
Particular negative	Some M are not P	(O)

Figure 7.1: The four moods of syllogistic predicate

The order of terms in a syllogism's premisses is significant. As there are two possible orders for each of the major and minor premisses, this gives rise to four possible arrangements, or "figures", as shown in Figure 7.2. Although the order in which terms are presented within the two premisses might vary, the order 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 7.2: The four figures of a syllogism

Figure 7.3 shows a syllogism with the form AA1. Aristotle would consider this to be a "perfect" syllogism, that is, one whose necessity can be seen by novice reasoners without logical expertise (Adams, 1984), and "one that needs nothing other than the premisses to make the conclusion evident" (Aristotle, in Ross, 1949, p.287). Figure 7.4 shows a syllogism of the form EO1. The conclusion drawn here is fallacious because one cannot say for certain whether its relations follow necessarily

from the information specified in the given premisses. No logically valid determinate conclusion follows by necessity in this case.

All humans are mortal
All Greeks are humans

All Greeks are mortal

Figure 7.3: A "perfect" syllogism

No Greeks are immortal
Some men are not Greeks

Some men are not immortal

Figure 7.4: An invalid syllogism

7.2 Error and Bias in Syllogistic Reasoning

There is general agreement in the cognitive community that syllogistic reasoning involves at least three stages, all of which are prone to error and bias: premiss interpretation, premiss combination and response generation (Erickson, 1974; Evans et al., 1993). It is also argued that a fourth stage exists during which reasoners test speculative conclusions (Johnson-Laird and Steedman, 1978). In order that correct conclusions may be drawn it is imperative that reasoners adhere to deductive principles. Syllogistic studies suggest that reasoners are frequently prone to depart from such principles, however, and that there are dominant causes for their erroneous responses. The cognitive literature has been keen to speculate possible explanations for these trends.

Atmosphere Effects

According to "atmosphere theory" (Woodworth and Sells, 1935), syllogistic premisses create a global impression, or "atmosphere", depending upon how they are quantified and qualified. The quantity of a premiss can be universal ("all") or particular ("some"). The quality of a premiss can be affirmative ("are") or negative ("are not"). Atmosphere theory, as reformulated by Begg and Denny (1969), makes two specific predictions. First, whenever the quality of at least one premiss is negative, the quality of the conclusion drawn will be negative; when both premisses

are affirmative, the quality of the conclusion drawn will be affirmative. Second, whenever the quantity of at least one premiss is particular, the quantity of the conclusion drawn will be particular; when both premisses are universal, the conclusion drawn will be universal. In short, contemporary atmosphere theory predicts that, where the relationship between S and P is less than obvious, a reasoner will draw a conclusion which shares the same qualifiers and quantifiers as those contained in the given premisses, with little or no regard for the logic of the syllogism. It would appear that the atmosphere effect is not restricted to syllogistic reasoning. Sells (1936, p.7) argues that, in those situations where the range of possible solutions is limited, atmosphere bias leads reasoners to endorse the solution "most similar to the general trend or tone of the situation set up".

Implicit Conversion Theory

"Implicit conversion theory" argues that reasoners attempt to simplify complex premisses to forms that are more amenable to mental representation or processing (Revlin and Leirer, 1980). The construction of transitive relations, for example, between a conclusion's end terms, S and P, can clarify the form of conclusion to be drawn. Illicitly converted forms, however, can form a basis from which erroneous conclusions are drawn. Conversion is logically permissible for the I and E premiss forms because "Some S are M" can be replaced by "Some M are S", and "No S are M" can be replaced by "No M are S". Conversion of the A and O forms in this manner, however, is not logically permissible; "All S are M" does not necessarily imply "All M are S", and "Some S are not M" does not necessarily imply "Some M are not S". Studies suggest that reasoners often fail to recognise the conditions under which conversion is acceptable (Chapman and Chapman, 1959; Dickstein, 1981; Newstead and Griggs, 1983b; Politzer, 1990; Wilkins, 1928). It is claimed not only that reasoners have a tendency to convert syllogistic premisses, but that conversion

is actually the preferred method of interpreting premisses not given in transitive form (Revlis, 1975a; 1975b; Revlin and Leirer, 1980). If Revlis' claim were true then, with the exception of first figure syllogisms, reasoners would never attempt to reason from the same premisses that were presented to them! Strong evidence exists to suggest that Revlis' claim is too strong and does not reflect the way in which people normally approach the syllogistic task (Johnson-Laird and Bara, 1984; Newstead and Griggs, 1983b; Traub, 1977). Illicit conversion would, nevertheless, appear to account for many of the errors committed.

Figural Effects

Despite having no logical bearing on the syllogistic task, the possible psychological repercussions of manipulating term and premiss order has been a focus of concern. Although there is general agreement that changing premiss order alone does not influence reasoning performance significantly (Dickstein, 1975; Wetherick and Gilhooly, 1990), it is claimed that the order of terms within premisses is significant (Begg and Harris, 1982). Johnson-Laird and Steedman (1978), for example, report strong correlations between syllogistic figure and the types of conclusion drawn. "Figural bias" theory claims that syllogistic figure determines the order in which people relate end terms during premiss integration, and that a directional bias in our mental processes makes it easier to scan the represented information in certain directions (Johnson-Laird and Bara, 1984).

Determinacy Bias

It is argued that "determinacy bias" misleads reasoners into interpreting or combining premisses in ways that can only lead to determinate conclusions, or causes them to discount hypothetical possibilities that lead to indeterminate conclusions (Dickstein, 1975; 1978b; Revlis, 1975a). In other words, reasoners would generally

prefer to draw a valid conclusion rather than say that nothing follows from a given premiss pair. Some experimenters attribute this response bias to the disproportionate number of determinate and indeterminate arguments that tend to occur in the syllogistic task. Assuming 64 possible premiss combinations, as in Dickstein's studies, less than one third lead to determinate conclusions. Errors therefore become attributable to reasoners' expectations that a greater proportion of the given premiss pairs will lead to determinate conclusions.

Set-theoretic Representations

It is argued that people reason about syllogisms in ways analogous to those in which set-theoretic tools, such as Venn diagrams or Euler circles, are used in mathematics (Adams, 1984; Ceraso and Provitera, 1971; Erickson, 1978; Traub, 1977). A central tenet of this argument is that the interpretation of premisses involves creating a combined mental representation showing the set relations that may exist between terms. In order that the correct conclusion may be deduced, every possible combination of set relation that follows from a given premiss pair must be explored. Non-logical errors then become explainable as a consequence of reasoners' use of inappropriate representations or their failure to consider all hypothetical combinations. The possible Euler representations that are consistent with individual syllogistic premisses are shown in Figure 7.5 (adapted from Erickson, 1974, p.310; Evans et al., 1993, p.220). It should be noted that the combination of premisses gives rise to many more representations than those shown here. That reasoners seem inclined to consider only a few of these is perhaps understandable given the mental effort that it would require to consider the entire set and the demanding nature of the syllogistic task.

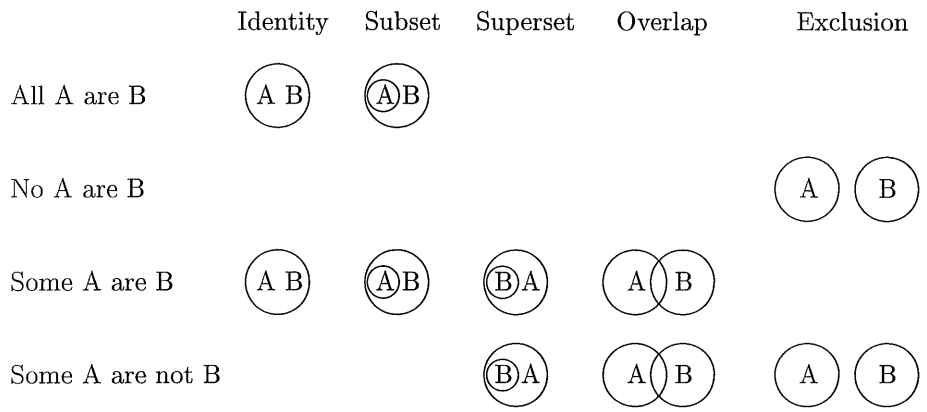

Figure 7.5: Possible Euler representations of syllogistic predicates

The quantifiers "some" and "some ... not", as they appear in natural language based syllogistic studies, are ambiguous because the individual is not told whether to adopt an everyday (partitive) interpretation or a logical (partitive, but possibly universal) interpretation. Support for this hypothesis is gained from the results of Chapman and Chapman (1959), who propose that the qualifier "are" in syllogistic predicates encourages reasoners to assume an identity relation, "is equal to", between terms when an inclusion relation, "is included in", would be more appropriate from a set-theoretic perspective, and that this encourages unwarranted assumptions of symmetry between terms, in a manner similar to that demonstrated by Tsal (1977). The experimenters ascribe assumptions of this form to reasoners' prior experience of elementary mathematical algebra or geometry where identity relations are commonplace. One might expect that the substitution of formal operators, with precise mathematical meanings, for the supposedly ambiguous qualifiers, "are" and "are not", would dispel any such ambiguities from the task and cue reasoners into interpretations which conform with the dictates of logic rather than conventions of everyday language.

Analogical Representations

It is argued that people reason about syllogisms as if constructing symbolic analogical representations of premiss information (Johnson-Laird and Steedman, 1978). According to this theory, the "classes" in syllogistic premisses are represented by imagining arbitrary instances of their exemplars. Although no explicit claims regarding the symbols or their interrelations are given, the four possible premiss relations might be represented as shown in Figure 7.6, where "↓" indicates a link, "⊥" indicates a negative link, and parentheses indicate a possible exemplar.

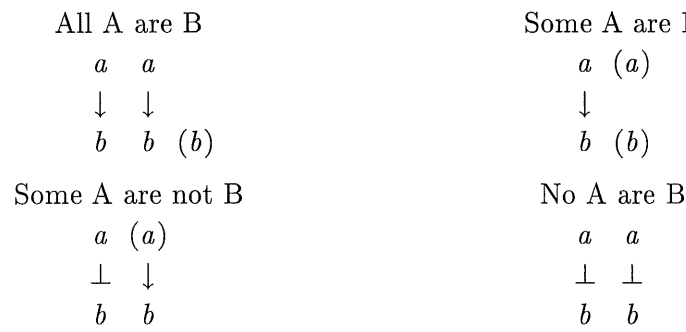

Figure 7.6: Analogical representations of syllogistic predicates

Syllogistic reasoning by analogy involves four stages: semantic representation of the given premisses, heuristic combination of each premiss' representation, formulation of a conclusion from the premiss combination, and testing of the conclusion (which can lead to modified representations). Like set-theoretic models, analogical theories are informal in the sense that they make no specific claims for the order in which representations are constructed or the tests performed.

Content Effects and Belief Bias

The view that formal logic abstracts away all extraneous details and allows reasoners to concentrate solely on the underlying logical form of arguments derives from Kantian philosophy (Kant, in Smith, 1993). Based on this assumption, one

might expect people to reason more logically when a task is expressed in abstract, as opposed to thematic, content. Studies of conditional reasoning, however, suggest that performance improves as logical tasks become less abstract (Dominowski, 1995; Gilhooly and Falconer, 1974; Van Duyne, 1974; Wason and Shapiro, 1971). It is claimed that the apparent facilitatory effects of thematic content can, in some cases, be explained by the possibility that reasoners simply read off responses from memory without performing the kind of logical analysis appropriate to the task (Griggs and Cox, 1982). These findings suggest that, despite the strictly logical requirements of laboratory based tasks, the semantic associations of thematic content can encourage reasoners to favour non-logical heuristics based on guesswork.

Studies of the syllogistic task report that thematic premiss content which elicits close associations with information stored in reasoners' semantic memories is more likely to elicit responses that accord with prior belief, albeit sometimes at the expense of logical necessity (Begg and Harris, 1982; Janis and Frick, 1943). The theory of "belief bias" claims that reasoners accept believable conclusions uncritically and only resort to logical analysis when premisses suggest unbelievable conclusions (Barston, 1986; Evans et al., 1993; Revlis, 1975a). Belief bias effects tend to become more evident as task content becomes more closely related to the personal beliefs of reasoners, because firmly held convictions are likely to be held in spite of evidence against them (Morgan and Morton, 1944). Specifically, it is argued that reasoning performance deteriorates when logic and pragmatic beliefs point towards different conclusions, but improves when they concur (Revlin and Leirer, 1980; Wilkins, 1928). The claims for belief bias are not, however, upheld universally. Other studies report no discernible differences in correctness for syllogisms phrased in abstract material and logically equivalent forms in thematic material leading to believable conclusions (Henle and Michael, 1956; Newstead, 1995).

