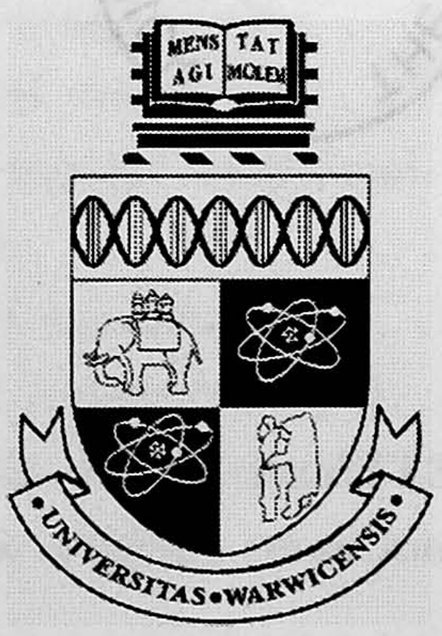
Dual Processes in Mathematics:

Reasoning about Conditionals

Matthew Inglis

A thesis submitted for the degree of
Doctor of Philosophy



University of Warwick
November, 2006

Contents

List of Figures

List of Tables

Acknowledgements

I have received valuable assistance from several sources during the course of writing this thesis.

In particular, I would like to thank Adrian Simpson, who has been an excellent supervisor throughout. Since his move to Durham, he has, despite the pressures of a new job, provided an exemplary model of how to supervise doctoral students from a geographically remote location, and deserves great credit for this. I am also extremely grateful to Derrick Watson, who has invested unreasonable amounts of time and effort to help me with various aspects of my work; it is important to recognise that he has never been in any way obliged or expected to spend *any* time working with me, and thus I am all the more appreciative of his efforts.

There are, of course, many others who have helped with various aspects of this work through suggestions and discussions. Although I am not going to list them all, they too deserve my thanks.

Finally, I would like to thank the Economic and Social Research Council for generous financial support throughout the 3 years of this study.

Declaration

I, the author, declare that the work presented here is my own and has not been submitted for a degree at any other institution. None of the work has previously been published in this form. However, aspects of this thesis have been published in various papers:

Inglis, M. & Mejia-Ramos, J. P. (2006). Applying informal logic to mathematics. *Proceedings of the 3rd International Conference on the Teaching of Mathematics at the Undergraduate Level*, Istanbul, Turkey: Turkish Mathematical Association.

Inglis, M., Mejia-Ramos, J. P., & Simpson, A. (in press). Modelling Mathematical Argumentation: The Importance of Qualification. To appear in *Educational Studies in Mathematics*.

Inglis, M. & Simpson, A. (2004). Mathematicians and the Selection Task. In M. Johnsen Høines and A.B. Fuglestad (Eds.), *Proceedings of the 28th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 3, pp. 89-96). Bergen, Norway: IGPME.

Inglis, M. & Simpson, A. (2005). Characterising Mathematical Reasoning: Studies with the Wason Selection Task. In *Fourth Congress of the European Society for Research in Mathematics Education.* Sant Feliu de Guíxols, Spain: ERME.

Inglis, M. & Simpson, A. (2005). Heuristic Biases in Mathematical Reasoning. In H.L. Chick & J.L. Vincent (Eds.), *Proceedings of the 29th Conference of the International Group for the Psychology of Mathematics Education.* (Vol. 3, pp. 177-184) Melbourne, Australia: IGPME.

Inglis, M. & Simpson, A. (2006). The Role of Mathematical Context in Evaluating Conditional Statements. In J. Novotná, H. Moraová, M. Krátká & N. Stehlíková (Eds.), *Proceedings of the 30th Conference of the International Group for the Psychology of Mathematics Education.* (Vol. 3, pp. 337-344) Prague, Czech Republic: IGPME.

Abstract.

This thesis studies the reasoning behaviour of successful mathematicians. It is based on the philosophy that, if the goal of an advanced education in mathematics is to develop talented mathematicians, it is important to have a thorough understanding of their reasoning behaviour. In particular, one needs to know the processes which mathematicians use to accomplish mathematical tasks. However, Rav (1999) has noted that there is currently no adequate theory of the role that logic plays in informal mathematical reasoning. The goal of this thesis is to begin to answer this specific criticism of the literature by developing a model of how conditional "if... then" statements are evaluated by successful mathematics students.

Two stages of empirical work are reported. In the first the various theories of reasoning are empirically evaluated to see how they account for mathematicians' responses to the Wason Selection Task, an apparently straightforward logic problem (Wason, 1968). Mathematics undergraduates are shown to have a different range of responses to the task than the general well-educated population. This finding is followed up by an eye-tracker inspection time experiment which measured which parts of the task participants attended to. It is argued that Evans's (1984, 1989, 1996, 2006) heuristic-analytic theory provides the best account of these data.

In the second stage of empirical work an in-depth qualitative interview study is reported. Mathematics research students were asked to evaluate and prove (or disprove) a series of conjectures in a realistic mathematical context. It is argued that preconscious heuristics play an important role in determining where participants allocate their attention whilst working with mathematical conditionals. Participants' arguments are modelled using Toulmin's (1958) argumentation scheme, and it is suggested that to accurately account for their reasoning it is necessary to use Toulmin's full scheme, contrary to the practice of earlier researchers. The importance of recognising that arguments may sometimes only *reduce* uncertainty in the conditional statement's truth/falsity, rather than *remove* uncertainty, is emphasised.

In the final section of the thesis, these two stages are brought together. A model is developed which attempts to account for how mathematicians evaluate conditional statements. The model proposes that when encountering a mathematical conditional "if P then Q", the mathematician hypothetically adds P to their stock of knowledge and looks for a warrant with which to conclude Q. The level of belief that the reasoner has in the conditional statement is given by the modal qualifier which they are prepared to pair with their warrant. It is argued that this level of belief is fixed by conducting a modified version of the so-called Ramsey Test (Evans & Over, 2004). Finally the differences between the proposed model and both formal logic and everyday reasoning are discussed.

Chapter 1

Plan of the Thesis

This thesis is about mathematical reasoning, concentrating particularly on the types of reasoning involved in evaluating mathematical conditionals. This topic is interdisciplinary. During the course of this thesis, papers are referred to, and ideas adapted from, several different disciplines: the mathematics education, psychology of reasoning and informal logic literatures are all heavily cited. Despite these disparate influences, the goal from the outset is clear: an integrated theory of the evaluation of mathematical conditionals.

The thesis falls into several distinct parts.

1.1 Literature.

The thesis begins by situating itself within the mathematics education field. Chapter 2 briefly reviews previous work which has looked at the cognitive skills required for mathematics students in order to understand proof. It is argued that the role of logic in proof is not sufficiently understood, and that, in particular, there is currently no satisfactory theory of logic in informal mathematical argumentation.

Chapter 3 looks in detail at the various models which have been proposed for understanding conditional statements. The chapter concludes by discussing three influential tasks from the psychology and mathematics education literatures which have been used to produce these models: the Maze Task, the Truth Table Task and the Conditional Inference Task.

A fourth task, the Wason Selection Task, is by far the most influential instrument in the history of reasoning research, and this forms the subject of Chapter 4. The task is described, the main empirical results reviewed and each of the major theories of reasoning that have been proposed in relation to it are introduced. Although these theories of reasoning are discussed with particular

reference to the Selection Task, they are all intended to be general theories of reasoning, and claim domains of applicability far wider than merely the Selection Task alone.

1.2 Methodology.

Chapter 5 discusses the methodologies available to conduct an investigation into mathematical reasoning. The validity and reliability of both quantitative and qualitative studies, in the form of interviews and standardised tasks, are discussed and compared. Finally the philosophy behind the methodology adopted in later sections of the thesis – the so-called quasi-judicial approach to case study analyses – is discussed.

1.3 Empirical research.

The empirical research reported in this thesis falls into two parts. Firstly, in Chapter 6, the various theories of reasoning discussed in Chapter 4 are critically evaluated by comparing the performance of mathematics students with the general well-educated population on the Wason Selection Task. Using an inspection time eye-tracker based methodology, it is argued that only the heuristic-analytic dual process theory of reasoning can successfully account for the behaviour of successful mathematicians on the Selection Task.

Having demonstrated that the heuristic-analytic dual process theory of reasoning is the most suitable framework within which to study mathematical reasoning, Chapter 7 reviews the theory in greater detail. The heuristics and biases research programme in Decision Making is briefly reviewed, and the intuitive/analytical distinction introduced by earlier mathematics education researchers is compared and contrasted with the dual process framework.

Chapter 8 reports an qualitative interview study which attempts to apply the dual process framework to the specific research question that this thesis set out to answer: how do mathematicians evaluate conditional statements? There are two parts to this study. Firstly, the role of preconscious heuristics in realistic mathematical contexts are examined; and secondly, the conscious processes involved in the evaluation of mathematical conditionals are discussed with reference to Toulmin's (1958) argumentation scheme.

1.4 The theory.

Finally, in Chapter 9, the strands of the thesis are drawn together and synthesised to form one coherent evaluative model of mathematical conditionals. This evaluative model is compared with both formal logic, and models that seek to explain how day-to-day non-mathematical indicative conditionals are evaluated. The thesis concludes by discussing the open research questions which the empirical work reported here has raised.

The full structure of the thesis is shown in Figure 1.1.

