Applying Software Metrics to Formal Specifications: A Cognitive Approach

Rick Vinter
EDS, Systems Assurance Group, UK

Martin Loomes and Diana Kornbrot
University of Hertfordshire, Hatfield, UK

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

KEYWORDS
Reasoning, formal, specification, software, metrics.

1 Introduction

There are three main questions which form the basis for the research discussed in this paper:

- How does the actual performance of software engineers differ from that which is expected (or claimed) when they are undertaking reasoning tasks using formal notations?
- Do the results of experiments conducted by cognitive psychologists into the reasoning performance of general populations with natural language carry across into a population of software engineers using formalisms?
- Can we find systematic reasoning errors that can be predicted by attributes of the task undertaken or the person performing the task and, if so, can we use this information to build predictive metrics for reasoning performance in software specification?

The research reported here was supported by the UK Economic and Social Research Council Grant No. J00429434043.

These questions are related, in that they seek to address aspects of the problem noted by Fenton and Kaposi [7, p.293] nearly a decade ago:

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

Moreover, the methodology adopted for seeking tentative answers to these questions attempts to address the concerns of Craigen et al. [2, p.407]:

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

It is important to stress that the scope of this research is novel and ambitious, and that the current results comprise more of a "proof of concept" than a set of findings to be widely used by the community of practitioners. It is our hope that the experiments we have carried out are scientific in the sense that they are repeatable (which is usually not the case for anecdotal, case study based exercises), and also in the sense that we have used an existing "tried and tested" body of knowledge to permit predictions of hypotheses to be refuted or supported. It is also scientific in the sense that we make no claims for "truth", rather we present some results that others might attempt to refute, thereby extending knowledge and understanding in the field.

In the rest of this paper we will outline the methodology for the research, present a small sample of the results, and illustrate how we have used these to produce a predictive measure for human error in reasoning about specifications.

2 Methodology

In order to render the research questions sufficiently concrete for empirical analysis, a notation had to be selected in which to express the tasks. A detailed review of twenty possible notations was undertaken to inform this decision [16]. Z was selected from these as a notation with a stable standard for syntax and a sufficiently large user community from which to recruit participants for the experiments. Whilst this inevitably leads to some loss of generality, Z is sufficiently similar to many other notations (at least, in the features we use in our research) that we would anticipate the results holding if other notations are used, and checking this empirically would be a simple (if resource intensive) matter.

A small scale pilot study was used to refine both the experimental methodology and the questions being explored [17]. Two tentative results that came out of this study are of relevance here:

- Even when experienced software engineers knew that they were participating in an experiment to study their reasoning errors, errors occurred. For example, in one task every participant made the same error when translating a simple piece of formal specification into English. This endorsed the view of the researchers that claims for the use of formal methods as generally applicable technology to improve reasoning performance should not be taken at face value.
- The errors made in reasoning performance were systematic, in the sense that there appeared to be some property in each task which led to several participants making the same error. Moreover, there appeared to be a high correspondence between the performance we observed and that predicted by the numerous experiments reported in the psychological literature on human reasoning, even though we were formalising the tasks and using participants experienced in these sorts of tasks.

As a result of this pilot study several larger scale empirical studies were proposed, and to date four have been conducted. The decision was taken to focus attention initially on three factors which the pilot study suggested might influence reasoning performance.

- The length of time that the participants had spent using Z. This was coupled with a reflective judgement on expertise level.
- The extent to which the specification is "abstract", that is, devoid of thematic content, rather than suggesting realistic concepts.
- The logical constructs used in the specification.

With respect to this third factor, logical operators were chosen with roughly equivalent meanings to those natural language constructs which have been shown to evoke erroneous reasoning in cognitive studies: $\Rightarrow$ (if), $\wedge$ (and), $\vee$ (or), $\neg$ (not), $\exists$ (some) and $\forall$ (all). The studies focussed on reasoning with: conditionals, disjunctives, conjunctives and quantifiers. Negation was not treated separately, but was incorporated into all four studies in a systematic fashion.