Influence of Pragmatic Knowledge

Pragmatic laws and conventions guide our interpretation of written and spoken language, enabling us to look beyond what is said explicitly in order to gain an appreciation of a speaker's intentions (Levinson, 1983). It is argued that syllogistic errors are often attributable to reasoners' attempts to "treat logical statements as if they are obscure attempts at communication, and interpret them by the same conventions they would use in normal discourse" (Begg and Harris, 1982, p.596). Many recent findings from syllogistic studies have been discussed in relation to Grice's (1975) seminal work on conversational implicature.

Grice's "Cooperative Principle" aims to explicate some of the universally accepted rules and conventions which govern everyday spoken and written communication. Under this principle there are four maxims. First, the maxim of quantity states that speakers should make their contribution as informative as is required for its purpose and to not withold information they know to be true. Second, the maxim of quality states that speakers should only say that which they believe to be true and supported by adequate evidence. Third, the maxim of relation states that speakers should keep the content of their contribution as relevant as possible. Fourth, the maxim of manner states that the contribution made by speakers should be clear and unambiguous. There is much evidence to suggest that reasoners' predisposition to apply Gricean conventions contributes to their downfall in the syllogistic task (Begg and Harris, 1982; Newstead, 1989; 1995; Politzer, 1986).

The theory of the "Same M" fallacy claims that, whenever the subject and predicate of a speculative conclusion are related by a seemingly common middle term (the same M), reasoners will accept this conclusion at face value according to the maxim of relation, irrespective of its logical necessity (Chapman and Chapman, 1959; Dickstein, 1975; 1976). If in everyday conversation one were to say "Some

politicians are lazy, and some lazy people are wealthy", then the listener is clearly being invited to conclude "Some politicians are wealthy". The speaker could otherwise be accused of violating the maxim of relation and being deliberately deceitful. Although this form of probablistic inference often leads to correct conclusions in everyday experience, it does not follow logically because the middle term in each premiss might not necessarily refer to the same class members. A similar phenomenon to the "Same M" fallacy is reported by Dickstein (1978b), who argues that, in premisses where no relations are specified between S and M, and between P and M, a reasoner might still draw a conclusion from S to P because both premisses seemingly share the common property of not being related to M. The maxim of relation can therefore explain the tendency to give determinate conclusions where none are warranted, especially from II and OO premisses, because it is assumed that experimenters would not intentionally make two unrelated statements in sequence.

In ordinary speech it is normally taken for granted that speakers abide by the Gricean maxim of quantity and divulge as much useful information as necessary; they will not say "some" when "all" is applicable, and they will not say "some ... not" when "no" is applicable. In ordinary speech, therefore, the particular quantifier "some" is given the partitive interpretation "at least one, but not all", and "some ... not" is given the partitive interpretation "at least one is not, but not none". The syllogistic task, however, sometimes requires reasoners to entertain counter intuitive notions such as "Some apples are fruits", even when they know that in fact "All apples are fruits". A failure to comply with this requirement is evident in the results of Woodworth and Sells (1935), who report a non-logical "caution bias", that is, a tendency to accept "Some ... are" more readily than "All ... are", and "Some ... are not" more readily than "None ... are". This inclination to accept weak conclusions, when a stronger version might exist, suggests that reasoners often fail to consider hypothetical possibilities and that they are generally conservative estimators.

7.3 Aims and Methodology

In summary, it appears that many of the errors observed during the syllogistic task are attributable to participants' application of similar rules and conventions to those which govern their communication of quantified statements in everyday language. The present study aims to test whether the trained users of formal methods are liable to commit the same kinds of error when reasoning about the formal quantifiers, "$\exists$" and "$\forall$", as those observed for their English counterparts, "some" and "all". It aims also to identify those particular syllogistic forms which give rise to systematic reasoning errors. Since formal notations are not governed by the same linguistic principles which govern everyday communication in natural language, one would not expect the same non-logical tendencies to transfer over into the formal domain, especially when the tasks are presented explicitly in formal logic and all participants have the relevant logical training.

In order to help explicate any differences between reasoners' treatment of the quantifiers from natural language and formal logic, the well established framework of the syllogistic task is employed. The variables investigated are the type of syllogism (comprising mood, figure, strength and determinacy) and the degree of thematic content used. Having noted that participants were particularly susceptible to error when reasoning with counter intuitive material during the initial study, a third experimental variable is included; the believability of the conclusion to be inferred.

7.3.1 Participants

A total of forty computing scientists volunteered to take part in the experiment. These comprised staff and students from academic institutions and computing professionals from industrial software companies, all of whom were recruited by personal invitation. All participants were native English language speakers and were

randomly selected. Participants were divided equally into two linguistic groups: Abstract Formal Logic (AFL) and Thematic Formal Logic (TFL). The groups were loosely matched, first, according to participants' personal ratings of Z expertise and, second, according to their lengths of Z experience. The AFL group comprised 15 staff, 3 students and 2 professionals. Their mean age was 34.65 years ($s = 8.79$) and all had studied a system of formal logic beforehand. Their mean level of Z experience was 5.84 years ($s = 4.55$). According to their personal ratings of expertise, the group comprised 8 expert, 11 proficient and 1 novice users of the Z notation. The TFL group comprised 13 staff, 1 student and 6 professionals. Their mean age was 33.25 years ($s = 9.79$) and all had studied a system of formal logic beforehand. Their mean level of Z experience was 4.43 years ($s = 3.89$), and the group comprised 5 expert, 10 proficient and 5 novice users.

7.3.2 Design

The study had a three factor mixed design. The first, between groups, factor was the degree of realistic material, abstract or thematic, corresponding to the two linguistic groups, AFL and TFL. The second, repeated measures, factor was the type of syllogism and comprised 30 levels. Various mood, figure and strength combinations were tested within this factor. The third, repeated measures, factor was the believability of the conclusion to be inferred and had two levels which applied only to the TFL group: intuitive, and counter intuitive.

A systematic variation of 16 moods, 4 figures and 2 levels of believability would normally yield 128 possible thematic tasks. For the practical purposes of this study, however, the tasks included only a representative sample from this range of possible syllogism types. The abstract tasks comprised 30 syllogisms (15 with determinate and 15 with indeterminate conclusions). 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 7.1 shows the forms of syllogism presented during the present study.

TABLE 7.1
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 given in parentheses.

In Dickstein's (1978a) study, where a systematic variation of sixteen moods and four figures yields 64 premiss combinations, this gives rise to 19 possible determinate conclusions. Owing to strong typing imposed by the Z notation, however, four of the premiss pairs which normally lead to determinate conclusions led to indeterminate conclusions in the present study.[5] Although it would have been possible

[5]The four tasks AA3, AA4, EA3 and EA4 lead to determinate conclusions in natural language studies, yet lead to indeterminate conclusions in the present study. This is because the terms of a formalised syllogism must be assigned Z types, or mathematical sets, and that any two universal premisses cannot give rise to a particular conclusion when the premiss terms might be assigned to empty sets. The following example contrasts natural language and Z versions of an EA3 syllogism. The conclusion in the Z version is indeterminate because the possibility that $Food = \varnothing$ acts as a counter example to any possible determinate conclusion.

No oranges are apples	$\neg \exists f : Food \bullet orange(f) \land apple(f)$
All apples are fruits	$\forall f : Food \bullet apple(f) \Rightarrow fruit(f)$
Some fruits are not oranges	No determinate conclusion

to overcome these restrictions and achieve a design that would allow for the same number of determinate conclusions to be drawn as in Dickstein's study, this would not have been possible without compromising the complexity or the consistency of the manner in which the tasks were presented, either of which would have jeopardised the practicality of the study. The design therefore gives rise to a lower ratio of determinate to indeterminate syllogisms; approximately 2:1. This is worthy of note because it is argued that a bias towards determinate responses may be introduced as a consequence of the extreme imbalance between determinate and indeterminate tasks (Chapman and Chapman, 1959; Dickstein, 1975; 1976; Revlis, 1975a; Traub, 1977). A more balanced design in this respect, therefore, has favourable implications for achieving unbiased responses.

7.3.3 Materials

To simplify the formalisation of the syllogistic task, a methodical approach was used to translate natural language based categorical syllogisms into logically equivalent forms in Z. Two obstacles had to be overcome in this respect. First, it was necessary to find formal operators which corresponded to the natural language quantifiers and qualifiers without altering the logical structure of the original task. Second, it was necessary to assign appropriate types to the variables, or "terms", of the formalised syllogism so as to avoid violating Z's strong type checking rules. Figure 7.7 shows the method used to translate the four possible forms of natural language premiss into equivalent forms in Z. All task sheets were computer generated.

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

Figure 7.7: Z translations of the four syllogistic predicates

So as to minimise the possibility of interference from prior knowledge for the abstract tasks, arbitrary single letter identifiers were used in place of meaningful function names. In order to facilitate the recall of relevant information from memory, meaningful identifiers were used for function names in the thematic versions of the tasks. These names were chosen to refer to concepts with which participants would be familiar including: social groups, occupations, animals, foods and materials. The experiment's materials are exemplified by the abstract and thematic versions of the AA1 syllogistic task shown in Figures 7.8 and 7.9 respectively.

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

(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 7.8: Abstract AA1 task

$\forall p : Person \bullet human(p) \Rightarrow mortal(p)$
$\forall p : Person \bullet Greek(p) \Rightarrow human(p)$

(a) $\exists p : Person \bullet Greek(p) \wedge mortal(p)$
(b) $\forall p : Person \bullet Greek(p) \Rightarrow mortal(p)$
(c) $\neg \exists p : Person \bullet Greek(p) \wedge mortal(p)$
(d) No valid conclusion

Figure 7.9: Thematic AA1 task

In order to test for possible effects of participants' personal beliefs on their reasoning performances, five thematic tasks were designed to lead to believable conclusions and five logically equivalent tasks were designed to lead to unbelievable conclusions. The nature of these tasks are exemplified by Figure 7.10, in which the reasoner is required to draw a believable conclusion corresponding to "No rich people are poor", and Figure 7.11, in which the reasoner is required to draw an unbelievable conclusion corresponding to "Some communists are capitalists".

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

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

Figure 7.10: Believable EA1 task

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

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

Figure 7.11: Unbelievable IA4 task

7.3.4 Procedure

Before starting the experiment, 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 personal rating of their Z expertise (novice, proficient or expert), and a description of any systems of formal logic studied beforehand.

Before completing the main tasks, participants were asked to show their understandings of the four possible forms of quantified formal expression by completing four corresponding "background tasks". Each task prompted the participant to select the closest natural English translation of a given formal Z expression corresponding to one of the four possible forms of syllogistic expression: A, E, I or O. It was hoped that the background tasks would help to explain some of the erroneous trends that might arise in participants' responses during the main syllogistic tasks. Participants were then shown the following instructions.

> "In each of the tasks that follow, you will be shown two Z predicate expressions taken from an operational schema. You may assume that all of the named functions have been defined. You will be asked to determine which one of four given statements follow from the information given. Please circle the letter of your choice. You will then be asked to give a confidence rating, which should indicate how far you believe your answer to be correct. Please complete all tasks to the best of your ability, without reference to textbooks. The experiment should take around one hour to complete."

For each main task participants were shown two Z predicates representing the premisses of a categorical syllogism, three Z predicates representing possible determinate conclusions, labelled "(a)" to "(c)", and a fourth predicate representing a possible indeterminate conclusion, labelled "(d)". Participants were asked to select the one conclusion that followed from the given premisses by circling the appropriate letter, then to give a rating of the extent to which they believed their response

was correct by ticking one of the corresponding boxes shown below. These boxes were coded 1, 2 and 3 respectively for the purposes of analysis. Task sheets were distributed to participants and completed anonymously then mailed back to the experimenter (these are shown in Appendix A). All participants were tested on an individual basis.

Confidence rating: ☐ Not confident ☐ Guess ☐ Confident

7.4 Results

Background Tasks

Table 7.2 contains the four background tasks presented prior to the main tasks. For each task, four possible natural language translations are shown along with the response rates for each option.

TABLE 7.2
Frequencies of selections during the background tasks ($N = 40$)

$\forall t : T \bullet A(t) \Rightarrow B(t)$		$\exists t : T \bullet A(t) \wedge \neg B(t)$	
*All As are Bs	38	At least one A is not a B	29
At least one (possibly all) As are Bs	0	*At least one (possibly all) As are not Bs	9
Possibly all As are Bs	2	Exactly one A is not a B	0
Some As are Bs	0	Some As are not Bs	2
$\exists t : T \bullet A(t) \wedge B(t)$		$\neg \exists t : T \bullet A(t) \wedge B(t)$	
At least one A is a B	29	*None of the As are Bs	40
*At least one (possibly all) As are Bs	9	At least one (possibly none) of the As are Bs	0
Exactly one A is a B	0	Possibly none of the As are Bs	0
Some As are Bs	2	Exactly one A is not a B	0

Note: Unambiguous set-theoretic translations are marked with an asterisk.