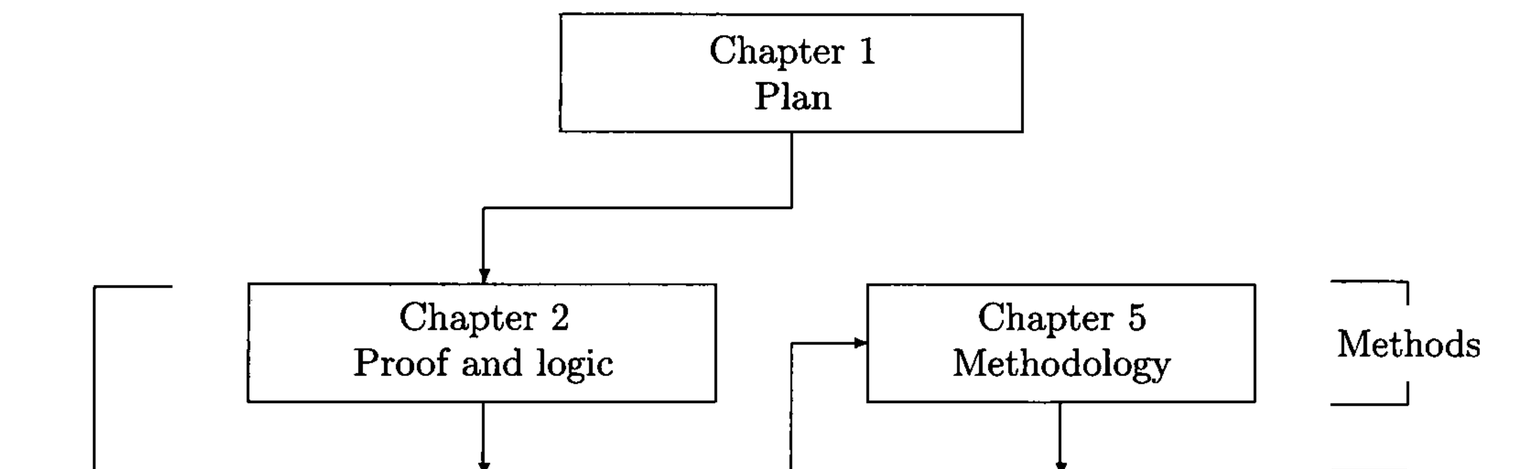


Figure 1.1: The structure of the thesis.

Chapter 2

Proof and the Role of Logic

The concept of proof is one which leads to notorious difficulties for students. A wide range of literature has found that students (from primary school up to undergraduate level) have difficulty constructing arguments and proofs, often fail to understand what proofs are and have difficulty in judging whether an argument is a proof or not (e.g. R. Moore, 1994; Recio & Godino, 2001; Selden & Selden, 2003). This chapter reviews various discussions about what constitutes proof, and provides a summary of the research that has been conducted on students' difficulties with the concept.

2.1 What is a mathematical proof?

According to popular conceptions of the subject, mathematics is held together by the notion of formal proof. Mathematicians prove theorems using logically correct arguments; once proven the theorems are true, and cannot be challenged. In his popular science book 'Fermat's Last Theorem', Simon Singh summed up this idea:

> "Mathematical proof is far more powerful and rigorous than the concept of proof we casually use in our everyday language ... once proven [theorems] are true until the end of time. Mathematical proofs are absolute" (Singh, 1997, p.21).

This essential quality of proofs – that they are the final arbiter of truth – has a long history. John Locke believed that mathematics (along with ethics) was one of only two disciplines where truth can be firmly and indisputably established; he noted that "mathematical proofs, like diamonds, are hard as well as clear" (cited in Dunham, 1994). This view of proof has been characterised as "rightwing" (Devlin, 2004), and is supported by some mathematics educators.

Selden and Selden (2003), for example, claimed that "one neither examines the life and times of a proof's author nor the sophistication of its readers in judging the truth of a theorem" (p.7).

Although it is possible to provide an explicit formal definition of a "rightwing" proof (e.g. Nagel & Newman, 2001), Thurston (1994) argued that it is important to recognise that few mathematicians actually construct such entities. In practice, much of the formal logic and trivial deductions are omitted. Some might argue that any mathematical proof could be translated into a complete formal sequence of logical deductions, but it is clear that this would be a non-trivial task, and may even be impossible. It has even been argued, however, that not only might it not be possible to do such a thing, but also that it would be undesirable (Fallis, 2003).

Contrary to Selden and Selden's (2003) view, many mathematicians, mathematics educators and philosophers believe that what constitutes a proof is, to a large extent, dependent on the community within which you are operating. The notoriously non-rightwing Bertrand Russell, for example, noted that you can never hope to write down the entire reasoning process behind a mathematical result. Instead, he believed, you must write what "is sufficient to convince a properly instructed mind" (Russell, 1961, p.163). This so-called "leftwing" conception of proof (Devlin, 2004) relies upon an agreed standard of what constitutes a "properly instructed mind". It is people-centric; far from being an absolute guarantor of the truth of a mathematical statement, a proof is whatever the mathematical community agrees a proof is (Balacheff, 1987).

The mathematics education literature has a lot to say on the nature of proof. Mason believed that in order to justify a conjecture, there are three stages you have to go through: convincing yourself, convincing a friend and convincing an enemy (Mason, Burton, & Stacey, 1982). Tall agreed with this analysis, but noted that mathematical proof was something more. In order for an argument to be a proof, not only does it need to convince both friends and enemies, it needs to do so in a certain agreed manner involving mutually acceptable procedures that transmit the truth of one statement to another (Tall, 1989).

As with Russell and Devlin, Tall placed the emphasis on the mathematical community agreeing on what steps can be used in a proof. The mathematical community, however, is a diverse thing, and the agreed conception of what constitutes a proof in topology, say, may not be the same as in fluid dynamics (P. Davis & Hersh, 1983; Thurston, 1994). Suggesting that it is difficult to precisely pin down exactly what a proof is, Tall aligned himself with Carroll's Humpty Dumpty who famously noted that "when I use a word it means just what I choose it to mean – neither more nor less" (Carroll, 1988; Tall, 1989).

Although the mathematical community seems unable to come up with a better definition of a proof than "it is what we say it is", there have been numerous attempts in the maths education literature to describe the characteristics of proof.

Justification. Variously described as 'verifying' (Bell, 1976) or 'convincing' (Hersh, 1993; Mason et al., 1982), the primary purpose of a proof is to justify (either to yourself or others) that a theorem is correct.

Explanation. Some mathematicians and mathematics educators argue that a proof should provide some idea to the reader *why* the theorem is true. Hanna (1991, p.55) claimed that a proof that fails to accomplish this "is likely to add very little to an understanding of its subject and ironically may not even be very convincing".

Communication. The 'language' of proof is the way the mathematicians communicate their ideas to each other, allowing new research to be built on old (Knuth, 2002).

Systemisation. Proof may be used to organize the results into a coherent theory of axioms and theorems (Bell, 1976; de Villiers, 1990).

Given that it is so hard to pin down exactly what the role and nature of proof is, it is perhaps not surprising to discover that many students have serious difficulties coming to terms with it. In the next section some of the research findings regarding the difficulties of teaching proof are discussed.

2.2 Students' difficulties with proof.

There has been plenty of research on how exactly students go about trying to convince themselves of a statement's truth. Note that this question is related to proof validation[1] and conviction, as opposed to proof production, although these two types of student interaction with the notion of proof have often been confused in the mathematics education literature. Harel and Sowder (1998, p.275) defined a 'proof scheme' as referring to "what convinces a person, and to what the person offers to convince others". In an exhaustive article they classified some of the common proof schemes that students may use into three general areas: external, empirical and analytical/deductive. In this section these proof schemes are discussed in turn, together with an overview of Tall and Vinner's (1981) influential concept image/concept definition framework.

[1]Proof validation is process of checking that a purported proof is correct (Selden & Selden, 2003; Weber & Alcock, 2004).

It is worth noting that since Harel and Sowder's (1998) original article their proof scheme taxonomy has been further developed and refined based mainly on epistemological, philosophical and historical analyses (Harel & Sowder, 2005; Harel, 2001, in press). The following summary of the different proof schemes uses the more recent terminology:

- External proof schemes (§2.2.1).
  - Ritual.
  - Authoritarian.
  - Symbolic.
- Empirical proof schemes (§2.2.2).
  - Inductive.
  - Perceptual.
- Deductive proof schemes (§2.2.4).
  - Transformational.
  - Modern axiomatic.

2.2.1 External proof schemes.

Harel and Sowder (1998) noted that some students can be convinced and persuaded by some other aspect of the proof other than its contents. They labelled the first such method 'ritual': the actual structure or presentation of the proof may be sufficient to convince a student of its correctness.

Another plausible way of being convinced of the veracity of a proof is by some external authority. Some students can be convinced of the correctness of a proof simply by virtue of who presents it to them. If a proof is in a book or is being presented in a lecture, then it has been checked and verified by a mathematician of greater powers than yourself and this, perhaps, immediately boosts its credibility.

In the third of Harel and Sowder's external proof schemes, conviction is derived from the routine manipulation of symbols without meaning. They described this as "approaching the solution of a problem without first comprehending its meaning" (p.251). This way of doing mathematics has a long history; mathematicians of the eighteenth and nineteenth centuries often manipulated algebraic symbols in bizarre and (from a modern perspective) illegitimate ways, and in doing so derived some of the more important results in calculus. It is worth questioning why Harel and Sowder included this scheme in the 'external'

category. Where is the external source that is providing the authority? They appear to be suggesting the symbols in the proof are external to the proof itself, leaving us wondering what exactly they consider the proof to be.

2.2.2 Empirical proof schemes.

The next category of proof scheme that Harel and Sowder (1998) identified is the so-called 'empirical' scheme. Here, the student becomes convinced of the truth of a conjecture by appealing to observations and experiences. For Harel and Sowder, this can be done in one of two ways: inductively and perceptually.

In a perceptual proof scheme, conviction is obtained by reference to observations of static mental (or physical) imagery. Some authors have noted that this form of reasoning is important when investigating new problems (Dreyfus, 1991). But others have pointed out that using diagrams in formal proof is dangerous; Tall (1995) cautioned us that a diagrammatic proof may only be valid for a range of situations where the diagram is prototypical.

When using an inductive proof scheme students attain conviction by testing one or more specific instances of the conjecture. Chazan (1993) characterised this sort of reasoning as confusing evidence with proof. Balacheff (1988) further subdivided this scheme into "naive empiricism" (randomly picking a few example cases to test), using a "crucial experiment" (picking carefully an example to test) and a "generic example" (where an example is chosen to be representative of the general case). Tall (1979) found that significantly more students preferred a generic proof of the irrationality of $\sqrt{\frac{5}{8}}$ to the standard proof by contradiction. It should be noted here that a generic example based proof is not necessarily considered illegitimate by the mathematical community and that examples of such arguments can be found in certain advanced mathematical texts (e.g. Aigner & Ziegler, 2000). Indeed, it could be argued that proof by generic example should be considered under the deductive proof scheme.