A total of one hundred and twenty people took part in the experiments, all of whom had some knowledge of the Z formal notation and training in logical deduction. These consisted of staff and students from various academic institutions, and computing professionals from industrial software companies. Participants comprised 72 staff, 26 students, and 22 professionals. Their mean age was 33.55 years ($s = 9.83$) and 110 had studied at least one system of formal logic beforehand, such as the propositional or predicate calculi. Their mean level of Z experience was 4.63 years ($s = 4.04$). According to participants' ratings of expertise, the experimental groups comprised 31 novice, 54 proficient and 35 expert users of the Z notation. The groups were counter balanced, first, according to these ratings of expertise and, second, according to their lengths of Z experience.

Participants were divided into two groups for each experiment, abstract formal logic (AFL) and thematic formal logic (TFL), with twenty different participants completing the tasks under each linguistic condition. The stimuli for each task comprised a prose description and several Z statements, at least one of which contained a logical rule. These represented the premisses necessary for the inference to be drawn. Participants were also shown four Z statements representing possible conclusions, one of which denoted "No valid conclusion" or "Nothing", depending on the context.

For illustrative purposes we will present an overview of the experiment carried out on conditional reasoning, and provide two examples of tasks. There are four classical forms used in the discussion of conditional reasoning. Two of these, Modus Ponens (MP) and Modus Tollens (MT), permit valid conclusions to be drawn. The other two do not permit valid conclusions, and are usually labelled by the fallacies that arise if conclusions are drawn erroneously, Denial of the Antecedent (DA) and Affirmation of the Consequent (AC). For each of these four forms, the two terms involved can be presented in affirmative (A) or negative (N) polarity. This gives rise to the sixteen tasks covering all possible combinations in the following table.

TABLE 1
Logical forms of the conditional inference tasks

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

Note: Conclusions shown for DA and AC are fallacious.

Two versions of each task were formulated in order to test for possible effects of realistic material; one expressed in abstract terms and one in thematic terms. All abstract tasks described relations between colours and shapes, whereas the thematic tasks described realistic computing applications including: a library database system, a flight reservation system, a missile guidance system, a video lending system, and a vending machine operation. The abstract version of the AC-AN task, for example, was presented as shown below. An asterisk indicates the logically correct conclusion.

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

SetColour

$\Delta ShapeAndColour$

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

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

Figure 1: Abstract conditional AC-AN inference

The thematic version of the AC-AN task was expressed terms of a nuclear reactor control program, as shown below.

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

ReactorTempCheck

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

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

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

Figure 2: Thematic conditional AC-AN inference

The same basic form was used for all four studies, which were administered separately over a period of two years. All participants were recruited via personal invitation. They were not supervised in completing the tasks, and were not put under time pressure. We should acknowledge the role played by these participants, who were prepared to give up their time to make this research possible, effectively sitting an examination which was intellectually quite demanding with no obvious benefit to themselves.

3 Results

Each of the four experiments yielded data on the performance of participants undertaking a set of tasks, for either the abstract or thematic condition. This was analysed together with the participants' length of experience and level of expertise. Predictions were made regarding the expected outcomes, mainly by reference to the psychological literature. Many of these predictions were confirmed, although some were refuted - a detailed account can be found elsewhere [17]. To illustrate the sort of results obtained, here are some holistic features for the conditional reasoning experiment, and a few snapshots of results that we considered "interesting", although we would warn of the dangers of taking these out of context. It is important to stress that, as well as analysing the numbers of correct responses, we were also interested in the occasions where several participants made the same incorrect response, as that would indicate systematic errors, such as those suggested in the psychological literature. Moreover, such systematic errors may well be overlooked within project teams using techniques such as walk-throughs or peer review.

Participants' levels of correctness under each experimental condition in the conditional reasoning task revealed overall rank orders as follows: AFL (79%) < TFL (90%) for group type, DA (73%) < MT (83%) < AC (85%) < MP (98%) for inference type, and NA (81%) < NN (85%) = AA (85%) < AN (88%) for polarity type. Analyses of variance revealed a significant main effect of inference type ($F_{(3,114)} = 7.53, p < 0.01$), and a significant interaction between inference and group type ($F_{(3,114)} = 2.64, p = 0.05$). Analyses by linear re-

gression revealed significant correlations between participants' lengths of Z experience and their levels of correctness ($R = 0.41, F_{(1,39)} = 7.82, p < 0.01$), and between their Z expertise ratings and levels of correctness ($R = 0.45, F_{(1,39)} = 9.40, p < 0.01$). A summary of participants' valid conditional inferences is given in the appendix of this paper.