A series of one way chi-square tests revealed that participants' selections significantly differed from chance in all cases: "All" ($\chi^2_{(3)} = 104.80, p < 0.01$),

"Some" ($\chi^2_{(3)} = 52.60, p < 0.01$), "Some ... not" ($\chi^2_{(3)} = 52.60, p < 0.01$), and "None" ($\chi^2_{(3)} = 120.00, p < 0.01$). The table suggests that nearly all participants succeeded in selecting natural language translations corresponding to unambiguous set-theoretic interpretations of the two universal expressions, quantified by "$\forall$" (All) and "$\neg\,\exists$" (None), but nearly three quarters of participants failed to select unambiguous set-theoretic translations of the two particular expressions, quantified by "$\exists$" (Some) and "$\exists \ldots \neg$" (Some ... not).

Group Correctness

A one way between factors analysis of variance revealed no significant differences in overall group correctness for the TFL group ($\bar{x} = 90\%$) and AFL group ($\bar{x} = 93\%$). Inspection of Figure 7.12, however, shows that there were 12 individual syllogisms with perfect scores for the AFL group, but only 4 perfect scores for the TFL group. A two way chi-square analysis revealed that this result significantly differed from chance ($\chi^2_{(1)} = 5.45, p = 0.02$).

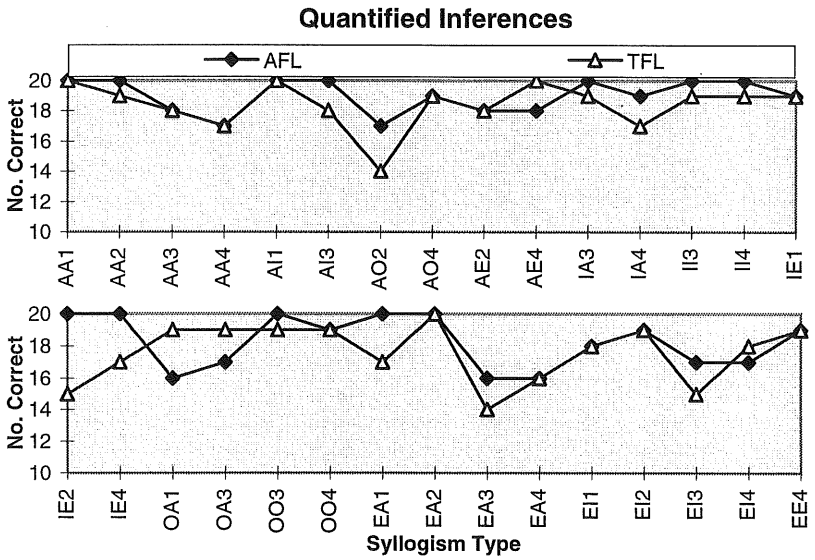

Figure 7.12: Frequencies of quantified syllogisms solved correctly ($n = 20$)

Syllogism Type

A one way repeated measures analysis of variance revealed a significant effect of the 30 syllogism types on participants' correctness ($F_{(1,29)} = 2.78, p < 0.01$). Hence, several planned comparisons were performed within this factor concerning: mood, figure, determinacy, believability, and strength of conclusion.

Mood and Figure

A planned comparison contrasting correctness for syllogisms with matching moods ($\bar{x} = 91\%$) and unmatching moods ($\bar{x} = 95\%$) was significant ($F_{(1,29)} = 9.17, p < 0.01$). A planned comparison contrasting correctness for syllogisms with two affirmatives ($\bar{x} = 95\%$) and the other mood combinations was significant ($F_{(1,29)} = 10.70, p < 0.01$). A planned comparison contrasting correctness for syllogisms with just one negative mood ($\bar{x} = 89\%$) and the other mood combinations was significant ($F_{(1,29)} = 18.26, p < 0.01$). A planned comparison contrasting correctness for syllogisms with two negative moods ($\bar{x} = 96\%$) and the other mood combinations approached significance ($F_{(1,29)} = 3.68, p = 0.06$).

A series of planned comparisons contrasting correctness for syllogisms in the first ($\bar{x} = 93\%$), second ($\bar{x} = 91\%$), third ($\bar{x} = 91\%$) and fourth figure ($\bar{x} = 91\%$) revealed no significant effects.

Determinacy, Believability and Strength of Conclusion

A planned comparison contrasting correctness for syllogisms with determinate conclusions ($\bar{x} = 92\%$) and indeterminate conclusions ($\bar{x} = 91\%$) was not significant. A total of 61 erroneous determinate conclusions were given in response to 35 indeterminate tasks, and 45 indeterminate conclusions were given in response to 35 determinate tasks.

A series of planned comparisons contrasting correctness for those ten syllogisms with believable conclusions ($\bar{x} = 92\%$), unbelievable conclusions ($\bar{x} = 92\%$), and abstract conclusions ($\bar{x} = 94\%$) revealed no significant effects. The frequencies of thematic syllogisms with believable and unbelievable conclusions solved correctly are shown in Figure 7.13.

A planned comparison contrasting the rates of selection for strong and weak conclusions, in those tasks where both options were possible, was significant ($F_{(1,9)} = 1560.53, p < 0.01$). A comparison of the rates of strong and weak conclusions drawn is shown in Figure 7.14.

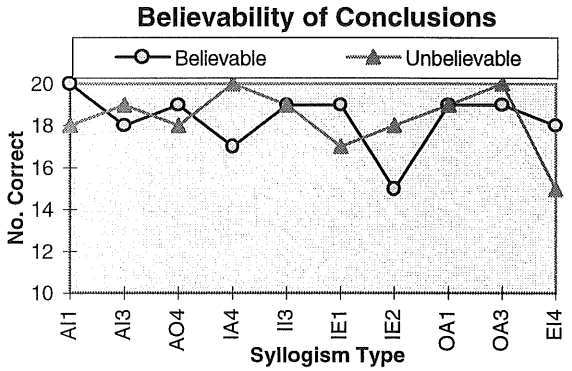

Figure 7.13: Believable and unbelievable ($n = 20$)

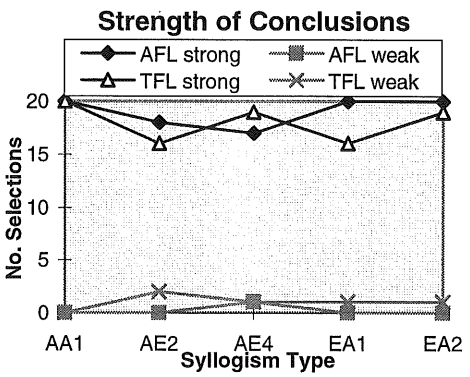

Figure 7.14: Strong and weak ($n = 20$)

Experience and Expertise

A linear regression analysis revealed no significant correlations between participants' levels of correctness and their ratings of expertise, their levels of experience or their ages. This suggests that participants' increased levels of experience or expertise with the Z notation was not related to their levels of performance.

Confidence Ratings

A one way between factors analysis of variance revealed no significant effects of linguistic group type on participants' confidence. The mean confidence

ratings for group type were as follows: AFL (2.79), TFL unbelievable (2.93), TFL believable (2.94). A series of planned comparisons for syllogisms with matching moods, two affirmatives and two negatives revealed no significant effects. A series of planned comparisons contrasting confidence for syllogisms in the first, second and third figure revealed no significant effects. But a planned comparison contrasting syllogisms in the fourth figure with those in the other figures was significant ($F_{(1,29)} = 8.50, p < 0.01$). A planned comparison contrasting confidence for syllogisms with determinate conclusions and indeterminate conclusions was significant ($F_{(1,29)} = 22.05, p < 0.01$). A planned comparison contrasting confidence for syllogisms with believable and unbelievable conclusions revealed no significant effects. That participants were highly confident in the correctness of their responses is evident from the mean confidence ratings shown in Figure 7.15.

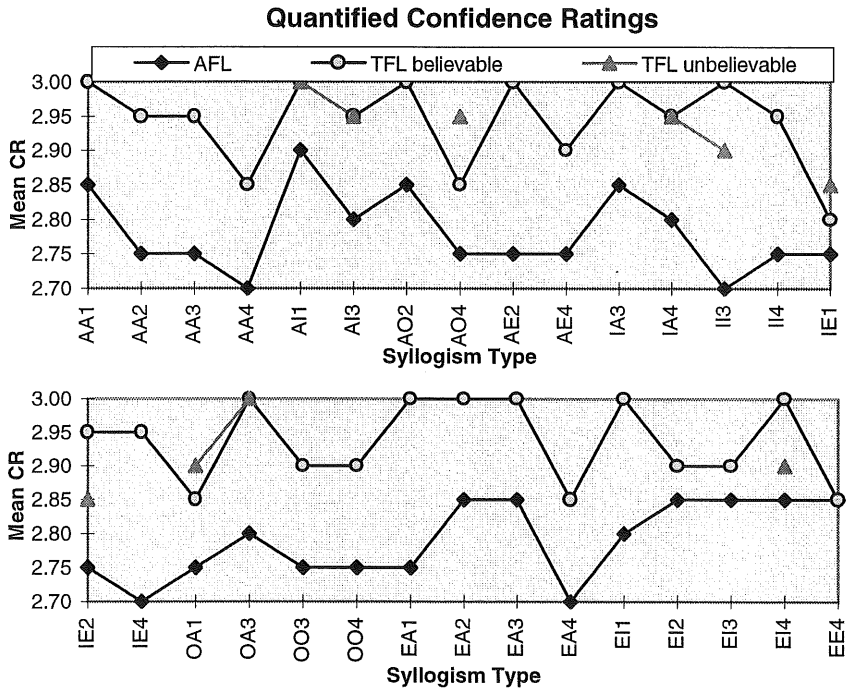

Figure 7.15: Confidence ratings for quantified inferences ($1 \leqslant CR \leqslant 3$)

7.5 Discussion

Influence of Pragmatic Knowledge

The results for the background tasks suggest that nearly all participants drew the most precise, set-theoretic interpretations of the universal statements corresponding to A and E premisses. This suggests that participants adhered to the Gricean maxim of quantity because they preferred to say "all" where "possibly all" and "some" were also possible, and "none" where "exactly one is not" and "possibly none" were possible. High rates of precise set-theoretic interpretations of universally quantified statements are also reported by Neimark and Chapman (1975).

The Gricean maxim of quantity appeared to lead participants away from the strongest possible set-theoretic translation for the particular I and O formal statements. A particular affirmative statement, "$\exists\, t : T \bullet A(t) \wedge B(t)$", does unquestionably entail the assertion "At least one A is a B", as endorsed by nearly three quarters of participants. But it also entails the possibility that "All of the As might be Bs", which, providing participants had abided by the Gricean maxim of manner and made their interpretations as informative and unambiguous as was necessary for the purposes of the study, would have led them to the strongest set-theoretic interpretation, "At least one (possibly all) As are Bs", as endorsed by only 22.5% of participants. Similarly, the meaning of a particular negative, "$\exists\, t : T \bullet A(t) \wedge \neg B(t)$", does entail the assertion "At least one A is not a B". But it also entails the possibility that "All of the As might not be Bs", as endorsed by only 22.5% of participants.

Participants' seemingly ambiguous interpretations of the formal statements with particular moods may be attributable to two possible causes. First, it may be ascribed to people's strong inclination to draw partitive interpretations of the "some" and "some ... not" quantifiers in everyday communication, where, according to

Gricean convention, the truth of one should imply the truth of the other (Newstead, 1989). Second, it may be ascribed to the way in which the existential quantifier is introduced in textbooks teaching the Z notation. Several popular undergraduate texts (for example: Diller, 1994; Potter et al., 1996) state that the "$\exists$" quantifier can be paraphrased as the English expression "there is" or "there exists", and is easily memorisable from its "reverse E" symbology. But these texts fail to mention that, unless explicitly predicated not to do so, the existentially quantified element of a set might refer to the entirety of the set that it represents. So when an individual perceives the "$\exists$" quantifier, he or she is inclined to assume a singular, rather than multiple referent, "there exists one or several, but not all", which contrasts with the "$\forall$" quantifier, where "all", or "for every", is definitively asserted. But when an individual perceives a statement beginning with "$\neg\,\exists$", he or she is inclined to assume the equally definitive "there does not exist", or "none".

The Gricean maxim of relation can explain reasoners' tendency to give determinate responses to indeterminate syllogisms because it is assumed that experimenters would not intentionally make two consecutive statements without there existing some relation between them. Similarly, the theory of determinacy bias claims that reasoners expect a greater proportion of determinate tasks than there actually are, and that this expectation contributes to their downfall on indeterminate syllogisms (Revlis, 1975a). Numerous studies confirm that reasoning improves for tasks with determinate rather than indeterminate conclusions (Dickstein, 1976; 1978b; Evans et al., 1983). Roberge (1970) reports 51.2% correctness for determinate syllogisms versus 35.8% correctness for indeterminate syllogisms, whilst Dickstein (1975) reports 72.6% and 58.2% respectively. The higher rates of correctness in the present study might be ascribed to the explicitly logical nature of the tasks and participants' experience of mathematical logic, where it appears to be part of the accepted norm that any two consecutive statements may be unrelated.