Although mathematicians agree that a (non-generic) inductive argument isn't sufficient to prove a theorem, the importance of inductive thought in creating mathematics has been commented on before. Bickley (1966) suggested that in "creative mathematics" deductive thought plays a subsidiary role to "the dominance of the inductive" (p.7). Tall (1997) wrote that in order to succeed at university level mathematics, "the individual must make an almost schizophrenic separation between the intuitive appeal to the concept image that senses mathematical truth and the formal deduction processes that establishes it." (p.16) Thus, for Tall, the process involved in producing the mathematics is significantly different from that involved in proving it.

The inductive-empirical proof scheme has been found to be widespread.

Across a wide range of mathematical topics, it has been found amongst secondary school pupils (Porteous, 1990; Coe & Ruthven, 1994; Edwards, 1998; Healy & Hoyles, 2000; Küchemann & Hoyles, 2004), secondary school teachers (Knuth, 2002) and undergraduates (R. Moore, 1994; Goetting, 1995; Recio & Godino, 2001).

Given that inductive reasoning appears to be so widespread, some researchers have tried to suggest reasons why. Recio and Godino (2001) drew (slightly superficial) parallels between such inductive proof schemes in mathematics and the kind of reasoning that are the norm in scientific subjects. Drawing on work in cognitive science, Alcock and Simpson (2002) noted that reasoning using prototypical examples is commonplace in day-to-day thought. When it comes to formal mathematics however, they argued that the student must develop "the rigour prefix" to emphasise the importance of working with the formal definition rather than their prototypical examples (Alcock & Simpson, 1999). This idea drew upon Tall and Vinner's (1981) important distinction between an individual's concept image and their concept definition.

2.2.3 Concept image and concept definition.

The idea that an individual has both a concept image and a concept definition has been an influential theoretical framework in mathematics education. It is a simple yet powerful idea:

> "We shall use the term *concept image* to describe the total cognitive structure that is associated with the concept, which includes all the mental pictures and associated properties and processes." (Tall & Vinner, 1981, p.152)

An individual's concept image then, is a potentially huge collection of structures, properties, pictures or processes that are associated with the particular concept. A concept image might be completely informal, it might not be coherent. Parts of the concept image might not agree with other parts. This won't cause problems, however, unless the conflicting parts of the concept image are evoked simultaneously.

In contrast the concept definition is

> "We shall regard the *concept definition* to be a form of words used to specify that concept." (Tall & Vinner, 1981, p.152)

A further subdivision was made between a personal concept definition and a formal concept definition, the latter being "a concept definition that is accepted by the mathematical community at large."

Tall and Vinner discussed some common problems that can arise with undergraduate mathematics. For example, a first year undergraduate may have a very strong concept image of a 'familiar' concept that was introduced at A-Level, but a weak understanding of the concept definition that they have only just met. This can cause problems for the student:

> "the difficulty of forming an appropriate concept image, and the coercive effects of an inappropriate one having potential conflicts, can seriously hinder the development of the formal theory in the mind of the individual student." (Tall & Vinner, 1981, p.169)

In Alcock and Simpson's (1999) terms, in order to be successful, the student must develop the ability to 'turn on' the rigour prefix. That is to say that they must learn to reason with the concept definition and not their, possibly misleading, concept image.

2.2.4 Analytical/deductive proof schemes.

The last category of proof scheme that Harel and Sowder discussed is known as 'analytic'. However, in a later revision of the proof scheme taxonomy, Harel (in press) renamed the scheme, referring to it as the 'deductive' proof scheme. Either way, the scheme is

> "one that validates conjectures by means of logical deductions" (Harel & Sowder, 1998, p.258).

It is the proof scheme that would result in what most mathematicians would regard as a standard mathematical proof. They further subdivided this category into two: transformational schemes (those that use arguments based upon dynamic imagery rooted in the real world) and modern-axiomatic schemes.

Harel and Sowder (1998) noted that an axiomatic proof scheme relies heavily upon mathematical logic. A proof's starting point is certain undefined terms and axioms, and it proceeds by making logical deductions until the conjecture has been reached. For some students the axioms must be linked to their intuitive understanding of the situation (Harel and Sowder referred to this as an intuitive-axiomatic scheme), for others they may be able to study the axiomatic structure itself (a structural proof scheme). A yet deeper understanding of mathematics may lead to an axiomatising scheme, here the student is able to reflect on the consequences of varying the axioms.

Even when students have an analytic/deductive proof scheme, the process of producing proofs is far from straightforward. It has been noted that students who understand what is required of them in a proof have great difficulty in

'getting started' (R. Moore, 1994). It has been suggested that one reason why is that students haven't operationalised the definitions of the concepts they are dealing with (Bills & Tall, 1998).

Some authors have disputed Harel and Sowder's (1998) understanding of what a analytic/deductive proof actually is. Rav (1999), for example, pointed out that formal logic is not a fundamental part of constructing mathematical proofs at all. Rav wrote:

> "One does not even think about rules of logic in writing or reading a proof [...] A proof in mainstream mathematics is set forth as a sequence of claims, where the passage from one claim to another is based on drawing consequences on the basis of meanings or through accepted symbol manipulation, not by citing rules of predicate logic" (Rav, 1999, p.13).

Arbib (1990, p.55) suggested that one statement in a mathematical proof followed from another not through formal logic but through "formal technique and intuitions about the subject matter at hand".

So instead of agreeing with Harel and Sowder (1998) that mathematicians validate conjectures "by means of logical deductions", Arbib (1990) and Rav (1999) suggested that some kind of poorly understood "informal" logic is the dominant force (related arguments were made by Thurston, 1994).

Rav (1999, p.14) summarised his view of the situation:

> "As things stand now, we have remarkable mathematical theories of formal logic, but inadequate logical theories of informal mathematics."

The meaning here is clear. Whereas the formal understanding of the branch of mathematics known as 'proof theory' is well developed, this has little or nothing to do with how mathematicians actually create proofs. What is needed is a theory of how logic is used in "informal" mathematics.[2]

In terms of Tall and Vinner's (1981) distinction, Rav (1999) was suggesting that whereas the formal concept definition of argumentation – formal logic – is well understood, there is currently no adequate understanding of mathematicians' concept images of argumentation. This thesis seeks to address this gap in the literature by investigating in detail the role that logic plays in mathematics. Specifically, the goal of this thesis is to investigate the manner in which successful mathematics students reason with conditional "if...then" statements.

[2]Here, of course, Rav (1999) is using "informal" to mean normal day-to-day mathematics as done by algebraists, topologists etc. Not that done by formal logicians or proof theorists.

2.3 Summary of Chapter 2.

- Proof is recognised to be one of the most vital components of 'coming to know' advanced mathematics.
- Students from all levels of education have difficulty in dealing with proofs.
- The role of logic in proof construction and proof evaluation is uncertain. Some researchers argue that formal logical deductions are vital to develop a sophisticated 'proof scheme'; whereas others suggest that whilst *formal* logic has little or no role in mathematical proofs, *informal* logic does have an important, but not currently well understood, role.
- The goal of this thesis is to fill this gap in the research literature. Specifically, to investigate the manner in which successful mathematics students reason with conditional "if... then" statements.

The next chapter begins to look in detail at literature which explores how conditionals are understood by the general population, and by mathematicians.

Chapter 3

Logical Implication

Logical thinking has long been assumed to be a vital part of mathematical reasoning. This chapter summarises various different models of how conditional statements are understood in mathematics (and elsewhere) and introduces several important research tools that have been used to investigate this subject.

The idea that learning mathematics develops clear and logical thinking has a long history. In the early part of the 18th century, the liberal philosopher John Locke wrote that mathematics ought to be taught to "all those who have time and opportunity, not so much to make them mathematicians as to make them reasonable creatures" (Locke, 1706/1971, p.20).

This sort of belief was once the rationale for placing mathematics at the heart of the school curriculum. When discussing the utility of studying mathematics, C. Davis (1850/1970) wrote that it was important to

> "point out and illustrate the value of mathematical studies, as a means of mental improvement and development... [studying mathematics] aids the memory at the same time that it strengthens and improves reasoning powers." (pp.60-61).

Oakley (1946) took this idea further:

> "The study of mathematics cannot be replaced by any other activity that will train and develop man's purely logical faculties to the same level of rationality." (p.19).

Similar beliefs still pervade the mathematical world today. For example, the QAA, the UK quality assurance agency for higher education, states that

> "[Mathematics] graduates are rightly seen as possessing considerable skill in abstract reasoning, logical deduction and problem solving,

> and for this reason they find employment in a great variety of careers and professions." (QAA, 2002).

There is a widespread belief then, that studying mathematics improves your reasoning skills, teaches you how to think, and makes you more rational. But is this really correct? Although research has been conducted that has looked at whether subject specific knowledge can be transferred across domains (Lave, 1988), there has been surprisingly little that has looked at whether mathematics actually does develop these kinds of reasoning skills at all.

Some of the popular mathematics literature takes a clear stand on the issue of role of logic in mathematical thinking:

> "the ability to construct and follow fairly long causal chains [and] a step by step logical argument... is fundamental to mathematics." (Devlin, 2001, p.15).

Mathematical textbooks take a similar view:

> "Everyday language is full of generalities which are vaguely true in most cases, but perhaps not all. Mathematical proof is made of sterner stuff. No such generalities are allowed: all the statements involved must be clearly true or false... [we must] be sure that our mathematical logic is flawless." (Stewart & Tall, 1977, p.110).

Whilst philosophers have discussed the question (Arbib, 1990; Rav, 1999) there has, apparently, been little *empirical* research on how successful mathematicians behave when they are doing mathematics; and on to what extent logic is a part of a mathematician's thinking. Of particular interest within the field of logic is the role of *logical implication.*

In mathematics education circles, it has been argued that an understanding of logical implication is one of the most important prerequisites for understanding and constructing proofs. Rodd (2000), for example, suggested that modus ponens reasoning is crucial to establishing mathematical truth, and Küchemann and Hoyles (2002, p.242) emphasised logical implication's "importance for success" in mathematics.