The results suggest that few participants experienced any difficulty whatsoever in drawing the MP inference, with near ceiling levels observed for all combinations of premiss polarity across both groups. This is supported by natural language based studies in which participants rarely erred when drawing MP inferences, irrespective of the term polarities [3, 5, 14] and the realistic content involved [8, 12]. The lower rates of correct MT inferences drawn in comparison to MP is also well supported [5, 14]. The results suggest that participants experienced some difficulty in drawing DA inferences, where up to 30% succumbed to the fallacy, and that participants experienced most difficulty in drawing the AC inferences where up to 45% succumbed to the fallacy. The high rates at which participants succumbed to these fallacies when reasoning about abstract material, in particular, is also supported in the cognitive literature [4, 14]. A brief summary of the results in relation to cognitive theories of conditional reasoning biases is given below - readers are referred to [17] for a detailed discussion.

- Signs of matching bias [3] were evident for the fallacious DA and AC inferences in the abstract group.
- Signs of negative conclusion bias [6] were evident for the MT, DA and AC inferences in the abstract group.
- Signs of facilitation by realism [9] and belief bias [1] were evident in the thematic group.
- No signs of affirmative premiss bias [6].

It is important to stress that the results above relate only to one study. It is tempting to overgeneralise these findings. For example, one might use these results to endorse the prejudice that formalisation which loses thematic content is damaging to reasoning performance in general. Unfortunately, the other studies would not support this view as, on many tasks, the abstract group performed better than the thematic one. These studies served to confirm our predicted answers to two of the research questions:

- Even experienced software engineers make large numbers of errors when reasoning about specifications expressed in a formal notation.
- The nature of the errors seem to correspond very well in general with the systematic errors reported by psychologists working with a general population using natural language.

It is argued that people find it easier to reason with formal expressions than their informal counterparts [15, 18]. Our results suggest that claims for formalisation such as this, which rest upon strong psychological assumptions, are difficult to support. In some specific areas, however, the use of formal methods (or the background of the participants in this study) does seem to improve matters.

4 A Predictive Model of Error

The results obtained so far allow us to proceed to the final part of the research program: using the data collected to explore the possibility of constructing predictive models which will allow us to predict the likelihood of errors being made by a particular individual working on a particular problem. Once again we will illustrate the technique by reference to the study of conditional reasoning.

In order to quantify how far each variable contributed to participants' levels of performance, a logistic regression analysis was used to model the data points generated. The table below shows that the variance in participants' correctness was accounted for, first, by the reasoner's level of expertise then, second, by the type of inference to be drawn and, third, by the degree of meaningful content in task material. The χ^2 values may be interpreted as the improvements to the accuracy of the model's predictions each time a significant variable was added as a parameter to the model in a forward stepwise manner. Although the accuracy of a logistic regression model's predictions generally increases along with the number of input parameters it allows, there comes a point at which the inclusion of new parameters does not improve the accuracy of the model significantly. This explains why polarity type has been excluded as a parameter from the model and a "fit" to the observed data has been achieved using only three parameters: expertise level, inference type and material type.

TABLE 2
Improvements made to the conditional model by stepwise addition of variables

Step	χ^2 Improvement	DF	p	Variable Added
1	53.64	2	< 0.01	Expertise Level
2	47.37	3	< 0.01	Inference Type
3	12.55	1	< 0.01	Material Type

One measure of how well a regression based model fits its data is to classify the proportion of predictions given by the model which are consistent with the observed data points from which the model was generated [11]. The

"Classification-fit" for our model is 87.81%. Given that only 12.19% of our data points were misclassified, this suggests that the model provides a reasonable fit to the data. An alternative measure, called the "Goodness-of-fit", compares the observed probabilities with those predicted by the model [11]. Using this measure 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 input parameters' improvements to the model by the Goodness-of-fit value for the model with no explanatory input parameters. This calculation tells us that 18% variance in the observed data is predictable by our model.