The responses to the abstract syllogisms AA4, EA3, EA4 and OA1, and thematic syllogisms AA4, IE2, IE4, EA3 and EA4, show that up to one quarter of participants mistakenly endorsed determinate responses when indeterminate ones were appropriate. This apparent tendency to perceive logical relations between logically unrelated premisses is consistent with predictions which stem from the maxim of relation and determinacy bias, but only in the case of these few tasks. The fact that no significant effects of determinacy were found overall might be attributed to the greater proportion of determinate tasks in the present study, which may have curbed participants' predisposition to give determinate responses. It is interesting to note that participants' application of the maxim of relation did not cause systematic errors on those tasks where this was particularly expected. It was hypothesised that the effects of the "Same M" fallacy would become most evident in responses to II syllogisms, where the middle terms seemingly share the common property of being related to both end terms, and OO syllogisms, where the middle terms seemingly share the common property of being unrelated to both end terms (Chapman and Chapman, 1959; Dickstein, 1975; 1976). However, only three people gave responses consistent with these trends for the II3, II4, OO3 and OO4 tasks, and most others gave indeterminate responses.

People's shared pragmatic knowledge of the Gricean maxim of quantity encourages them to divulge as much useful information as necessary in ordinary conversation. They will not say "some" when "all" is applicable, and they will not say "some ... not" when "no" is applicable. Participants' willingness to apply the maxim of quantity, even in formalised contexts, is evident in their responses to the main experimental tasks. Figure 7.14 shows that a total of only six weak conclusions were endorsed where one hundred stronger versions were possible. This significant preference for universal conclusions also counts against Woodworth and Sells' (1935) theory of "caution bias". It may be worthy of note that five of these weak conclusions

were chosen by the thematic group, which suggests that the degree of meaningful content can influence the strength of conclusion endorsed by reasoners.

Content Effects and Belief Bias

Evidence that the abstract group outperformed the thematic group is reflected in the fact that the latter achieved three times as many perfect scores for individual syllogisms. Wilkins (1928, p.77) ascribes 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 similarly high mean rates were observed for the two groups is also supported throughout the cognitive literature (Henle and Michael, 1956; Newstead, 1995). It is suggested that, when reasoners' beliefs are not held with a sufficient degree of conviction, or are indifferent to the real world referents of the task, they are unlikely to distort logical reasoning. Performance is likely to be similar for abstract and thematic content under such conditions.

The more sporadic rates of correctness observed within the thematic group suggests that the presence of meaningful content affected performance in some tasks but not in others. This is supported by findings which suggest that any facilitatory or inhibitory effects caused by changed material are entirely specific to the task and the extent to which its content relates to the reasoner's prior beliefs (Barston, 1986; Traub, 1977). Evans et al. (1983) report rates of correct inference as high as 97% when logic accords with belief and as low as 27% when logic conflicts with belief, Revlin et al. (1980) report respective rates of 83% and 67%. The fact that the mean scores for those ten syllogisms with abstract, believable and unbelievable conclusions were much higher and more evenly balanced in the present study suggests that the beliefs elicited by the chosen thematic materials were not sufficiently strong to lead reasoners away from logical rules.

Mood and Figure Effects

Inspection of Figure 7.12 shows that the lowest scores in both linguistic groups were for syllogisms with just one negative premiss mood. At least one fifth of participants gave erroneous responses to the abstract syllogisms EA4, OA1 and OA3, and the thematic syllogisms AO2, IE2, EA1, EA4, EI3, EA3, and EI4 (unbelievable). Participants achieved higher rates of correctness on syllogisms with matching premiss moods, even when both were negative. These findings are consistent with results from our study of disjunctive reasoning, which supports the claim that premisses containing one negative term are more difficult than those containing two (Evans and Newstead, 1980; Roberge, 1976b; 1978).

Atmosphere bias theory makes several specific predictions: AA premisses yield A conclusions, AE, EA or EE premisses yield E conclusions, II, AI or IA premisses yield I conclusions, and OO, AO, OA, EI or IE premisses yield O conclusions (Sells, 1936; Simpson and Johnson, 1966; Woodworth and Sells, 1935). The results offer mixed support for these predictions. The perfect scores observed for the following syllogisms suggest that performance was facilitated where logic and atmosphere theory pointed to the same conclusion: abstract tasks AA1, EA1, AI1, IA3, and thematic tasks AA1, AI1, IA4 (unbelievable), OA3 (unbelievable). There were cases where many participants failed to draw the correct conclusion, however, even where this was dictated both by logic and atmosphere theory: abstract tasks AE4, AO2, EI3 and EI4, and thematic tasks EA1, AO2, EI3 and EI4 (unbelievable). For those cases where logic and atmosphere theory pointed to different conclusions, the high rates of correctness suggest that logic exerted a dominating influence on reasoning. Only the responses to the abstract EA3, OA1 and OA3 tasks and thematic EA3 task, are consistent with the predictions of atmosphere bias. These findings do not concur with those of Sells and Koob (1937), for example, who report cases in which

atmosphere bias seemingly led to error rates exceeding 90%. Such errors might be attributed to the severe time pressures imposed by the experimenters, however, which could have dissuaded participants from conducting full logical analyses of the tasks. The fact that there was no strict time limit imposed during the present study might therefore be partly responsible for participants' increased logicality.

Cognitive studies report significant differences in reasoning performance under each of the four syllogistic figures (see for example: Erickson, 1974; Johnson-Laird and Steedman, 1978). The fact that the rates of correctness were generally higher and much more evenly balanced across the four figures in the present study suggests that figure did not account for the same degree of variation in participants' responses. We must be careful not to generalise from this between studies comparison, however, because the subset of syllogistic tasks varied in each study. Although the performance differential under the four figures was not significant, possibly owing to a ceiling effect, the higher mean rate of first figure syllogisms supports the hypothesis that performance may be facilitated by first figure syllogisms (Dickstein, 1978a; Johnson-Laird and Bara, 1984), where the correct conclusion can be exposed simply by scanning the given premisses in a forwards direction.

Implicit Conversion

Implicit conversion theory proposes that errors can arise as a result of reasoners' attempts to simplify given premisses into forms more amenable to representation or reasoning (Revlin and Leirer, 1980). Natural language based studies suggest that illicit conversion of universal affirmatives, in particular, is responsible for many errors (Newstead, 1989; Newstead and Griggs, 1983b). Illicit conversion of the A premiss in the indeterminate thematic EA3 task might explain why one quarter of participants endorsed determinate non-logical E responses. This is supported by the theory of "conversion by addition" which claims that reasoners are inclined to

convert "All A are B" to "All A are B and all B are A" (Dickstein, 1981) or to "All B are A" (Politzer, 1990). Conversion in this case may have been facilitated by participants' social knowledge pertaining to the task materials, which appears almost to invite an illicit conversion of the indeterminate EA3 task into a determinate EA1 task. The corresponding natural language forms of these tasks are illustrated in Figures 7.16 and 7.17 respectively. The slightly lower rate of E conclusions given in response to the abstract version of the same task might therefore be attributed to the fact that conversion of the A premiss was not invited by the intuitive plausibility of the relation between terms in the resulting conclusion.

No churchgoers are atheists	No churchgoers are atheists
All churchgoers are devout people	All devout people are churchgoers
Nothing	No devout people are atheists

Figure 7.16: Original EA3 task

Figure 7.17: Converted EA1 task

It is postulated here that illicit conversion of a universal affirmative was also responsible for the four incorrect A responses to the abstract OA1 task. The fact that high rates of correctness were observed for both thematic versions of the same task suggests that the thematic relations created by conversion of the A premiss in these cases may have ran contrary to participants' prior beliefs and blocked any attempts to draw determinate conclusions. That conversion for these tasks would have led to the counter intuitive assertions "All birds are owls" and "All mammals are dogs" supports this hypothesis. Evidence of illicit conversion blocked by counter intuitive real world associations is reported throughout the cognitive literature (Ceraso and Provitera, 1971; Evans et al., 1983; Newstead, 1990; Revlis, 1975a; Revlin et al., 1980; Tsal, 1977). When premisses are couched in abstract material, a reasoner is unlikely to have strong dispositions towards the terms and is liable to regard them as being interchangeable with alternative forms. Figures 7.18 and 7.19 show how

illicit conversion appeared to evoke erroneous responses to the abstract OA1 task, where thematic connotations were not sufficiently strong to block conversion.

Some B are not C	Some B are not C
All A are B	All B are A
Nothing	Some A are not C

Figure 7.18: Original OA1 task Figure 7.19: Converted OA3 task

The results suggest that the likelihood of an illicit conversion being accepted in formalised contexts depends upon two factors. First, it may depend upon the degree of perceivable symmetry that exists between the end terms of a syllogism. This was evident in the EA3 task where participants seemingly converted the given premisses into the first syllogistic figure before generating putative conclusions. Participants who believe that a premiss pair is not presented in an ordered and symmetrical manner will try to convert it to another syllogistic figure, where relations between the end terms in the conclusion are more readily apparent. This is supported by results from natural language based studies (Begg and Harris, 1982; Dickstein, 1978a). Second, the likelihood of a conversion being accepted appears to depend not only on the degree of thematic material used, but on whether this material establishes believable relations in converted premisses or putative conclusions, according to participants' conceptions of the real world. This was evident in the thematic EA3 task, where conversion led to thematic relations which conformed with popular social beliefs, and in the OA1 task, where conversion was seemingly blocked when it led to forms which contradicted popular zoological knowledge. Given that participants' errors are only ascribable to illicit conversion in several isolated cases, however, the results do not support Revlin and Leirer's (1980) hypothesis that conversion is a routine part of the syllogistic task. It seems worthy of note that the syllogisms in which conversion appears to have caused most errors lead to logically indeterminate conclusions. It is therefore possible that many non-logical conversions were endorsed

as the consequence of a general bias towards determinate conclusions.

Set-theoretic Representations

The results are now interpreted in light of Erickson's (1974; 1978) theory that people represent and process premiss information in ways analogous to those in which Euler circles or Venn diagrams are used in mathematics. Inspection of those tasks for which three or more participants gave the same non-logical response suggests two marked trends which could account for many errors. Where the syllogism is indeterminate, errors may be attributed to participants' failure to find counter examples to putative conclusions. The systematic process of constructing representations and searching for counter examples, however, can require more mental effort than a reasoner is willing to expend (Barston, 1986; Johnson-Laird and Bara, 1984). This trend appears evident in participants' responses to the abstract OA1 task and thematic IE2 and IE4 tasks. Figure 7.20 suggests how failure to consider a counter example may have led one quarter of participants to the erroneous conclusion, "No drunkards are scientists."

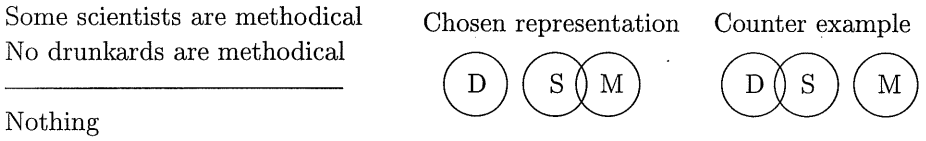

Figure 7.20: Set-theoretic representations of the IE2 syllogism

Where a syllogism is determinate, errors may be attributed to participants' failure to adopt the appropriate A, E, I or O interpretation of correctly represented premiss combinations. This trend is evident in the results for the thematic syllogisms AO2, EA1, EI3 and EI4 (unbelievable). One might expect participants not to recognise all five of the valid set-theoretic representations that follow from combined EI3 premisses, but to respond based on only a subset of these possible representations, given the effort that a full analysis would require (as shown in Figure 7.21).

This hypothesis may account for the five participants who failed to derive the correct interpretation, "Some conductors are not woods", which is consistent with all five possible representations. This finding supports the view that the difficulty of a syllogism increases along with the number of ways in which its premisses can be represented (Ceraso and Provitera, 1971; Erickson, 1974; Johnson-Laird and Steedman, 1975). It also suggests that participants experienced particular difficulty in drawing particular negative interpretations of represented premisses.

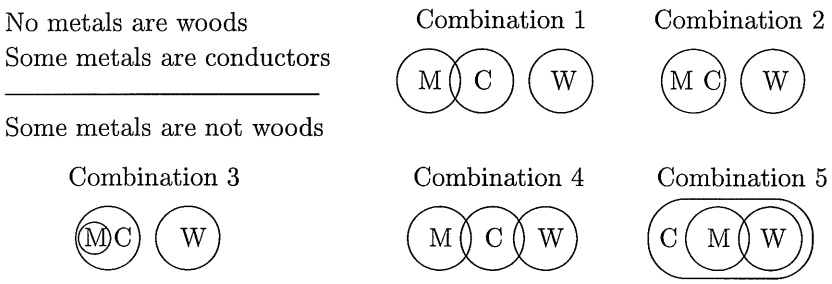

Figure 7.21: Possible set-theoretic representations of the EI3 premisses

Analogical Representations

A Spearman rank order comparison between the rates of correctness obtained by the present study and those by Johnson-Laird and Steedman (1978) revealed a significant correlation for the 30 tasks common to both studies ($r = 0.37, p = 0.05$). According to Johnson-Laird and Steedman, the form of conclusion generated from analogical representations depend upon a heuristic relating to the polarity of links in represented paths: one negative link yields an O conclusion, two negative links yields an E conclusion, one positive link yields an I conclusion, two positive links yields an A conclusion, otherwise the conclusion will be indeterminate. This heuristic is now discussed in relation to several tasks which elicited systematic errors.