However, it has also been recognised that logical implication is a topic that causes difficulties for students (e.g. Deloustal-Jorrand, 2002; Hoyles & Küchemann, 2002; O'Brien, 1973). One of hypothesised reasons for this apparent difficulty is the different models of implication[1] that are common in different

[1]There is a subtle philosophical distinction (described further below) between an implication and a conditional. However, this thesis argues that, for the current purposes at least, the distinction is not *psychologically* important. For this reason the words are used interchangeably.

P	Q	Material Conditional	Defective Conditional
t	t	t	t
t	f	f	f
f	t	t	i
f	f	t	i

Table 3.1: Truth tables for '$P \Rightarrow Q$' in the cases (a) the material conditional and (b) the defective conditional. Here t = true, f = false and i = irrelevant.

contexts. In this section several distinct models that are mentioned in the mathematics education, psychology and philosophy literatures are described, and comparisons are drawn between them. Note that some of these models of the conditional have been described by different authors, and several of them overlap. These issues will be discussed further in a later section (§3.1.9).

3.1 Different models of the conditional.

3.1.1 The material conditional (T1).

The material conditional[2] model comes from the formal definition that is commonly taught in first year undergraduate courses. A material conditional "if P, then Q" is true if and only if either $\neg P$ or Q is true. It is often introduced via a truth table (see Table 3.1), from which the equivalence

$$P \Rightarrow Q \equiv \neg P \vee Q$$

can be deduced.

In their undergraduate textbook, Stewart and Tall (1977) introduce the material conditional using the example 'if $x > 5$, then $x > 2$'. Arguing that 'if $x > 5$ then $x > 2$' is obviously correct, they consider the case of $x = 4$. For this value of x, '$x > 5$' is false, but '$x > 2$' is true. Strictly speaking, of course, this sentence is actually a generalised conditional (see §3.1.3) as neither '$x > 5$' or '$x > 2$' have truth values unless x is specified.

Here it is worth pointing out the contrast that some logicians emphasise between the material conditional and the (material) implication. Quine (1966) draws a philosophical distinction between the implication 'P implies Q' and the conditional '$P \Rightarrow Q$'. For Quine the latter is a mathematical statement in its own right whereas the former is a sentence that talks about the two statements P and Q using only their names. 'P implies Q' is not, in itself, a mathematical

[2]Also known as the 'propositional connective'.

statement. However it is valid "when and only when the conditional is valid" (p.37). Quine writes:

> "[We may] write: 'dreary' rhymes with 'weary', but here again we are using names of the rhyming words in question – the names being in this case formed by adding single quotation marks. It would not be merely untrue but ungrammatical and meaningless to write: Dreary rhymes with weary." (Quine, 1966, p.37)

This distinction is also hinted at in some undergraduate textbooks (e.g. D. L. Johnson, 1998), and is mentioned by Durand-Guerrier (2003) as one of her four "notions of the conditional", although she uses the term "logically valid conditional". Although this distinction is important from the philosophical position adopted by Quine (1966), it is largely irrelevant from the psychological standpoint that this thesis takes; namely a position that is attempting to *empirically* analyse how mathematicians use conditionals in their work.

There are four common ways of using a material implication, two legitimate and two fallacious. Given $P \Rightarrow Q$

Modus Ponens is the deduction of Q from the assumption P.

Modus Tollens is the deduction of $\neg P$ from the assumption $\neg Q$.

Affirming the consequent is the incorrect deduction of P from the assumption Q.

Denying the antecedent is the incorrect deduction of $\neg Q$ from the assumption $\neg P$.

Modus ponens appears to be an easier deduction to make than modus tollens, despite the potentially serious results of failing to deduce $\neg P$ from $\neg Q$, although the reasons why this might be so are controversial (see Chapter 4). Incredibly, it is possible that the Chernobyl disaster can be attributed to the failure of a workman to make this deduction (Johnson-Laird, 1999).[3]

Many studies have found that use of affirming the consequent and denying the antecedent are widespread, even amongst highly educated populations. O'Brien (1973) denoted consistent use of affirming the consequent and denying the antecedent deductions as 'child logic', as opposed to 'maths logic'. Using inference tasks (such as that in Figure 3.1), he found that between 40 and 50% of undergraduate mathematics students[4] consistently used child logic as opposed

[3]The Chernobyl plant had the safety regulation "if the test is to continue, then the turbine must be rotating fast enough", but despite the fact that the turbine wasn't rotating fast enough, no one deduced that the test should be stopped.

[4]O'Brien's participants had all just completed an 'introduction to mathematics' course. It is not at all clear whether maths was their major subject.

Here is a rule:

"if the car is shiny, then it is fast".

For each of the following answer *yes*, *no* or *can't tell.*

1. The car is shiny. Is the car fast?
2. The car is fast. Is the car shiny?
3. The car is slow. Is the car shiny?
4. The car is not shiny. Is the car fast?

Figure 3.1: A standard inference task.

to only 5% who consistently used maths logic. This compared to a previous experiment that gave similar tasks to secondary school children which found that around 70% used child logic (O'Brien, Shapiro, & Reali, 1971). In both experiments O'Brien found that using familiar content in the questions facilitated modus tollens deductions.

The claim that the conditionals in natural language can be successfully modelled using the material conditional has been referred to as Theory 1, or T1, by Edgington (2003) and Evans and Over (2004). T1, whilst believed by the likes of Boole (1854/1958) and Inhelder and Piaget (1958), has been subject to serious challenge by modern philosophers and psychologists alike (see Chapter 4).

3.1.2 The defective conditional.

The defective conditional[5] occurs when "if P, then Q" is considered irrelevant if P takes the value false (Wason, 1966). This model's truth table is shown in Table 3.1 (see p.15). Quine (1966) anecdotally noted that this is the form of the conditional that is general used in day-to-day life. He wrote that "'if P then Q' is commonly felt less of an affirmation of a conditional than as a conditional affirmation of the consequent" (p.12).

Quine's observation has been experimentally investigated by psychologists. The so-called 'truth table task' involves asking participants to decide whether certain given information makes a conditional true, false or whether the infor-

[5]D. Mitchell (1962) and Durand-Guerrier (2003) use the terms "hypothetical proposition" and "the common understanding" respectively.

Which of the following situations (a) supports, (b) contradicts or (c) tells us nothing about the rule:

"if the shape is a square, then it is green".

1. A green square.
2. A red square.
3. A blue circle.
4. A green circle.

Figure 3.2: A standard truth table task.

P	Q	Material Equivalence	Defective Equivalence
t	t	t	t
t	f	f	f
f	t	f	i
f	f	t	i

Table 3.2: Truth tables for '$P \Rightarrow Q$' in the cases (a) material equivalence and (b) defective equivalence. As before, t = true, f = false and i = irrelevant.

mation is irrelevant to the truth or falsity of the rule (see Figure 3.2). Wason and Johnson-Laird (1969), for example, used a conditional with abstract content and found that subjects often regarded that the information that P was false made the rule irrelevant. At the time this finding was used to dispute the Piagetian claim that children reached the stage of formal operations by the age of 12 (see also §4.1).

Wason and Johnson-Laird (1969) and Evans, Newstead, and Byrne (1993) noted that along with the material conditional and defective conditional interpretations of implication, the closely related 'material equivalence' and 'defective equivalence' interpretations are also common. These occur when "$P \Rightarrow Q$" is mixed up with "$P \Leftrightarrow Q$" (see Table 3.2 for truth tables).

Hoyles and Küchemann (2002) have argued that the defective conditional is a "more appropriate" interpretation than the material conditional because

> "in school mathematics, students have to appreciate the consequence of an implication when the antecedent is taken to be true." (p.196)

This is a rather peculiar claim, as both the defective conditional and the material

conditional allow one to "appreciate the consequences" of "if P then Q" when P is true. The difference occurs when P is not true.

Durand-Guerrier (2003) criticised Hoyles and Küchemann by pointing out that knowing the truth value of a conditional when P is false is vital to understanding definitions such as, for example, that of a diagonal matrix:

> A $n \times n$ square matrix $[a_{ij}]$ is diagonal if and only if $\forall\, i, j \in \mathbb{N}$ such that $1 \leq i, j \leq n$, if $i \neq j$, then $a_{ij} = 0$.

But this argument is surely fallacious. Such a definition is perfectly understandable with a defective conditional interpretation: when $i \neq j$, "if $i = j$, then $a_{ij} = 0$"is irrelevant. That is to say, nothing can be concluded about the truth of $a_{ij} = 0$. Such an interpretation does not prevent one from fully understanding, and making operable (in the sense of Bills & Tall, 1998), the definition of a diagonal matrix.

3.1.3 The generalised conditional.

A generalised conditional is found in the domain of predicate rather than propositional logic. Such a conditional has the form "$\forall\, x \in X$, if $P(x)$, then $Q(x)$", where X is the set that we are working in. Here, neither $P(x)$ nor $Q(x)$ have a truth value until x is specified. This is the form that most mathematical theorems are written in, albeit sometimes deceptively. For example, the theorem "if the diagonals of a quadrilateral bisect one another, then the quadrilateral is a parallelogram" is a generalised conditional of the form "for all quadrilaterals Q, if Q has diagonals that bisect one another, then Q is a parallelogram". One might argue that most of the examples of conditionals discussed so far have implicitly been generalised.

3.1.4 The causal/temporal conditional.

Deloustal-Jorrand (2002) explained that she

> "understands by 'causal conception of the implication' all the rules, practices and knowledges [sic] related to the interpretation of the sentence 'A implies B' by 'A is the cause of B'." (p.284)

She went on to point out that this causality need not be a temporal relationship, although in many day-to-day cases it is. Here the word "cause" is problematic. Is it reasonable to say that event P can cause event Q even though it doesn't precede it? The word "cause" seems to carry with it implications of some kind of temporal order. The notion of causality is a hugely complex subject that has associated with it a vast research programme in philosophical and cognitive

science areas (e.g. Sperber, Premack, & Premack, 1995). Whilst noting that this research programme exists, it is of minor relevance to this thesis. Instead it suffices to remark that the question of what causality means in mathematics is even more problematic than in other settings. In the mathematical domain, temporal order is something of a meaningless notion. In no sense can "$\mathcal{T}$ is locally connected" be said to have occurred either before or after "$\mathcal{T}$ is connected", so it seems to be slightly odd to suggest that one is a cause and one is an effect. Later it is argued that a better way of characterising the notion of causality in mathematics is in terms of belief rather than truth (see §3.1.6). This naturally leads us to consider Toulmin's (1958) notion of the warrant and backing of an argument.