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

TABLE 3
Parameters in the model of conditional reasoning

Factor	Parameter	β	SE	DF	p	β_x
Abstract Material	$M1$	-0.8794	0.25	1	< 0.01	-0.4397
MP Inference	$I1$	2.9167	0.55	1	< 0.01	1.1197
MT Inference	$I2$	0.7010	0.30	1	0.02	1.1197
DA Inference	$I3$	0.8610	0.31	1	0.01	1.1197
Novice Expertise	$E1$	-1.7765	0.40	1	< 0.01	-0.5991
Proficient Expertise	$E2$	-0.0207	0.45	1	0.96	-0.5991
	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 the participants' likelihood of being correct under the corresponding experimental condition. These values represent the parameters for our conditional model of inferential complexity. According to [10], the "odds" of an event occurring are calculated by the probability that it will occur divided by the probability that it will not. The summation of the β estimates give the log of the odds, or "logit" value, as shown in the following general formula.

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

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

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

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

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

The model developed thus far provides a means by which the users of formal methods can predict the likelihood that a reasoner of a given expertise will draw an inference of a given type about a given type of logical statement expressed in a given degree of thematic material. At present the model yields logit values which appear to have little meaning in isolation. What we are lacking is a means for translating these values into absolute probabilities ($0 \leq p \leq 1$). The following formula performs the necessary translation [11].

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

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

Similar models have been produced from the results of the other three studies.

5 A Brief Demonstration

Let us consider how these predictive models might be used in practice. We can envisage Will Wise, a senior software developer working on a defence based project, having been presented with the operational specification for a guided missile system, shown as follows. 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 3: Thematic Z specification for a missile system

$[C]$

$M ::= m1 \mid m2$

MS

$p, q : C$

$r : M$

$r' = m1$

MC

ΞMS

$p \neq q \Rightarrow r = m2$

Figure 4: Abstract Z specification for the missile system

Given that the specification is expressed in realistic material, whose variable identifiers refer to rather fast moving animate objects, we can safely classify its material as being thematic in nature. Supposing Will had acquired a fair amount of Z experience by formally verifying part of a previous project, had studied several systems of logic at university and had even gone on expensive Z training courses, we might be inclined to regard Will as an expert Z user. If we were to analyse the logic of the terms involved we would see that the consequent of a conditional rule is being denied, which suggests that Will is being invited to draw a modus tollens inference. We now have the three parameters that we need to apply our model of conditional reasoning: the material type (Thematic), the type of inference to be drawn (MT), and the Z expertise of the reasoner (Expert). The question that we must ask is: How likely is Will to infer the logically correct conclusion, $current = target$, under these conditions? Application of the model predicts that Will is 95.6% likely to draw this conclusion, which is calculated as follows.

To calculate logit($Thematic, MT, Expert$):
$z = (Const + \beta_{I2}) - (\beta_{Mx} + \beta_{Ix} + \beta_{Ex})$
$z = (2.4588 + 0.7010) - (-0.4397 + 1.1197 - 0.5991)$
$z = 3.0789$

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 below.

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

MissileStatus

$current, target : COORDS$

$report : MESSAGE$

$report' = Hit$

MissileCheck

$\Xi MissileStatus$

$report = Hit \Rightarrow current = target$

Figure 5: Revised Z specification for the missile system

If we analyse the logic of the terms following the schema inclusion we see that the antecedent of the conditional is being affirmed, which suggests that a much simpler, modus ponens, inference is required. Supposing Will and Sam are asked to determine the implications of the schema inclusion, the model predicts that their potential for failing to draw the logical conclusion, $current = target$, has decreased to just 0.5% and 2.9% respectively. It 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.

6 Conclusions

We have hopefully illustrated that a scientific approach to the questions posed at the outset of this paper is both possible and fruitful. The experiments are repeatable and easily extended, to add confidence (or refutation) and generalisation. The predictive metrics appear to have most of the desirable properties suggested in the literature [13], although space has not allowed us to illustrate this, and are easily applied to simple problems.

There are, however, still many important issues that need addressing before this sort of approach can be seen as approaching an industrial strength tool. We have dealt with simple logical constructs, but have not yet tackled the problems of compound constructs, involving several layers of nested forms. Little work has been done on this in the cognitive psychology community either, and so we will be attempting experiments with little experience to build upon. Similarly, we need to address alternative types of formal method, such as process algebras, whose structure does not map so easily onto logical reasoning in natural language. We see no reason, however, why such areas should not be put on a firm scientific basis in the fullness of time. Hopefully this paper will also lead to the research community raising many more such questions that can be tackled in a scientific manner.