Figures 7.22 and 7.23 show combined premiss representations for the thematic EI3 task and thematic AO2 task, where "−" indicates a negative path, "+" indicates a positive path, and "?" indicates an indeterminate path. Application of the heuristic predicts, in both cases, an O conclusion for the first path and an indeterminate conclusion for the second. This prediction is born out in the results: 80.0% O conclusions and 17.5% indeterminate responses for the EI3 task, and 77.5% O conclusions and 17.5% indeterminate responses for the AO2 task.

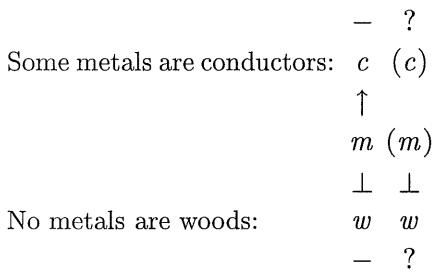

Figure 7.22: Thematic EI3

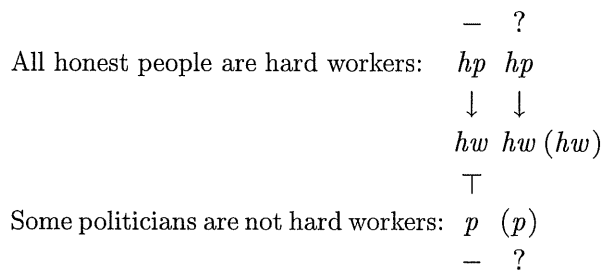

Figure 7.23: Thematic AO2

Figures 7.24 and 7.25 show symbolic analogical representations for the abstract and thematic EI4 tasks. Again, the 83.3% O conclusions and 16.7% indeterminate responses to the abstract EI4 task, and the 83.3% O conclusions and 16.7% indeterminate responses to the thematic EI4 task, suggest that participants applied the heuristic to one represented path only and subsequently failed to conduct exhaustive logical testing - the fourth stage in Johnson-Laird and Steedman's theory. It is interesting to note that participants appeared to experience particular difficulties with representations from which particular negative or indeterminate conclusions could be drawn from the represented paths. It is also interesting to note that the abstract and thematic EI4 tasks give rise to analogical representations with the same basic structure. This may account for the similar types and rates of response elicited by these tasks.

	–	?
No A are B:	a	a
	$\perp$	$\perp$
	b	(b)
	$\downarrow$	
Some B are C:	c	(c)
	–	?

Figure 7.24: Abstract EI4

	–	?
No disloyal people are married:	dp	dp
	$\perp$	$\perp$
	m	(m)
	$\downarrow$	
Some married people are traitors:	t	(t)
	–	?

Figure 7.25: Thematic EI4

Confidence Ratings

Participants' susceptibility to error, in spite of their high confidence ratings, suggests that many were overconfident in the correctness of their responses, particularly in the thematic group. A noteworthy link between correctness and confidence appears evident. Although the abstract group outperformed the thematic group overall, the mean confidence rating for every thematic task is higher than the corresponding rating for its abstract counterpart, with only one exception. One might expect a reasoner's confidence to increase along with their correctness. Given that the abstract group were consistently more correct but less confident than the thematic group, the results run contrary to this expectation. This trend may be attributable to the recognition of familiar everyday terms which led the thematic group to believe that non-logical everyday heuristics were sufficient for the tasks at hand, and the recognition of purely symbolic terms which led the abstract group to believe that a logical approach was more appropriate. As the use of non-logical heuristics is perceived to involve a less mentally intensive analysis as that required by a purely logical approach, this may explain the differences in group confidence.

Possible Explanations for Errors

A list of the possible causes for participants' non-logical responses is given in Table 7.3. It seems worthy of note that many of these heuristics and biases are hy-

pothesised in the cognitive literature to exert a central and dominating influence on human reasoning processes under experimental conditions or in everyday experience, and that the predictions of some are more specific than others.

TABLE 7.3
Possible causes of error in the study of quantified reasoning

1	Atmosphere bias
2	Figural bias
3	Z formalisation of the syllogistic task
4	Belief bias
5	Gricean maxims of quantity or relation
6	Determinacy bias
7	The "Same M" fallacy
8	Caution bias
9	Illicit premiss conversion
10	Inaccurate set-theoretic or analogical interpretations
11	Incomplete testing of represented premisses

Table 7.4 contains English translations of those abstract syllogisms in which three or more participants endorsed the same non-logical conclusion and gives a speculative list of possible causes for these errors, according to participants' responses. Inspection of the table suggests that most erroneous responses to the abstract tasks are explainable in terms of participants' misapplication of Gricean conventions, determinacy bias and inaccurate representations of the given premisses.

TABLE 7.4
Abstract syllogisms which elicited 3 or more non-logical responses

Task	*Premisses*	*Logical Response* [*Erroneous Response*]	*Possible Causes*
AA4	All A are B, All B are C	Nothing [Some C are A]	3, 5, 6
EA3	No B are C, All B are A	Nothing [No A are C]	1, 3, 5, 6, 9, 10
EA4	No A are B, All B are C	Nothing [Some C are not A]	3, 5, 6, 10
EI4	No A are B, Some B are C	Some A are not C [Nothing]	10, 11
OA1	Some B are not C, All A are B	Nothing [Some A are not C]	1, 5, 6, 9, 10
OA3	Some B are not A, All B are C	Some C are not A [Nothing]	9, 10, 11

Note: Numbers of possible causes relate to the list presented in Table 7.3.

Table 7.5 contains English translations of those thematic syllogisms in which three or more participants endorsed the same non-logical conclusion and gives a speculative list of possible causes for these errors, according to participants' responses. The table suggests that many errors are explainable in terms of the predictions of set-theoretic or analogical models, and participants' adherence to Gricean maxims. The errors are consistent in this respect with those observed for the abstract tasks. Table 7.5 also suggests, however, that belief bias were more prevalent during the thematic tasks. These findings are consistent with the view that meaningful syllogistic terms, when combined with prior beliefs relating to these terms and the tendency to use conventions of everyday linguistic usage, may have distorted the logical demands of some thematic tasks.

TABLE 7.5

Thematic syllogisms which elicited 3 or more non-logical responses

Task	*Premisses*	*Logical Response* [*Erroneous Response*]	*Possible Causes*
AA4	All bank managers are responsible, All responsible people are trustworthy	Nothing [Some trustworthy people are bank managers]	3, 4, 5, 6
AO2	All honest people are hard workers, Some politicians are not hard workers	Some politicians are not honest [Nothing]	10, 11
IE2	Some scientists are methodical, No drunkards are methodical	Nothing [No drunkards are scientists]	1, 5, 6, 9, 10, 11
IE4	Some edible foods are vegetables, No vegetables are minerals	Nothing [No minerals are edible]	1, 4, 5, 6, 9, 10
EA1	No millionaires are poor, All rich people are millionaires	No rich people are poor [Nothing]	10, 11
EA3	No churchgoers are atheists, All churchgoers are devout people	Nothing [No devout people are atheists]	1, 3, 4, 5, 6, 10, 11
EA4	No oranges are apples, All apples are fruits	Nothing [Some fruits are not oranges]	3, 4, 5, 6, 9, 10, 12
EI3	No metals are woods, Some metals are conductors	Some conductors are not woods [Nothing]	10, 11
EI4*	No disloyal people are married, Some married people are traitors	Some traitors are not disloyal [Nothing]	4, 10, 11

Note: Numbers of possible causes relate to the list presented in Table 7.3. Premisses leading to unbelievable conclusions are marked with an asterisk.

It seems worthy of note that tasks AA4, EA3, EA4, EI4 appear in both Tables 7.4 and 7.5. This suggests that participants from both linguistic groups experienced difficulties with these tasks, and that the causes for these errors were not directly related to the level of thematic content used.

7.6 Conclusions

Given that the response trends observed in the present Z based study are consistent with cognitive theories designed to explain human reasoning in natural language based syllogistic studies, it would appear that participants employed reasoning strategies common to both language domains. The fact that the trends are consistent with a wide range of such theories suggests that participants' errors were due to combinations of non-logical reasoning heuristics or biases, as suggested by Tables 7.4 and 7.5. Given that the responses appeared, in many cases, to depend on participants' beliefs towards the seemingly real world referents of the task materials, it also seems probable that those factors which elicited errors differed between participants. The errors are clearly consistent with many cognitive theories, but the question of which particular biases or heuristics caused these errors is not so clear.

It is argued that a failure to distinguish between the laws of everyday reasoning and the laws of logic causes many syllogistic errors (Politzer, 1986; 1990). It was perhaps because of this failure that participants seemed so strongly inclined to employ everyday linguistic conventions. This trend was particularly noticeable under the thematic condition, where the presence of realistic terms seemed almost to cue Gricean conventions and lead reasoners away from the logic of the tasks. Specifically, many errors seem attributable to participants' failure to recognise that the Gricean maxims of quantity and relation are not universally applicable. It was participants' adherence to the maxim of relevance, for instance, which seemed to elicit large numbers of determinate responses to indeterminate tasks. Although the predisposition to conform with Gricean convention apparently led participants away from the logic of some tasks, it appeared to encourage the correct conclusion in others. Adherence to the maxim of quantity during the background tasks, for example, seemingly led most participants to unambiguous set-theoretic interpretations of the

formal A and E expressions. During the main tasks, this maxim also appeared to lead participants to endorse strong conclusions where weaker ones were also possible.

> "Aristotle, after all, invented the syllogisms as a means of enabling people to extract the logically necessary information from discourse and thus to loose themselves from the interpretive aquiescence that language invites. To the extent that people nevertheless perceive and treat the syllogisms as discourse, the system cannot serve its purpose. How could this problem be remedied? One possibility might be to make the system less seductively language-like. One might recast it on terms of, say, propositional logic. The major drawback to this solution is that the logic would remain relatively inaccessible except to reasoners with special training" (Adams, 1984, p.303-304).

Given their logical training and the explicitly logical nature of the tasks, one might have expected computing scientists to have been cued into using the laws of logic throughout. The fact that their levels of correctness were generally higher than those observed in logically equivalent natural language guises suggests that the laws of logic exerted a dominating influence on their reasoning. The non-negligible rates of observed errors, however, suggest that the users of formal methods are liable to disregard logic in favour of non-logical biases and heuristics, including those based on pragmatic convention that occur regularly in everyday communication. With regard to Adams' (1984) hypothesis, therefore, formalisation in terms of a notation with strong foundations in propositional logic appears to provide only a partial remedy for people's errors in the syllogistic task; it does not seem to prevent reasoners from applying inappropriate language conventions from ordinary discourse, including those reasoners with "special training".

7.7 Summary

This chapter began with a discussion of how the categorical syllogism provides a means for explicating some of the core processes underlying quantified reasoning and, moreover, those underlying human reasoning in general. This led onto a review of the various forms of systematic error and bias exhibited during natural language based studies of syllogistic reasoning. This review formed a basis for design of the formalised study. The chapter concludes with a discussion of the results, which again suggest that many of the non-logical heuristics that people exhibit when reasoning about logically equivalent statements in everyday language are liable to transfer over into the formal domain. The next chapter explores a method for reformulating the results of the three main empirical studies in terms of a single probablistic model for predicting human reasoning performance in formalised contexts.

Chapter 8

Theories to Models to Measures

> "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).

Our empirical studies have enabled us to identify a range of factors, including grammatical constructs and linguistic conditions, which are liable to affect human reasoning performance in formalised contexts. This chapter recasts the results of these studies into a descriptive statistical model for measuring the levels of psychological complexity likely to be experienced by users when reasoning about formal expressions in Z specifications. It demonstrates how such a model might be applied in software engineering contexts so that corrective actions can be taken to reduce the likelihood of inaccurate development decisions being made. It is important to recognise that the model is demonstrated in order to illustrate the concepts involved in assessing the psychological complexity of formal specifications, rather than to present a "tool" which is ready for general application. The chapter concludes with

a description of how far the methods used during the model's construction satisfy generally accepted software measurement validation criteria.