3.1.5 Informal logic.

Informal logic is an attempt by philosophers to accurately describe the structure of arguments (R. H. Johnson, 1999), and is largely based upon the work of the philosopher Stephen Toulmin. Writing in the fifties he suggested, against the prevailing orthodoxies of the time, that the best way to study argumentation was not to use the material conditional of formal logic (Toulmin, 1958; Toulmin, Rieke, & Janik, 1984). Toulmin's *The Use of Arguments* was very poorly received, with one of his colleagues branding it "Toulmin's anti-logic book" (Toulmin, 2001, p.11). Since it's publication, however, Toulmin's work has become an important idea within of many academic disciplines, include rhetoric theory, linguistics and perhaps mathematics education.

For Toulmin, contrary to the formal logician's beliefs, the purpose of an argument is to convince an audience of the conclusion's veracity. He suggested that an argument consists of several parts, all designed to convince the audience. The arguer starts by putting forward the data (D) and showing, via the warrant (W), that the conclusion (C) follows. A warrant tends to be a statement of the form "given D, one can take it that C" (Toulmin, 1958, p.99). If the warrant is not immediately obvious the some justification, or backing (B), for it is required. The qualifier (Q) gives an indication of the level of certainty contained in the argument (of course, in mathematics arguments are traditionally seen as aiming to establish the full certainty of claims rather than a level of probability in them). The final part, the rebuttal (R), occurs when the conviction in the argument is non-absolute. Toulmin's scheme is illustrated in Figure 3.3.

An example of an argument expressed in this form is given in Figure 3.4. Here, the arguer is suggesting that Hislop was at fault for the goal (C), a close range header scored whilst he was the goalkeeper (D), because he ought to have caught the cross from which the goal was scored (W). This is because

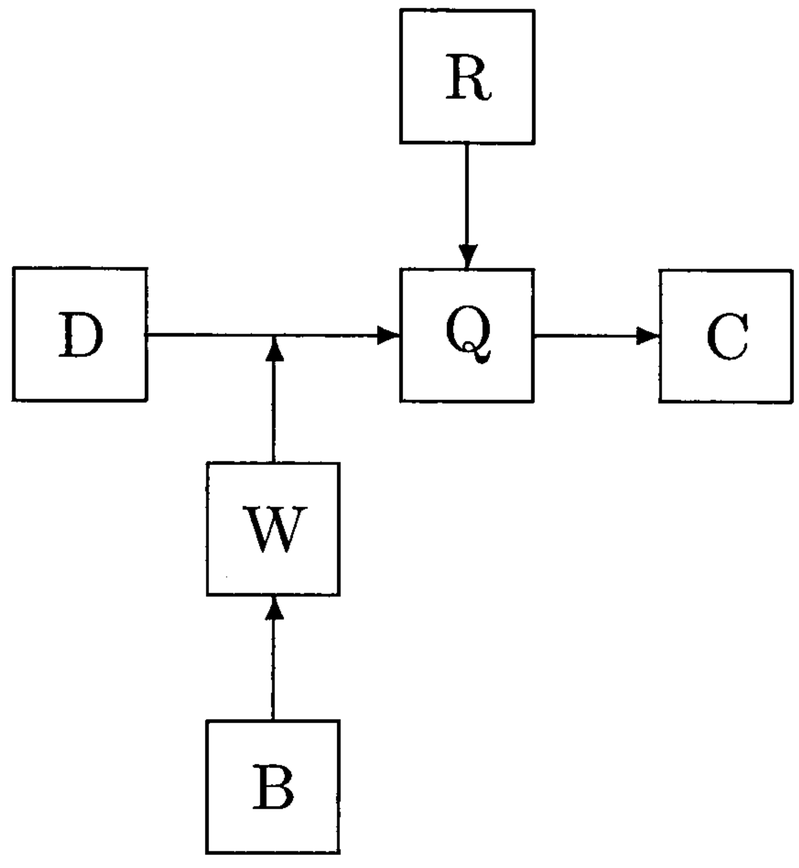


Figure 3.3: Toulmin's model of a general argument.

goalkeepers are expected to be able to catch crosses that are close to them (B). However, the arguer accepts that the argument could be rebutted if Hislop was fouled as the ball was crossed (R). Note that the backing in this argument, as in most arguments, is a general statement of "if...then" form; that is to say, the backing is "if a cross is close to a goalkeeper then he should catch it".

Toulmin's (1958) argumentation scheme, and how it has been applied to mathematical arguments, will be discussed in greater detail in §8.6.1. For now, however, it suffices to note that Toulmin's work has been used by some mathematics education researchers to put forward a model of the conditional.

3.1.6 The warranted conditional.

Using Toulmin's (1958) scheme, Weber and Alcock (2005) introduced the notion of a warranted conditional.[6] They wrote:

> "When one evaluates whether the implication 'if P, then Q' is warranted, P is seen as the data and Q as the conclusion. ...In determining whether 'if P, then Q' is warranted, the reader must not only evaluate the truth of P and Q, but also judge the soundness of this possibly inferred warrant." (p.36).

An implication is warranted in the sense of Weber and Alcock if it allows you to deduce the conclusion Q from the data P, i.e. it justifies the modus ponens inference. Modus tollens plays, at most, a subsidiary role.

[6] Actually, they used the word 'implication' instead of 'conditional' throughout.

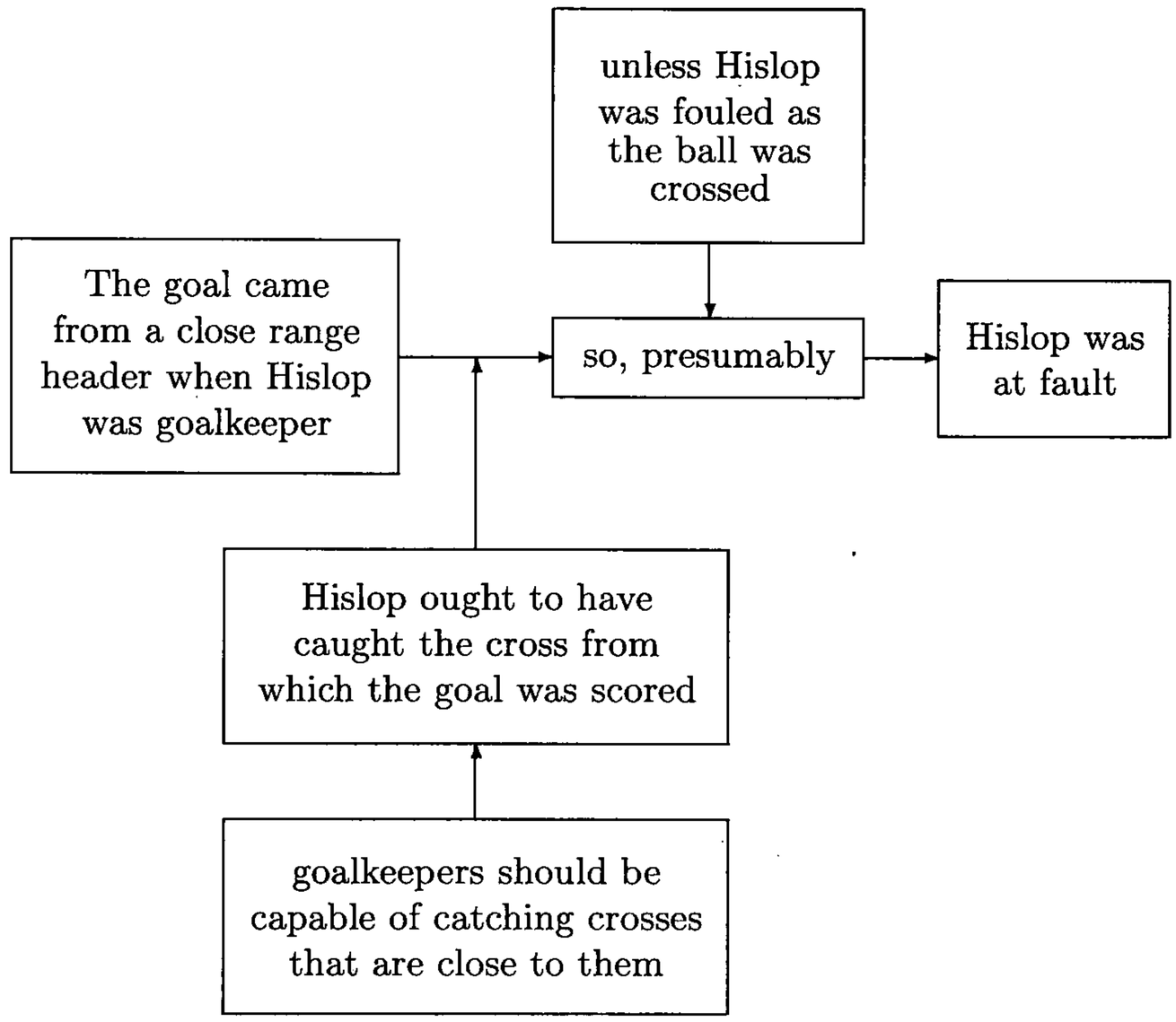


Figure 3.4: An argument expressed using Toulmin's structure.

Weber and Alcock argued that warranted implications are vital for validating proofs (in the sense of Selden & Selden, 2003). They pointed out that if a conditional is materially true but unwarranted, it cannot be used in a proof. They gave the example of a student who, when asked to prove that 1007 is prime writes "if 7 is prime, then 1007 is prime". When seen as a material conditional, this statement is true. However, it is not warranted. The implicit warrant (although, really this should be called a backing), that "if n is prime, $1000 + n$ is prime", is not justified (for a relevance theoretic explanation as to why this particular warrant is inferred, see Inglis, 2004).