REFERENCES

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

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

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

[4] Evans, J.St.B.T. Reasoning with negatives. *British Journal of Psychology*, *63*, 213-219, 1972.

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

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

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

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

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

[10] Kleinbaum, D.G. *Logistic Regression. A Self-learning Text.* New York: Springer, 1994.

[11] Norušis, M.J. *SPSS Advanced Statistics 6.1.* Chicago: SPSS Incorporated, 1996.

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

[13] Sheppard, M. and Ince, D. *Derivation and Validation of Software Metrics.* Oxford: Oxford University Press, 1993.

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

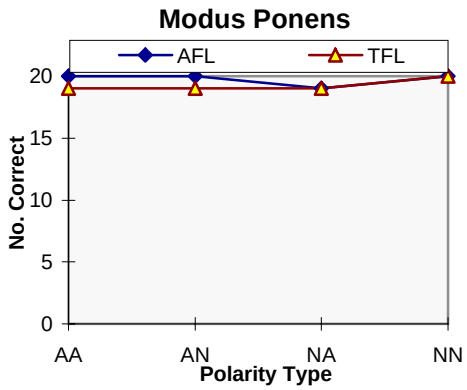
[15] Thomas, M. Formal methods and their role in developing safe systems. *High Integrity Systems Journal*, *1*, 5, 447-451, 1995.

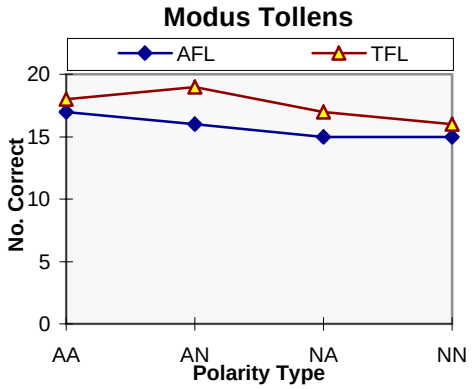
[16] Vinter, R.J. *A Review of Twenty Formal Specification Notations.* University of Hertfordshire, Division of Computer Science, Technical Report No. 240, February 1996.

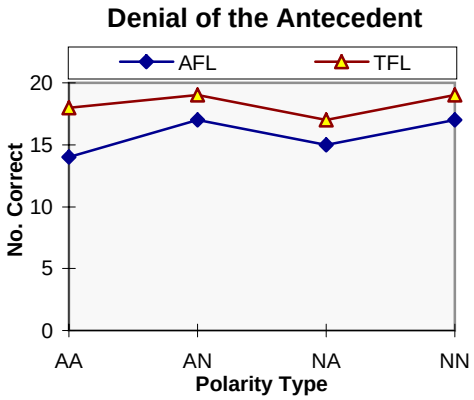
[17] Vinter, R.J. *Evaluating Formal Specifications: A Cognitive Approach.* PhD Thesis, University of Hertfordshire, Division of Computer Science, June 1998.

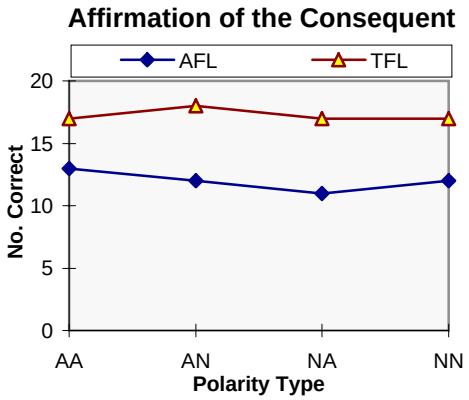
[18] Wing, J.M. A specifier's introduction to formal methods. *IEEE Computer*, *23*, 9, 8-24, September 1990.

APPENDIX


Modus Ponens
AFL
TFL
No. Correct
0
5
10
15
20
AA
AN
NA
NN
Polarity Type


Modus Tollens
AFL
TFL
No. Correct
0
5
10
15
20
AA
AN
NA
NN
Polarity Type


Denial of the Antecedent
AFL
TFL
No. Correct
0
5
10
15
20
AA
AN
NA
NN
Polarity Type


Affirmation of the Consequent
AFL
TFL
No. Correct
0
5
10
15
20
AA
AN
NA
NN
Polarity Type