8.1 A Model of Conditional Reasoning

A logistic regression analysis was used to model the data generated during the formalised study of conditional reasoning (see Chapter Five). Table 8.1 shows that the greatest variance in participants' levels of correctness was accounted for, first, by the reasoner's level of expertise, second, by the type of inference drawn and, third, by the degree of meaningful content in the task material. The χ^2 values may be interpreted as the improvements made 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. DF refers to the degrees of freedom associated with these values. Although the accuracy of a logistic regression model's predictions generally increases along with the number of input parameters that 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 8.1
Improvements made to the conditional model by stepwise addition of variables

Step	χ^2 *Improvement*	*DF*	*Significance*	*Variable Added*
1	53.635	2	< 0.01	Expertise Level
2	47.396	3	< 0.01	Inference Type
3	12.546	1	< 0.01	Material Type

A standard measure of how closely a regression based model reflects its underlying data is to classify the proportion of predictions given by the model that are consistent with the observed data points from which the model was generated (formula in Norušis, 1996). The "Classification-fit" for our model of conditional reasoning is 88%. Given that only 12% of our data points are misclassified, this suggests that the model provides a reasonable fit to the data.

Another 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. Using this value it is possible to calculate the "Percentage of variability" in the data accounted for by the model. This is obtained by dividing the sum of the parameters' improvements to the model by the Goodness-of-fit value for the model with no explanatory parameters. This calculation tells us that 18% variance in the observed data is predictable by our model. Dawes' (1971) model, in comparison, accounts for 16% variance. The mathematical formulae necessary for these calculations are shown below (adapted from Norušis, 1996, p.10).

Percentage of variability	$\frac{Sum\ of\ improvements}{Goodness\text{-}of\text{-}fit}$	
Sum of improvements	$\sum \chi_i^2$	where i is each parameter in the model
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

Calculation for the model of conditional reasoning $\frac{113.577}{640.225} = 18\%$ variance

A logistic regression analysis generated the results shown in Table 8.2. 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 statistical

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 from the influence of other variables.

TABLE 8.2
Parameters in the model of conditional reasoning

Factor	*Parameter*	β	*SE*	*DF*	*Significance*	β_x
Material-Abstract	$M1$	-0.8794	0.25	1	< 0.01	-0.4397
Inference-MP	$I1$	2.9167	0.55	1	< 0.01	
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	< 0.01	-0.5991
Expertise-Proficient	$E2$	-0.0207	0.45	1	0.96	
	Const	2.4588	0.20	1	< 0.01	

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 a participants' chances of drawing a logically correct conclusion 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 gives the log of the odds, or "logit" value, as shown in the following general formula.

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

The following examples demonstrate how this formula can be applied to calculate the logit values under a range of conditions. These examples illustrate how the calculations are always performed relative to the regression constant and the β_x parameter means.

$$\text{logit}(Abstract, MP, Novice) = (Const + \beta_{M1} + \beta_{I1} + \beta_{E1}) - (\beta_{Mx} + \beta_{Ix} + \beta_{Ex})$$

$$\text{logit}(Abstract, DA, Expert) = (Const + \beta_{M1} + \beta_{I3}) - (\beta_{Mx} + \beta_{Ix} + \beta_{Ex})$$

$$\text{logit}(Thematic, MT, Expert) = (Const + \beta_{I2}) - (\beta_{Mx} + \beta_{Ix} + \beta_{Ex})$$

$$\text{logit}(Thematic, AC, Expert) = Const - (\beta_{Mx} + \beta_{Ix} + \beta_{Ex})$$

8.2 A Model of Disjunctive Reasoning

A logistic regression analysis was used to model the data generated during the formalised study of disjunctive reasoning (Chapter Six). Table 8.3 shows that the greatest variance in participants' correctness was accounted for, first, by the reasoner's level of expertise and, second, by the degree of meaningful content. A fit to the data (Classification-fit = 91%, Percentage of variability = 6%) was achieved by excluding the following parameters: the type of inference to be drawn, the polarity of premisses, and the position of the term denied or affirmed.

TABLE 8.3
Improvements made to the disjunctive model by stepwise addition of variables

Step	χ^2 *Improvement*	*DF*	*Significance*	*Variable Added*
1	33.272	2	< 0.01	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 the study of disjunctive reasoning are shown in Table 8.4.

TABLE 8.4
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	< 0.01	-1.4719
Expertise-Proficient	$E2$	-1.9421	0.54	1	< 0.01	
	Const	2.5742	0.20	1	< 0.01	

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

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

8.3 A Model of Conjunctive Reasoning

A logistic regression analysis was used to model the data generated during the formalised study of conjunctive reasoning (Chapter Six). Table 8.5 shows that the greatest variance in participants' correctness was accounted for, first, by the degree of meaningful content and, second, by the reasoner's level of expertise. A fit to the data (Classification-fit = 94%, Percentage of variability = 4%) was achieved by excluding the following parameters: the type of inference to be drawn, the polarity of premisses, and the position of the term denied or affirmed.

TABLE 8.5
Improvements made to the conjunctive model by stepwise addition of variables

Step	χ^2 *Improvement*	*DF*	*Significance*	*Variable Added*
1	8.190	1	< 0.01	Material Type
2	11.261	2	< 0.01	Expertise Level

The β estimates quantifying the degree of influence exerted by each of these variables on participants' correctness during the study of conjunctive reasoning are shown in Table 8.6.

TABLE 8.6
Parameters in the model of conjunctive reasoning

Factor	*Parameter*	β	*SE*	*DF*	*Significance*	β_x
Material-Abstract	$M1$	1.5017	0.41	1	< 0.01	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	< 0.01	

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

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

8.4 A Model of Quantified Reasoning

A logistic regression analysis was used to model the data generated during the formalised study of quantified reasoning (Chapter Seven). Table 8.7 shows that the greatest variance in participants' correctness was accounted for, first, by the reasoner's level of expertise, second, by the first premiss mood type and, third, by the degree of meaningful content. A fit to the data (Classification-fit = 92%, Percentage of variability = 3%) was achieved by excluding the following parameters: figure type, second premiss mood type, and the believability of logical conclusions.

TABLE 8.7
Improvements made to the quantified model by stepwise addition of variables

Step	χ^2 *Improvement*	*DF*	*Significance*	*Variable Added*
1	24.476	2	< 0.01	Expertise Level
2	10.305	3	0.02	First Mood Type
3	6.897	1	< 0.01	Material Type

The β estimates quantifying the degree of influence exerted by each of these variables on participants' correctness during the study of quantified reasoning are shown in Table 8.8.

TABLE 8.8
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	$E2$	-0.8050	0.24	1	< 0.01	
First Mood-A	$F1$	-1.8700	0.34	1	0.58	
First Mood-E	$F2$	-0.7561	0.33	1	0.02	-0.2300
First Mood-I	$F3$	0.0233	0.36	1	0.95	
	Const	2.8894	0.16	1	< 0.01	

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

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

8.5 Conversion to Absolute Probabilities

The model developed thus far provides a means by which the users of formal methods can predict the likelihood that a reasoner of given expertise will draw an inference of a given type about a given type of logical statement under specific linguistic conditions. 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, given by Norušis (1996), performs the necessary translation.

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

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

8.6 Demonstrating the Model

So far in this chapter we have reviewed a procedure for formulating a predictive model based on the results of our empirical experiments. We now turn to explore how the model might be employed to reduce the potential for human error and influence development decisions in everyday software engineering contexts.

8.6.1 A Missile Guidance System

We can envisage Will Wise, a senior software developer on a defence based project, having been presented with the operational specification for a guided missile system, as shown in Figure 8.1. Suppose that Will is asked by his team leader to determine the implications of including schema *MissileStatus* within schema *MissileCheck*.

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

MissileStatus

$$current, target : COORDS$$
$$report : MESSAGE$$

$$report' = Hit$$

MissileCheck

$$\Xi MissileStatus$$

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

Figure 8.1: Thematic Z specification for a missile system

Given that the specification is expressed in realistic material, whose variable identifiers refer to 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 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 Z expertise of the reasoner (Expert), and the type of inference to be drawn (MT). 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($Thematic, MT, Expert$):

$$z = (Const + \beta_{I2}) - (\beta_{Mx} + \beta_{Ix} + \beta_{Ex})$$

$$z = (2.4588 + 0.7010) - (-0.4397 + 1.1197 - 0.5991)$$

$$z = 3.0789$$

To translate into an absolute probability:

$$p = e^z/(1 + e^z)$$

$$p = e^{3.0789}/(1 + e^{3.0789})$$

$$p = 0.9560$$

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 8.2.

$[C]$

$M ::= m1 \mid m2$

MS

$p, q : C$

$r : M$

$r' = m1$

MC

ΞMS

$p \neq q \Rightarrow r = m2$

Figure 8.2: 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 probably make a subjective, educated guess based on their feelings towards Sam and the specification. The scope of our model is fortunately sufficient to account for these

conditions and can provide us with a much more quantifiably precise estimate.

To calculate logit($Abstract, MT, Novice$):

$z = (Const + \beta_{M1} + \beta I2 + \beta E1) - (\beta_{Mx} + \beta_{Ix} + \beta_{Ex})$

$z = (2.4588 - 0.8794 + 0.7010 - 1.7765) - (-0.4397 + 1.1197 - 0.5991)$

$z = 0.4230$

To translate into an absolute probability:

$p = e^z/(1 + e^z)$

$p = e^{0.4230}/(1 + e^{0.4230})$

$p = 0.6042$

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 8.3.

MissileStatus

$current, target : COORDS$

$report : MESSAGE$

$report' = Hit$

MissileCheck

$\Xi MissileStatus$

$report = Hit \Rightarrow current = target$

Figure 8.3: 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 con-

clusion, *current* = *target*, has decreased to just 0.5% and 2.9% respectively. These values are given by logit(*Thematic*, *MP*, *Expert*) and logit(*Thematic*, *MP*, *Novice*), and performing the necessary translations. It should now be clear that application of the model has strong implications for the ways in which formal specifications are written and for the levels of expertise acquired by those people who work with them.

8.6.2 A Security Door Alarm

We have seen how our model might be used to quantify and compare the levels of inferential complexity likely to be experienced by people with different levels of expertise when reasoning about formal expressions with different levels of meaningful material. We now turn to consider how the model might be used to discriminate between the logical forms of statements themselves, according to each one's propensity for eliciting erroneous decisions. As a further illustration of our model, we now consider an altogether different scenario.

Imagine that Will Wise is specifying a security alarm system operation, *SecurityCheck*, which is derived from the natural language requirements "The alarm must be set whenever the door is locked". The specification is intended for Sam Slow, a programmer with little formal methods experience, who is responsible for implementing the system. Although Will recognises that it is possible to write the specification in one of numerous possible ways, two particular candidates that Will is contemplating are *SecurityCheck*(a) and *SecurityCheck*(b), as shown in Figure 8.4. It is worthy of note that the conditional expression in *SecurityCheck*(a) and the disjunctive expression in *SecurityCheck*(b) are logically equivalent; $p \Rightarrow q \equiv \neg p \vee q$ (proof in Lemmon, 1993, p.59).

$$DOOR ::= Locked \mid Unlocked$$
$$ALARM ::= On \mid Off$$

DoorStatus

$$door\_status : DOOR$$
$$alarm\_status : ALARM$$

$$door\_status' = Locked$$

SecurityCheck(a)

$$\Xi DoorStatus$$

$$door\_status = Locked \Rightarrow alarm\_status = On$$

SecurityCheck(b)

$$\Xi DoorStatus$$

$$\neg door\_status = Locked \vee alarm\_status = On$$

Figure 8.4: Two specifications of *SecurityCheck*

In the absence of an independent measurement system, the criteria that Will uses to discriminate between the two candidates is liable to vary. It might, for example, be based on Will's personal writing style preferences, his previous experience in writing for audiences generally, or his recollection of expressions which have caused Sam to err in the past. Suppose that Will decides to favour the *SecurityCheck*(a) option based on the intuitive feeling that it is easier to see that $alarm\_status = On$ following the inclusion of schema *DoorStatus* in *SecurityCheck*, and because there are fewer grammatical constructs involved. The question we must ask is: Will this decision lead to a specification which is less likely to cause Sam to err? Application of the model predicts that the likelihood of Sam failing to make the necessary inference with *SecurityCheck*(a) is only 2.88%, as compared with 21.79% for *SecurityCheck*(b). Will's intuitive feelings therefore appear to be well founded and, based on these predictions, he would be well advised to favour the former option. Although application of the model may not have changed the specification author's decision making process in this instance, the main difference is that his intuitive feelings are now supported with concrete empirical evidence.

8.7 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). Sometimes, however, the theoretical or empirical foundations for software metrics are improperly considered prior to their formulation or are checked only as an afterthought. Roche (1994, p.80) states 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 which generated the data that populates our metrics (Vinter et al., 1997a; 1997b; 1997c; 1997d). So rather than construct a formal model and then subject it to empirical validation, our methodology has proceeded in the opposite direction by feeding data from empirical studies into a formal model.

It is argued in the 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, and the latter asks whether the model has been built correctly.