Specifically then, a person interprets "if P, then Q" as a warranted implication if they infer (or look for, but fail to find) a warrant that allows them to conclude Q from the data P. Notice that this use is subtly different to that adopted by both Toulmin and Weber and Alcock. For Weber and Alcock, a conditional is warranted only if there exists a valid warrant (although it is very hard to be precise about what 'valid' means in this context). In this thesis the term *warranted conditional* is used to refer a particular model of the conditional – in which, when evaluating the truth of a conditional sentence, one is directed towards looking for a warrant which may or may not be found.

Weber and Alcock not only argued that the warranted implication is vital for proof validation, but that it is not taught sufficiently in class. They wrote:

> "for students to gain conviction and understanding from ... proofs, they must consider the implicit warrants used to justify the assertations in the proof. However, it is not clear that students will naturally do this." (Weber & Alcock, 2005, p.38)

However, Inglis (2004) argued that in fact it *is* clear that students will naturally do this. That is to say that the warranted implication is an entirely natural way of understanding "if... then" statements: it requires no special training (Inglis, 2004; Reid & Inglis, 2005).

3.1.7 The Stalnaker conditional (T2).

The Stalnaker (1968) conditional, a terminology adopted by Evans and Over (2004), is subtly different to the conditionals described in the previous sections. When P is false the Stalnaker conditional '$P \Rightarrow Q$' may be either true or false. Edgington (2003, p.383) gave the following example:

> If you touch that wire, then you will get an electric shock.

She points out that if you don't touch the wire the conditional might still be true or false. It would depend, for example, on whether the wire was insulated or not. In this situation, Stalnaker (1968) argued, we must make the minimal changes necessary to ensure our beliefs are consistent after P has been hypothetically added to them. That is to say, suppose we know that we didn't touch the wire, but also that the wire was not insulated and we were not wearing gloves. The conditional in this case would be true: P has been hypothetically added to our beliefs and, as a consequence, we have evaluated Q to be true.

Evans and Over (2004) summed up the situation by explaining that a Stalnaker conditional $P \Rightarrow Q$ is true in the case where P is false and Q is true (denoted FT) "if and only if TT is a closer possibility to FT than TF is" (p.26). Note that because of this complication, modus tollens is not a valid deduction from a Stalnaker conditional. The hypothesis that the Stalnaker model of the conditional best represents day-to-day conditionals has been called Theory 2, or T2 as opposed to T1 discussed in §3.1.1 (Edgington, 2003; Evans & Over, 2004).

The Stalnaker conditional model is not the same as the warranted conditional model in the case when both P and Q are true. The Stalnaker conditional is true in this case, but the warranted conditional will not be if there is no warrant.

3.1.8 The suppositional conditional (T3).

The suppositional conditional is an altogether different beast to the previous models of the conditional mentioned. It has no outward truth values at all, but instead is evaluated in terms of the so-called Ramsey Test. This test, proposed by Ramsey (1931/1990), suggests that the probability that a conditional is true is equal to the conditional probability of the consequent given the antecedent. That is to say that people judge the probability of $P \Rightarrow Q$ by "adding hypothetically P to their stock of knowledge and arguing on that basis about Q" (Ramsey, 1931/1990, p.247). Probability here should be taken to mean the degree of belief one has in the conditional. This relationship can be expressed symbolically:

$$\mathbb{P}(P \Rightarrow Q) = \mathbb{P}(Q \mid P).$$

Specifically, the suppositional conditional approach (as presented by Edgington, 2003; Over & Evans, 2003; and Evans & Over, 2004) suggests that people fix their degree of belief in a conditional statement by performing a two-stage Ramsey test: The probabilities of $P \wedge Q$ and $P \wedge \neg Q$ are evaluated and then compared. If $\mathbb{P}(P \wedge Q)$ is *high* compared to $\mathbb{P}(P \wedge \neg Q)$, then $\mathbb{P}(Q \mid P)$ is high and so $\mathbb{P}(P \Rightarrow Q)$ is judged to be high. Similarly, if $\mathbb{P}(P \wedge Q)$ is *low* compared to $\mathbb{P}(P \wedge \neg Q)$, then $\mathbb{P}(Q \mid P)$ is low and so $\mathbb{P}(P \Rightarrow Q)$ is judged to be low.

Over and Evans (2003) point out that the manner in which the probabilities of $P \wedge Q$ and $P \wedge \neg Q$ are evaluated is highly varied. Sometimes these evaluations are implicit (highly influence by System 1 processes) and sometimes they are explicit (highly influenced by System 2 processes). Over and Evans write:

> "There are a number of ways in which people can [evaluate these two probabilities]. Sometimes they will know relevant frequency information, and they can use that to make an explicit comparison. [...] More widely, heuristics [such as the availability heuristic] will sometimes be engaged." (p.346)

Evidence for the role of probability judgements in evaluating conditionals was presented by Evans, Handley, and Over (2003). They concocted a situation where $\mathbb{P}(Q \mid P)$, $\mathbb{P}(P \wedge Q)$ and $\mathbb{P}(\neg P \vee Q)$ were all radically different. They then asked participants to judge "how likely" the rule $P \Rightarrow Q$ was. The correlation between participants' evaluations of $\mathbb{P}(P \Rightarrow Q)$ and $\mathbb{P}(\neg P \vee Q)$ was found to be low. In a further experiment it was found that participants' evaluations of $\mathbb{P}(P \Rightarrow Q)$ were much closer to $\mathbb{P}(Q \mid P)$ than their evaluations of $\mathbb{P}(P \wedge Q)$.

The suppositional conditional is different to the Stalnaker conditional. Suppose, on the conditional '$P \Rightarrow Q$', you have managed to rule out the case $P \wedge \neg Q$. Is the conditional true or false? T2 cannot say. The conditional may fail in the

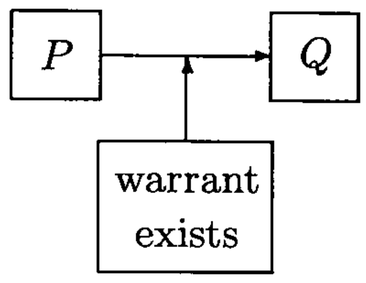


Figure 3.5: The warranted conditional in terms of Toulmin's scheme, the conditional asserts the existence of a warrant/backing.

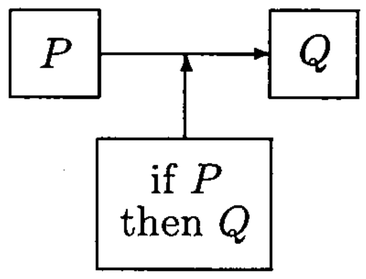


Figure 3.6: The material conditional in terms of Toulmin's scheme, here the conditional *is* the warrant.

cases where $\neg P$ is the case. T3, however, says that the conditional is true. P is hypothetically added to your stock of knowledge and the probability of Q is assessed, this is certain as we have ruled out $\neg Q$. Interestingly, when he developed the Stalnaker conditional, Stalnaker (1968) was attempting to combine the intuitive correctness of T3 with the advantages of a conditional determined entirely by truth values. This has since been shown to be impossible (Lewis, 1986).

3.1.9 Comparing the different conditionals.

There are many similarities worth noting between the these differing models. Firstly, it is worth pointing out explicitly the differences between the material and warranted conditionals. They both allow you to deduce Q from P. Phrasing both explanations in terms of (a reduced version of) Toulmin's (1958) scheme sheds some light on the matter. The warranted conditional understanding is shown in Figure 3.5. With this view, the sentence "if P, then Q" is making the claim that there exists a warrant that allows you to deduce Q from P; you are left to infer what the warrant might be. This is different from the material understanding (see Figure 3.6). Here, the conditional *itself* is the warrant that allows you to deduce Q from P; and, for that matter, $\neg P$ from $\neg Q$.

However, this is not to say that all the conditionals mentioned above are different. Weber and Alcock's (2005) warranted conditional and Deloustal-Jorrand's (2002) causal conditional are different names for similar underlying concepts. As discussed in §3.1.4, there are difficulties with the linguistic sub-

tleties of claiming that one mathematical fact causes another. However there are no such difficulties in saying that *believing* P is true causes you to believe that Q is true. In other words, the causal conditional acts as a warrant. It is a warranted conditional. Conversely, a warranted conditional causes to you to believe Q if you believe P: it is a causal conditional. Given the similarity between these views, and the difficulties with the word 'causal', this thesis uses 'warranted' throughout.

Notice that although the warranted and the generalised views of the conditional are not identical, in virtually all practical mathematical situations they coincide. After all, if something having property P allows you deduce that it also has property Q, then the same better be true for all things which have property P (i.e. all warranted conditionals are generalised). It is not the case that "for all x, $P(x) \Rightarrow Q(x)$" means there must be a warrant that links P and Q, but in the vast majority of cases a mathematician meets there is. Indeed, one might see most direct proofs as chains of warrants that indicate how having one property leads to having another. Without such links the only way of proving a statement such as "for all x, $P(x) \Rightarrow Q(x)$" would be to exhaustively test all the x's – a procedure that is very rarely seen in advanced mathematics.

So, to summarise, six apparently different models of the conditional have been identified:

The material conditional (T1). "If P, then Q" (or $P \Rightarrow Q$) is true if $\neg P \vee Q$ is true.

The defective conditional. "If P, then Q" is irrelevant whenever P is false.

The generalised conditional. "If $P(x)$, then $Q(x)$" is true if '$P(x) \Rightarrow Q(x)$' is true for all instances of x.

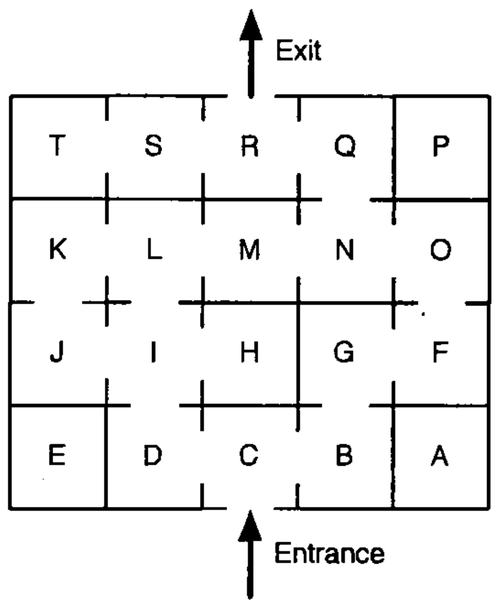
The warranted conditional, also known as the causal conditional, "If P, then Q" asserts the existence of a warrant that allows you to conclude Q from the data P.

The Stalnaker conditional (T2). Agrees with T1 for when P is true, but differs where Q is false.

The suppositional conditional (T3). The degree of belief in "if P, then Q" is given by $\mathbb{P}(Q|P)$. A world where P is true is imagined, and the likelihood of Q is evaluated.

Given these different models of the conditional, it is natural to ask which versions, if any, best match the way people understand and use conditionals, and what methods are there to measure this?

A person named X managed to pass through a maze and never used the same door twice. We can write down sentences about the situation. For each of the sentences you must decide whether it is *true*, *false*, or whether there are *not enough clues* to tell. For example, the sentence 'X crossed C' is a true sentence, as C is the only entrance to the maze.



Place each of the following into the categories: *true* (T), *false* (F) or *not enough clues* (N).

- X crossed P.
- X crossed N.
- X crossed M.
- if X crossed O, then X crossed F.
- if X crossed K, then X crossed L.
- if X crossed L, then X crossed K.

Figure 3.7: The maze task.

3.2 Standard logic tasks.

There are several standard logical tasks that have been used in the literature. The most famous and widely researched is the so-called Wason Selection Task, and this will be discussed at length in Chapter 4, together with the various theories of reasoning that have been developed to explain it, and other results from the literature. In this section, however, other common but less widely used logic tasks are briefly introduced and discussed.

3.2.1 The maze task.

The maze, or labyrinth, task was first given to 15-16 year old French schoolchildren, they were presented with the task shown in Figure 3.7 (APMEP, 1992; Durand-Guerrier, 2003).

The last sentence, "if X crossed L, then X crossed K", is the most interesting. Durand-Guerrier (1996) reported that 71% of the pupils answered "can't tell", which she argues is the correct answer.[7] For some routes (e.g. *CDIJKLMNQR*) the material conditional is true, and for some (e.g. *CDILMNQR*) it is false. Thus, suggested Durand-Guerrier, we cannot be sure which it is.

However, the teachers who administered the task thought that the answer to be "false". They explained that "the important matter is the link between the two sentences and not the particular truth value of each one" (Durand-Guerrier, 2003, p.9). Durand-Guerrier suggested that the teachers were viewing the sentence as a generalised conditional – they were inferring an illicit "for all" at the front of the sentence. When the sentence is seen like this, it is indeed false. However, in the teacher's quote, they speak of the link between the statements being of primary importance. This sounds more like a warranted conditional interpretation than the generalised conditional that is suggested by Durand-Guerrier.

Durand-Guerrier went further by concluding that all analyses of logical reasoning using propositional, rather than predicate, logic is bound to fail. She argued that such an approach (including that adopted by, for example, Toulmin, 1958) ignores the possibility of contingent sentences, and thus is necessarily inaccurate. Inglis and Simpson (2006) found that posing the maze task in a mathematical context biases participants away from responding with what Durand-Guerrier considered to be the mathematical correct answer, and suggested that Durand-Guerrier's focus on such narrow logical concerns missed some of the psychological subtleties of the task.

3.2.2 The truth table task.

The truth table task has been used by many experimenters. It comes in two flavours: evaluative and constructive. A standard evaluative version is given in Figure 3.2 (p. 18). Participants are given a rule "if P, then Q" (often with rotated negatives[8]) and are asked to identify which combinations of $P, \neg P, Q$ and $\neg Q$ support the rule, contradict the rule or are irrelevant to the rule (Wason & Johnson-Laird, 1969). In the constructive version, participants are asked to construct combinations that support or contradict the rule themselves (Evans, 1972). Wason and Johnson-Laird's (1969) findings regarding the defective truth table (see §3.1.2) were found in both versions of the task.

[7] Durand-Guerrier (2003) indicates that the "can't tell" answer was given more frequently by "those deemed good at mathematics" (p.9), although she offers no data to support this.

[8] Rotating the negatives here refers to using four different versions of each rule: instead of just using "if P, then Q", the rules "if P, then $\neg Q$", "if $\neg P$, then Q" and "if $\neg P$, then $\neg Q$" would also be used.

The vast majority of experiments with the truth table task have used abstract content (like that shown in Figure 3.2). But Newstead, Charles Ellis, Evans, and Dennis (1997) asked participants to solve the task with thematic content. They used rules that were classified as either promises, threats, tips, warnings, temporals, causals, and universals.[9] They found that the type of content had a significant effect upon the response. For example, promises and threats seemed to increase the frequency of the material equivalence interpretation. The effect of thematic content on the participants' responses to logical tasks is discussed further, with reference to the Wason Selection Task, in Chapter 4.

There are no examples of mathematics education researchers who have used the truth table task. However, another classic logical problem, the so-called inference task, has been used with various mathematical populations.

3.2.3 The inference task.

The inference task involves participants being asked to judge the validity of deductions from a conditional. O'Brien's (1973) version is shown in Figure 3.1.[10]

Typically almost all participants succeed in making the modus ponens, far fewer successfully use modus tollens (between 40-80%), and the numbers that incorrectly deny the antecedent and confirm the consequent varies considerably between studies. Evans, Newstead, and Byrne (1993) summarised the research in the field and noted that the percentage of participants found in different studies to deny the antecedent varied between 17 and 73%. Newstead et al. (1997) attributed this variation to subtle differences in the way the task was presented. In any case, they note that the frequency of denying the antecedent and confirming the consequent deductions is roughly equal. As with the truth table task, Newstead et al. (1997) found that the type of thematic content makes a significant difference to the results.

Hoyles and Küchemann (2002) gave a mathematically based inference task to schoolchildren in years 8 and 9. They were given the rule "if the product of two whole numbers is odd then their sum is even" and were told that the product of two numbers is 1271. 47% correctly made the modus ponens deduction, whereas 47% suggested that they needed to know what the numbers were before they would know.

---

[9]For example, one of the 'warning' rules was "if you wear Everton's colours to the match you'll be beaten up on the train". Options included: "Sandy didn't wear Everton's colours to the match; he wasn't beaten up on the train" and "Sandy did wear Everton's colours to the match; he wasn't beaten up on the train".

[10]O'Brien's version is slightly unusual in that it has neither abstract nor realistic-thematic content.

A slightly modified non-mathematical, thematic content inference task[11] was given to two mathematical populations, maths students and maths education students, by Stylianedes, Stylianedes, and Philippou (2004). They only tested the modus tollens deduction and the denial of the antecedent fallacy. 67% of maths education students and 76% of maths students successfully identified the modus tollens deduction as correct. Surprisingly the figures for identifying the incorrect nature of denying the antecedent were 76% and 60% respectively. These figures are only moderately higher than for the closest equivalent question given by Newstead et al. (1997), who found that 60% and 48% respectively correctly answered the denial of the antecedent and modus tollens questions (with causal content). Barring the (seemingly inconclusive) differences between the two groups it is difficult to see what can be concluded from Stylianedes et al.'s (2004) work, as they failed to have either a control group of non-mathematicians, or an isomorphic task with mathematical content.

Although, as discussed here, several mathematics education researchers have studied logical implication using the inference, truth table and maze tasks, few have used by far the most famous deductive reasoning instrument: the Wason Selection Task. The long history and literature surrounding this task forms the subject of the next chapter.

3.3 Summary of Chapter 3.

- Many researchers, philosophers and curriculum bodies have assumed that studying mathematics develops logical reasoning skills (C. Davis, 1850/1970; Locke, 1706/1971; Oakley, 1946; QAA, 2002).
- Conditional 'if...then' statements form one of the most important parts of logic.
- Several different models for how conditional statements are understood have been proposed by researchers from various communities. Many of these models overlap with one another.
- Several different 'standard' tasks has been used to investigate the issue of how conditional statements are used and understood. The most famous of these, the Wasón Selection Task, is the subject of the next Chapter.

[11]The rules were "if Costas suffered from pneumonia, he would have high fever" and "if the car doesn't have fuel, it will not move", both would seem to fall into Newstead et al.'s (1997) 'causal' category.

Chapter 4

The Wason Selection Task and Theories of Reasoning

The Wason Selection Task has been consistently used by reasoning researchers for the last forty years. Over this time it has influenced the development of many competing theories of reasoning, and led to the rejection of many others. The goal of this chapter is to set out the context in which the task was introduced, to describe the main empirical findings, and to review the major theories of reasoning that attempt to account for these findings. It is important to emphasise, however, that the reasoning theories discussed in this chapter are not intended to apply only to the selection task, their proponents would claim far wider domains of applicability.

4.1 The brain-computer metaphor.

Boolean logic was first described in detail by the British mathematician George Boole in the middle of the nineteenth century. His work is often considered as forming the basis of logic and computing. But at the time, following Aristotle, Boole believed that his logical laws were an accurate description of the thought processes of human beings. The title of his book – 'An Investigation into the Laws of Thought' (Boole, 1854/1958) – provides a succinct summary of his thesis. Although, of course, people sometimes make mistakes in their reasoning, Boole saw these as an aberration. On the whole, for him, the brain worked along the same lines as correct Boolean logic.

This view, that the brain was a some kind of logical machine, was widespread throughout much of the first half of the twentieth century. It was significantly boosted by the development of the first computers in the post-war years. In 1943

McCulloch and Pitts managed to prove mathematically that a sufficiently large network of formal neurons (equivalent to the base logical units of the brain) formed a Turing machine. That is to say that anything a computer could do, could, in theory, also be done by a network of neurons, in other words a brain (McCulloch & Pitts, 1943).