8.7.1 Theoretical Validation

The criteria specified by Sheppard and Ince against which a theoretical validation of software metrics may be performed, along with a discussion of 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 a 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 input parameters to the model is as objective as one could reasonably expect for a model of psychological complexity. For example, the extent to which the logical terms have real world referents determines the "material type" parameter, the type of reasoning to be performed lends itself to the "inference type" parameter, and the length or type of Z experience acquired by the reasoner lends itself to the "expertise type" parameter. 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 reduced through adherence to simple guidelines or the maintenance of historical employee records. It seems likely that a certain degree of subjective judgement will always be present, however, even if obscured by guidelines based on "deterministic" criteria.

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 question of whether users of formal methods are liable to err in ways similar to those observed for the users of natural language seems a reasonable one to ask in light of recent cognitive findings and some of the problems facing today's software developers. This hypothesis underlies the model, whose worth is evident from its ability to identify potential sources of development errors and its ability to provide empirical support for some of the claims associated with formal methods.

8.7.2 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 some systems is that the measures are either unsupported by empirical evidence or that the methods used to validate them are flawed or inadequate (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 formulae, the metrics do not scale

up to larger programs, and the experiments used to validate the metrics were flawed in their design (Coulter, 1983; Ince, 1989).

The method used to formulate our model in this thesis 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 has only 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 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 or extension of the empirical studies would not be of value. Further empirical studies would help to refine the probability data which populates the model; the greater and more representative the samples which underly the model, the more accurate its predictions are likely to be. Such studies may also call into question, via refutation, some of the theoretical assumptions which have hitherto been unrecognised, hence, refining the theoretical basis for the model.

There now follows a discussion of the criteria specified by Sheppard and Ince against which an empirical validation of software metrics may be performed, along with a description of how far our model meets these criteria.

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 significant results being derived from an unusable model. The aims and scope of our model, however, were defined at the outset of this research and are spelt out clearly in Chapter One of this thesis. The criteria against which the model may be validated should therefore be clear. Given that these early

discussions and our theoretical validation have shown the relevance of the hypotheses underlying the model to current software engineering concerns, it should be evident that the model can generate usable and meaningful results.

2. *The artificiality of the data used.* The data which populates the model is based on actual, rather than theoretical, instances of human reasoning by large numbers of staff, students and professional users, each with various levels of expertise. It is therefore representative of the full range of formal methods' users. This provides for a degree of flexibility in the model's predictions. Although this could, of course, be improved with more resources and unlimited access to a cooperative software engineering community, the data obtained seems adequate for the demonstration of our model. The null hypothesis which we sought to test during validation was whether the users of formal methods succumb to similar non-logical errors as those committed by users of natural language. 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 had a significant influence on 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 statistical tests were therefore appropriate for their purpose. Given that they could, and sometimes did, yield results which conflicted with intuition and the recommendations of published literature, these tests were applied objectively.

Having evaluated the model against Sheppard and Ince's criteria at an empirical level, we now scrutinise the relation between its predictions and the results of our empirical studies. We calculated earlier that the probability of drawing an AC conditional inference for an expert reasoner in thematic material is 92% ($p = 0.9151$). If we were to calculate the inferential complexity for the same inference and material type but for lesser experienced Z users, we would expect to obtain lower probability values. These calculations yield a slightly lower p value of 0.9135 for a proficient user and a much lower p value of 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 might 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.

Based on the results of our study of conditional reasoning, we would intuitively expect to see a similar incremental effect by maintaining the same material and expertise type then changing inference type from MP to AC to MT to DA, or by maintaining the same inference and expertise type then changing material type from TFL to AFL. So according to the model, a user's chances of drawing a logically correct conditional inference diminishes along with their level of expertise, the ease of the inference, or the amount of realistic material. Trends in the model's predictions, such as this, are entirely consistent with the results of our empirical studies which revealed significant correlations between participants' expertise levels and their levels of correctness. The real worth of our model, however, becomes evident in those cases where the rank orders of complexity in its predictions do not conform so strongly with our intuitive expectations. In such cases, the model may alert its users to potential problem areas in formal specifications which might otherwise be overlooked in design reviews.

8.8 Summary

This chapter has shown how closely our descriptive model reflects the experimental data upon which it is based, how the model might be applied in software engineering contexts, and how far the methods used in generating the model satisfy well accepted software measurement criteria. We have progressed from cognitive theories of human reasoning, to models of the ways in which people reason about formal specifications, which in turn yield predictive measures of inferential complexity. Our investigative line of inquiry has therefore proceeded hitherto from theories to models to measures.

Chapter 9

Conclusions

> "Only when certain events recur in accordance with rules or regularities, as is the case with repeatable experiments, can our observations be tested - in principle - by anyone. We do not take even our own observations quite seriously, or accept them as scientific observations, until we have repeated and tested them" (Popper, 1992, p.45).

This chapter begins with a discussion of how far the overall and subsidiary aims of the research programme were met. It reflects on the methodology, describing the problems that were encountered and how, with the benefit of hindsight, they might have been avoided or reduced. The research results, the proposed model and its underlying theories are discussed in relation to their contribution to empirical knowledge, and their implications are explored from the perspectives of the software engineering and cognitive science communities. Several possible directions are proposed that could usefully build upon these results. Finally, some concluding remarks summarise the main findings from the programme of investigation.

9.1 Meeting the Research Aims

The overall aim of this research was to explore a new approach for supporting software engineering claims with empirical evidence and help to reduce the numbers of defects that appear in software systems as a result of poorly written software specifications. In pursuit of this aim we have developed a prototype system of metrics based on empirical data. These metrics can be used, first, to evaluate some of the psychological claims associated with formal methods, second, to identify potential sources of reasoning difficulty in specifications and as a basis for engineering less error-prone designs. The overall aim of this research has therefore been met. It should be recognised, however, that the descriptive model developed in this thesis is a tentative one. It cannot account for all of the different forms of human inference that could lead to the introduction of software defects, nor all of the linguistic conditions which could evoke them. This was inevitable in view of the exploratory nature of the investigation and the fact that the overriding aim was to demonstrate the feasibility of the research approach, rather than the generality of the metrics.

Three subsidiary aims were identified under the overall aim at the outset of the project. We now consider how far these were met.

1. *Identify through a review of the cognitive science literature those key factors which significantly affect human reasoning performance in natural language contexts.* Links to empirical knowledge and theories stemming from cognitive research were used to identify properties of formal specifications liable to elicit human reasoning errors in formalised contexts. In searching for possible linguistic sources of reasoning difficulties in a given language, it is reasonable to focus on common properties which have been shown to cause reasoning errors elsewhere. The line of inquiry was therefore directed towards similar variables for which cognitive science had produced empirical evidence of people's sys-

tematic fallibility in natural language contexts. These variables included logical operators, such as the propositional connectives and predicate quantifiers from logic, and linguistic conditions, such as the degree of thematic material, the believability of speculative conclusions, the structure of the inference to be drawn, and the order of terms manipulated.

2. *Test through empirical study whether the non-logical errors and biases that reasoners exhibit in natural language contexts are liable to transfer into the formal domain.* Different combinations of those variables identified under the first subsidiary aim were manipulated under experimental conditions and the resulting effects on reasoning performance were noted. The results suggest that non-logical reasoning heuristics can be cued by the way in which logical statements are expressed and the linguistic conditions in which they appear. This finding is consistent with the results from natural language based studies which suggest that control of an individual's reasoning may be usurped by higher order, non-logical, language independent cognitive processes. The presence of this "transfer effect" is further supported by the fact that participants' erroneous responses were consistent with cognitive theories originally propounded to account for non-logical errors in natural language based studies and with trends in the empirical data yielded by these studies.

3. *Formulate a system of metrics for quantifying how far different combinations of these factors are likely to affect human reasoning performance in formalised contexts.* Having established that users of formal methods are prone to various forms of error and bias when reasoning about formal specifications, a means of assessing the potential of alternative design representations for admitting these non-logical heuristics was developed. This was achieved by synthesising the results of the main studies into a descriptive model which may be used

to quantify the levels of inferential complexity in given specifications. The model is not complete, of course, but it has proved adequate for its intended purpose by demonstrating a means for discriminating between alternative design representations in software engineering contexts, and by demonstrating an empirical means for testing some of the software engineering community's psychological claims relating to the use of formal methods.

9.2 Reviewing the Research Methodology

Given the lack of previous research into the cognitive processes involved in formal specification, the aims and scope of this project were far from clearly defined at its outset. Although the intention had been to help reduce the numbers of errors committed in the specification process as a result of erroneous human decisions, it was unclear exactly what an investigation of this nature would entail and how it could be performed. The initial study helped in this respect by explicating some of the key cognitive processes involved in formal specification, generating empirical data around which the research methodology could be refined, and instilling a sense of confidence that there were relevant and interesting findings to be made. It might be argued that our initial investigation could have been omitted and attention directed immediately towards the main studies, since it was only their results which formed the basis for our system of metrics. This programme of research could not have achieved its aims, however, without the knowledge gained from the initial study, because it was this knowledge which set a context for the main studies by pointing to likely sources of human reasoning errors in formalised contexts, and by stimulating empirical hypotheses which the main studies sought to address.

The programme of investigation has broken with two marked trends in traditional computing science research. This was necessary in view of the research aims.

First, it has focussed on the human users of software engineering technology, rather than on the technology itself, in the belief that users play a highly significant part in determining the performance of that technology. Second, it has adopted much more of an interdisciplinary stance than many computing research programmes, by applying theories and procedures from cognitive science to address software engineering problems. Computing research stands to benefit by adopting a cognitive stance, first, by learning the psychological implications of applying its technologies and, second, by supporting or refuting the claims relating to these technologies with empirical evidence. It is therefore not only the results or theories to emerge from this research which may be beneficial to the software engineering community, but the way in which the research was conducted.

Although computing research has proposed a range of statistical models to characterise various kinds of complexity in development contexts, the ways in which some models are formulated reflect only the personal views of their originators; these views are not shared by the wider computing or cognitive communities (Coulter, 1983; Ince, 1989; 1990; Ott, 1996). Consider, for example, the metrics proposed by Halstead (1977) and DeMarco (1982). A comparison with the methods used to generate our model of inferential complexity reveals some marked differences, perhaps the most notable of which is that all and only the data from our empirical studies populate the model's mathematical formulae; our calculations do not rely upon subjective weightings. Whilst it may be argued that subjective judgement decides which variables are input as parameters to the model, selection of these variables can now be refined, along with their underlying theories, in a scientific way, that is, in response to repeatable studies building on the results of this research. It is noteworthy that new research directions at other UK academic sites, such as the Empirical Assessment of Formal Methods project at Southampton University, are seeking further means for the assessment of claims relating to formal methods.

Although the descriptive model developed in this thesis provides an empirical means for assessing some of the claims associated with formal methods, there is no apparent reason why similar models should be restricted to the Z notation, to formal methods, to the specification process, nor even to human reasoning. By using theories from cognitive science as a basis for the design of appropriate studies and following the same procedures used in this research, similar models might be formulated to measure the psychological complexity of other software engineering technologies, such as program design or source code. As in this research, the process of generating data to support such models can help to identify unforeseen or overlooked problems associated with the use of software technologies, regardless of whether these technologies are emerging or well established. The resulting models may be used to mitigate against various forms of human error, and as an empirical basis for evaluating anecdotal claims or comparing competing technologies.

With the benefit of hindsight the prompt used to extrapolate confidence ratings during the three main studies could have been improved. Although the intention had been to present category titles on a linearly ordered scale, it is debatable as to whether such an order is suggested by: "Not confident", "Guess" and "Confident". The category "Guess" is problematic because it is unclear how far it falls between the other two categories or, indeed, whether or not it even does so. This problem might be resolved by selecting more appropriate category titles: "Not confident", "Slightly confident", and "Highly confident". Alternatively, the use of a numeric scale would ensure a fixed interval between participants' ratings.

A central component of the methodology used in this research was the application of theories and procedures from cognitive science to assess the performance of human participants under experimental conditions. Given that our experimental tasks required participants to exercise deductive forms of reasoning in formalised contexts to reach the correct answers, the decision to use logic as the criteria against

which to assess participants' responses was a natural one. This decision simplified the task of performance assessment by making the concepts of "logicality" and "correctness" synonymous in the context of our experiments. Under other criteria the correctness of participants' judgement might not be so clear-cut, particularly where this criteria is based solely on the experimenter's personal opinion. It should be recognised that the methodology may be unsuitable for experiments based on informal reasoning, for example, where the criteria used to evaluate the correctness of human decisions often cannot be formally defined and what is regarded as "correct" is liable to vary between individuals - as exemplified during our initial investigation, which incorporated a survey of writing style preferences. Ideally, the criteria for assessing performance will provide an independent, categorical assessment of correctness for all possible response types, in a manner analogous to logic for deductive reasoning. Such criteria might, for example, include the laws of arithmetic to assess the correctness of participants' mathematical calculations, or the grammar of a programming language to assess the syntactical validity of responses expressed in that language.

The problem of finding adequate numbers of suitably skilled participants is one which faces applied cognitive science research in general. This problem became pronounced during the latter stages of this research as the supply of willing and eligible volunteers gradually approached exhaustion. The problem was exacerbated by the need for adequate numbers of participants with specific levels of expertise in order that the experimental groups could be counter balanced and appropriate statistical tests performed. At present there is a relatively small number of regular Z users, even fewer industrial practitioners, and many of these are reluctant to participate in what are generally perceived to be tests of their competence. This is understandable because the question of whether formal methods do in fact lead to the benefits commonly purported in the computing literature remains a com-

mercially sensitive issue for the producers of formal methods technologies, for the organisations who purchase and apply them, and for the customers who rely upon systems developed from them. This was illustrated in the many queries raised by participants from industrial organisations about whether their "test scores" would be made available to their seniors or to external organisations. Although prospective volunteers were told that their anonymity would be preserved and that their responses would be treated in the strictest confidence, the aforementioned concerns may have dissuaded many other industrial users from participating in this research.

9.3 Further Implications and Future Directions

This thesis has focussed almost exclusively on variables for which cognitive science has produced empirical evidence of people's systematic fallibility, namely, specific combinations of logical constructs and linguistic conditions in natural language. It is possible, however, that reasoning in formalised contexts may be influenced significantly by many more independent variables than we have considered in this research. We must be prepared to take on board other relevant findings from cognitive science and investigate their effect on human reasoning performance. We have painted only a small part of the overall picture and it will take many more studies of the ways in which people reason in formalised contexts before we will be able to discover exactly what it is we are painting.

> "We may say that the most lasting contribution to the growth of scientific knowledge that a theory can make are the new problems which it raises" (Popper, 1974, p.222).

In the forthcoming discussions we explore the main implications of the research findings for both the software engineering and cognitive science communities, pausing occasionally to suggest possible directions for future research.

9.3.1 Testing the Formalists' Claims

Although the model of inferential complexity developed in this thesis is a tentative one and no claims are made regarding its immediate suitability for application, the methods used in its formulation demonstrate an approach via which anecdotal claims pertaining to software engineering technologies can be subjected to empirical examination. The approach has been used here to quantify how far the human potential for error is liable to remain after formalisation of the software specification process. Rather than being based on subjective belief, the lines of inquiry pursued in this research stem from well supported cognitive studies. Rather than using isolated case studies from which results can be difficult to extrapolate, we have borrowed experimental procedures from cognitive science 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 psychology in this manner (Coulter, 1983; Fenton and Pfleeger, 1996; Ott, 1996).

Although it was not a direct aim of this research, the same methodology may be used to generate a model of inferential complexity for natural language predicates. The feasibility of this idea was demonstrated in our study of conditional reasoning, where a natural language based version of the formalised tasks was presented to a separate experimental group. This research has generated the empirical data to populate a model of inferential complexity for a range of linguistic conditions in formal contexts. It remains as an exercise for future research to apply the same theories, follow the same methodology, and repeat the experiments to generate a model for quantifying inferential complexity in informal contexts. Once this has been achieved, the two models might be used to compare the inferential complexity of logically equivalent formal and informal specifications. One would expect the claims concerning the relative benefits of formal and informal specifications to gain

more credence providing they are based on empirical comparisons of this kind, rather than isolated case studies and subjective opinion.

9.3.2 Writing Formal Specifications

Although the cognitive processes in the creative process of writing formal specifications were not a direct focus of concern at the outset of this research, our analysis of the ways in which people interpret and reason about specifications has yielded implications for the ways in which specifications are written. This is because specifications containing high levels of inferential complexity are more likely to elicit errors of human judgement than those without.

> "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" (Gurr, 1995, p.395).

For every statement expressed in a formal notation it is always possible to find an alternative expression which conveys the same meaning, owing to the logical nature of their underlying grammars. It should be clear that our model of inferential complexity enables us to estimate how far each alternative is likely to admit errors of human reasoning and, hence, which properties of a formal notation are the "safest", or least error-prone, to use in given situations. It thereby provides a basis for discriminating between alternative ways of expressing designs, and for resolving development decisions such as: "Would it be safer to use $p \Rightarrow q$ or $\neg q \vee p$; negative or affirmative forms; abstract or thematic identifiers?" Application of the model is therefore likely to prove beneficial at the initial creative stage of the specification process, when a designer frequently makes numerous implicit decisions of this kind and where "there exists a multiplicity of potential designs for even the most trivial problem" (Sheppard and Ince, 1989, p.91).

Although formal grammars are generally much more restricted than those of natural languages, the style in which formal specifications are written is an active area of research. Following the development of formal notations, the software engineering community began publishing recommendations for particular linguistic styles and desirable properties of formal specifications. Gravell's (1991) claim, that communication may be improved by emphasizing preciseness rather than conciseness, was subjected to empirical analysis during our initial study. The results, although tentative, suggest that audiences exhibit no significant difference in their preference for precise, concise and verbose styles. Rather than revealing any general preferences, the results suggest marked links between audiences' ages, levels of experience and their style preferences. We focus now on two specific recommendations which illustrate how claims based on anecdotal evidence can give rise to oversimplified conceptions of the cognitive processes involved in formal specification.

1. Macdonald (1991, p.7) recommends that "it is usually better to avoid one-letter names for variables unless they are only used locally, such as in quantified expressions", while global identifiers "should be given meaningful names (often full words) in order to make understanding easier for the reader".

Rather than being based on results from purposely designed experiments, Macdonald's recommendation appears to be based on intuitive belief and personal experience. The predictions of our model suggest that, whilst human reasoning performance may indeed be facilitated for conditional rules containing meaningful names rather than abstract one-letter names, the opposite seems to be the case for disjunctive, conjunctive and quantified rules. Macdonald's recommendation is therefore consistent with our model's predictions only in the case of conditionals. It does not seem applicable to the other forms of logical rule because reasoning performance was observed in many cases to improve as a result of using abstract

content. If the users of formal methods follow Macdonald's recommendation uncritically then, despite their good intentions, the predictions of our model suggest they are liable to write formal specifications which cause erroneous decisions.

2. It is argued that a formal specification should contain both mathematical and natural language descriptions of the required system (Bowen, 1988; Hall, 1990). Gravell (1991, p.148) argues, moreover, that the "syntactic gap" between the two forms should be kept as narrow as possible "by choosing a mathematical formulation which closely mirrors a straightforward English description".

Gravell's recommendation is appealing, at least intuitively, because the expression of formal statements in ways which mirror their natural language counterparts could cue people in formalised contexts to favour those pragmatic reasoning procedures with which people are highly familiar and which tend to be used successfully on a frequent basis in everyday life. Our empirical studies suggest that reasoners' use of these same everyday heuristics in formalised contexts can be distractive, however, and can lead to the endorsement of logical fallacies. Gravell's recommendation might therefore be considered contentious. Jacky (1997, p.8) presents a counter argument to Gravell's recommendation in his claim that a good formal model is no mere paraphrase of a prose description, but "a different expression of the same behaviours, in a form that is better organized to serve as a guide for programming". The question arises: Which, if any, of these two contradictory recommendations should the users of formal methods employ? A cognitive approach could provide answers to empirical questions of this kind.

9.3.3 Extending the Model to Complex Predicates

The design of our main empirical studies has focused on the ways in which people reason about "atomic" predicates in the Z notation, that is, forms containing only

one logical connective or quantifier, plus negatives. This design was deemed a sensible starting point to test for similar errors and biases as those exhibited in natural language based studies of human reasoning, especially since these studies adopt a similar approach. The scope of our model of inferential complexity is relatively narrow as a result of this design, however, because the model in its present state is applicable to only a small subset of the possible predicates that can occur in Z specifications. It would be interesting to test how far the results from our studies, and also from the natural language based studies, generalise to more "complex" predicates containing different combinations of multiple logical operators. The scope of our model must be extended in this way before it will be applicable in industry. A starting point for such research might be an investigation into the formal equivalents of those few complex predicates for which cognitive science has already generated evidence of people's fallibility in informal contexts, such as multiply quantified statements like "Every man knows some woman, therefore, some woman is known by every man" (Johnson-Laird et al., 1989; Wason and Johnson-Laird, 1982).

Providing systematic reasoning errors can be elicited for an atomic predicate containing a single logical operator, these errors are unlikely to be rectified by adding extra operators to this predicate. It would be reasonable to expect measures of inferential complexity to increase for complex predicates, particularly where all of the additional operators are logically necessary for the conclusion to be drawn. The cognitive theories which underly our atomic model could therefore account for human reasoning errors with complex predicates, however, it is unlikely that there will be direct correlations between the measures yielded for atomic and complex predicates. We must be careful not to make any claims based on this intuitively obvious assumption, however, without adequate support from empirical data.

9.3.4 Extending the Model to Composite Inferences

The process of reasoning about a formal specification, and rigorous verification in particular, can require lengthy chains of deductive reasoning with numerous intermediate stages. Consider the premiss pair "$p \Rightarrow r;\ p$", where only one conditional modus ponens inference is necessary to reach the logically valid conclusion, "r". Contrast this with the premisses "$(p \vee q) \Rightarrow r;\ p$", where disjunctive elimination and conditional modus ponens inferences are necessary to reach the same conclusion, "r". We must consider how our statistical model can be extended to cope with "composite" inferences of this nature involving intermediate stages.

The feasibility of extending the model to cope with composite inferences was demonstrated during the study of disjunctive and conjunctive reasoning, where participants were required to draw two stage inferences, comprising application of a De Morgan's law followed by an introduction or elimination rule, in order to reach the correct conclusion. The results for these tasks suggest that performance increases along with the experience and expertise of the reasoner. It is left as an exercise for future research to determine whether increased exposure to formal notation has a similar facilitating effect for the other forms of composite inference that are typically drawn when users reason about formal specifications in industry.

9.3.5 Extending the Model to Other Formal Notations

The grammar of Z has been used in this research as a tool for experimentation and metrics formulation. The measures yielded by the resulting model are limited because they can only be interpreted in relation to Z specifications. The accuracy with which the same measures might also quantify levels of inferential complexity in other notations is an open question. Besides extending our model to account for more complex forms of predicate and composite forms of inference, we might

also question whether models of inferential complexity can be devised for formal notations other than Z. The scope of our metrics could generalise in this way, given that the logical calculi underlying the Z notation are the same as those underlying many other formal notations, but this remains to be tested.

Standard logic (that is, propositional logic with predicate calculus extensions) forms the grammatical basis for many other popular notations including: Gypsy (Ambler, 1977), Larch (Guttag et al., 1985), RAISE (RAISE Language Group, 1992) and VDM (Jones, 1989). Given that many of these notations share the same logical symbols as Z for denoting negation, condition, disjunction, conjunction and quantification, it seems likely that a repetition of our experiments in these grammatical contexts would yield similar results to populate a system of metrics which, in turn, would yield similar measures of inferential complexity. It would make for an interesting research exercise to investigate and quantify the ways in which users reason about specifications expressed in different notations, and one whose results could have far-reaching implications for the ways in which notations are selected for use on software projects if any significant differences were found.

9.3.6 Automated Tool Support

Our model of inferential complexity is aimed at the initial stages of software projects where formal specifications are usually constructed and, during the course of which, many key development decisions are made. A specification document might undergo numerous revisions, however, during the course of an entire project, and measures must be gathered and analysed following revisions in order to ascertain when these give rise to levels of complexity outside acceptable limits. It may be impractical to perform this process manually or on a regular basis, however, in view of the large numbers of calculations involved, its time consuming nature, and the risk of data collection or analysis errors. Experience has shown that the automation of

collection and analysis procedures enables metrics to be influential at the initial design stage of software projects (Bainbridge et al., 1990). A computerised tool, capable of automatically generating up to date measures of inferential complexity from given specifications, would clearly be helpful to developers in this respect.

A key component of many automated metrics collection tools is a language parser which scans given texts and applies mathematical formulae to yield meaningful measures (Heitkoetter et al., 1990; Whitty, 1997). An essential prerequisite for the development of this component is a precise and complete description of the grammatical rules comprising the language in which the texts are written (Roche, 1994). Such a description is often not available, however, or cannot be defined explicitly. Until the advent of formal methods, program source code had been one of the few tangible outputs from the software development process amenable to automated means of parsing and measurement (Ince, 1989). Given that the grammatical foundations of many formal notations are explicitly defined and that it is possible to write parsers for such grammars, formal specifications can now be treated in a similar manner, but at an earlier stage in the development process. A possible exercise for a future academic project would be to develop a tool capable of parsing Z specifications and applying our formulae for measuring inferential complexity in order that these may influence the ways in which developing specifications are expressed.

9.3.7 Knowing When to Apply the Model

One dilemma raised by the development of our model, and by measurement systems in general, is knowing when to favour a model's statistical predictions over contradictory predictions based on the learned experience of human practitioners. This dilemma is explored in the case of the clinical psychology community by Meehl's (1973) thesis, "When shall we use our heads instead of the formula?" Whilst intuitive and statistical methods have their relative merits and clinical psychologists seek