Piaget agreed with Boole, claiming that by the age of twelve children would have reached the stage of 'formal operations' and that their thinking would be abstract, formal and logical. He wrote:

> "reasoning is nothing more than the propositional calculus itself."
> (Inhelder & Piaget, 1958, p.305)

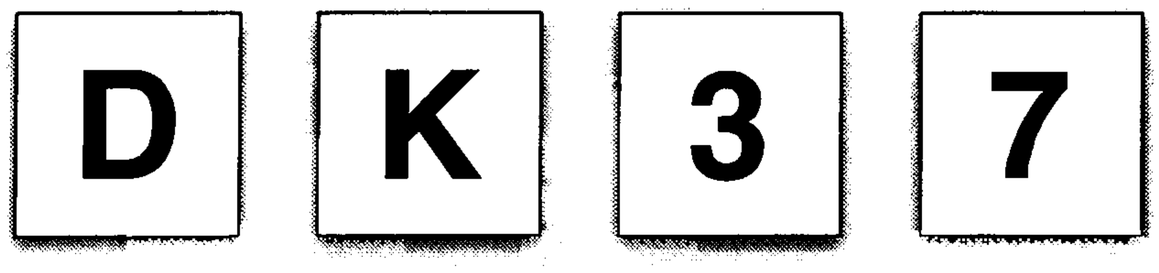
So, at the beginning of the sixties many cognitive scientists were convinced that the human brain reasoned along the lines described by Boole; in effect it was just a complicated logical machine. It took the work of psychologists such as Peter Wason to cast doubt upon this claim.

The Selection Task, one of the most widely studied experiments in psychological research, was piloted, and then reported, by Wason (1966, 1968). According to Johnson-Laird (2003), the Selection Task "has launched more investigations than any other cognitive puzzle". There have been literally thousands of papers written using data from various different versions of the task. By necessity then, the following review can only begin to give a broad overview.

4.2 The task.

Participants in the standard abstract version of the task are shown a selection of cards each of which have a letter on one side and a number on the other.

The participants can see:



They are given the following instructions:

Here is a rule:

"if a card has a D on one side, then it has a 3 on the other."

Your task is to select all those cards, but only those cards, which you would have to turn over in order to discover whether the rule is true or false.

selection	%
P	33
P, Q	46
$P, \neg Q$	4
$P, Q, \neg Q$	7
others	10

Table 4.1: Wason's initial results for the rule "if P, then Q" (Wason & Johnson-Laird, 1972, p.182).

The correct answer is to pick the D card and the 7 card, but across a wide range of published literature less than 10% of the general population do. Instead most choose D and 3 (denoted P, Q in Table 4.1).

Wason's results came as something of a surprise.[1] As noted above, Piaget had claimed that adult humans reason in ways that are abstract, formal and logical. If this were true, one would expect adults to perform nearly flawlessly on the Selection Task. Instead the 'failure' rate was over 90%. Wason wrote:

> "Some of the highly intelligent subjects tested in [my] experiments took a considerable time before they saw [the answer] was correct, and a few continued to dispute its correctness. And yet a computer could readily be programmed to solve the problem, as some subjects have been quick to point out after they had failed to solve it."
> (Wason & Johnson-Laird, 1972, p.173)

Wason was concerned that his results were due to some methodological problem. To counter this he began to tinker with the details of the task. To demonstrate that the task was sufficiently simple to be understood by the population he was dealing with, he reversed the experiment. Participants were given the question and the answer, and were asked to explain why the solution was correct. All twenty participants managed this successfully. From this result Wason concluded that the task is "deceptive rather than complex" (Wason & Johnson-Laird, 1972, p.174). Indeed it has since been found that participants will readily justify any answer given to them by the experimenter: correct or incorrect (Evans & Over, 1996a).

In another experiment the wording and symbols used in the question was tweaked. Instead of presenting the rule as "if a card has a D on one side then it has a 3 on the other", it was phrased "every card which has a red triangle

[1] In Wason's (1968) original experiment, participants (all psychology undergraduates) were interviewed in person. The instructor pointed to each of the four cards in turn and asked the participant whether knowing what was on the other side would enable him or her to find out whether the sentence was true or false. Administering the task on paper rather than verbally has been found not to affect the results and is now the norm in Selection Task research.

on one side has a blue circle on the other". The options were then red triangle, red circle, blue triangle and blue circle. No change in performance was detected (Wason, 1969).

Wason and Johnson-Laird (1970) modified the task so that all the information was included on one side of the cards, but so that some of it was covered up by a mask. Again, no improvement in performance was detected.[2]

Clearly then, Wason's worries about methodological flaws were unfounded. His abstract version of the Selection Task has been found to be highly robust.

4.3 Accepted results from the Selection Task.

The Selection Task is controversial, as some interpret the finding that participants fail to find the logically correct answer as being an attack on human rationality (see §4.5). Over the years, however, there have been several stable findings that are not in dispute, and in this section four of the most important are described.

4.3.1 Matching bias.

First noted by Evans and Lynch (1973), matching bias is the tendency to select cards that are mentioned in the rule, regardless of the presence of negatives. So with a rule "if P then Q", participants tend to select P and Q. However, if the rule is "if P then $\neg Q$" they also tend to select P and Q, which is in this case the correct answer. Evans and Lynch found, for example, that 61% of respondents answered correctly on the rule "$S \Rightarrow \neg 9$"[3] compared to only 13% on the rule "$S \Rightarrow 9$". Extraordinarily, the standard mistake, that of selecting P and Q, was made by *nobody* when confronted by a $S \Rightarrow \neg 9$ rule.

Evans, Clibbens, and Rood (1995) used three different types of rule, together with rotated negatives: "if P, then Q", "P only if Q" and "Q if P". All three showed significant matching bias effects. Evans et al. also found that instructing participants to verify or falsify the rule does not appear to reduce the effect of matching bias, and that the effect is more pronounced on abstract versions of the task (as opposed to thematic versions, see §4.3.2). Evans, Ellis, and Newstead (1996) found that varying the instruction type does not limit the

[2]Curiously, this experiment appears to have anticipated and dealt with Durand-Guerrier's (1996) later criticism. Durand-Guerrier argued that the reason why participants perform poorly on the task is that dealing with "hidden sides" complicates the logic involved. However, as we have seen, Wason and Johnson-Laird (1970) showed that placing all the information on one side does not improve performance.

[3]The implication symbol here is used to save space, in fact Evans and Lynch used the rule "if there is an 'S' on one side of the card then there will be a '9' on the other". A similar convention is adopted throughout this thesis.

effect of matching bias. Both instructing participants to verify the rule and to falsify the rule results in significant matching bias effects.

The logical structure of the rule used in the task appears to be important in determining whether matching bias is an important effect or not. Ormerod, Manktelow, and Jones (1993) found that matching bias was present on rules such as "P only if Q" and "Q if P". Roberts (2002) compared the effect between conditional rules ("if P then Q") and disjunctive rules ("P or Q"), both with rotated negatives. He found that although, in line with Evans's work, the conditional rules resulted in a large matching bias effect, the disjunctive rule appeared to produce a reverse matching bias effect. Roberts concluded that matching bias is not always generalisable beyond conditionals and that existing theories do not account for it.

Evans (1998b) noted that not all the existing theories of conditional reasoning can begin to offer an explanation of matching bias. As we shall see, pragmatic reasoning schemas (§4.4.3) and social contract theory (§4.4.5) do not even attempt to offer an explanation for Evans and Lynch's (1973) findings in their abstract contexts. In short, matching bias is an extremely robust finding when applied to abstract conditionals, and many theories have trouble explaining it.

4.3.2 The thematic effect.

Wason and Shapiro (1971) found that performance on the Selection Task could be improved by phrasing the task using thematic materials. Instead of using an abstract rule that refers to letters and numbers, Wason and Shapiro used the rule "every time I travel to Manchester I travel by train", with the cards 'Manchester', 'Leeds', 'train' and 'car'. On this task 10 out of 16 participants selected the correct answer: Manchester and car. A similar result was found by Johnson-Laird, Legrenzi, and Legrenzi (1972), who asked participants to pretend they were postal workers and gave them rules such as "a letter is sealed only if it has a $5d$ stamp on it". This rule elicited a correct response rate of 81%.

The naive assumption in the 1970s was that any thematic materials improved performance. However, later studies threw doubt upon this belief when they failed to find such an effect despite using identical materials to Wason and Shapiro (e.g. Griggs & Cox, 1982; Manktelow & Evans, 1979). It has since become clear that only certain types of materials robustly facilitate: in particular it seems that only rules which are *deontic* (those which convey rules, permissions, duties or obligations) produce higher success rates. Conversely, indicative rules (those which merely describe the world) seem not to facilitate

Rule: "If a person is drinking alcohol then they must be aged 18 or over."

You may check how old people are and what they are drinking. Which people in the bar would you need to check? There is a beer drinker, a coke drinker, a 14 year-old and a 22 year-old.

Figure 4.1: The drinking age version of the task.

performance.

The first robust finding of thematic facilitation came from Griggs and Cox (1982) and is shown in Figure 4.1. They found that many more people answered correctly on this version ($\approx$ 70%) than on the original abstract task. Griggs and Cox also looked at a transportation problem but found little facilitation. They concluded that when the rule was not familiar to the subjects it did not improve performance.[4]

Since Griggs and Cox (1982), thematic effects have been observed in many studies. However, as we shall see, the precise nature of the thematic content required for facilitation, and the reason why it occurs, remains highly controversial (see §4.4.3–4.4.5).

4.3.3 The training/education non-effect.

Cheng, Holyoak, Nisbett, and Oliver (1986) compared the performance of their participants on the task before and after six different types of training, including a full term's course in logic. Intriguingly, they found that only two of their training types made a difference: training in abstract logic coupled with explicit examples of selection problems, and training in the nature of obligations and the procedures needed to check whether violations of these obligations had taken place. The full term's course in logic had no significant effect upon performance.

In another experiment, the Selection Task was given to two groups of participants, those with Bachelor degrees and those with doctorates (from four different subject areas). No difference in performance was found between the two groups (Jackson & Griggs, 1988).

Of particular interest for the current study is the subject areas used by Jackson and Griggs. Twenty participants were mathematicians (10 with bachelor

[4]The large effect found by Wason and Shapiro (1971) can perhaps be put down to a small sample size. Interestingly, the rule used by Johnson-Laird et al. (1972) was actually a postal regulation in Britain prior to 1968. Griggs and Cox, therefore, argue that the effect was down to the familiarity with the rule rather than the thematic materials.

degrees and 10 with doctorates). On this very small sample, they found that 5 participants with bachelors and 7 participants with doctorates found the correct answer, significantly higher percentages than those found in other subject areas. There was no significant difference between the two levels of education for mathematicians. As the difference between mathematics and other subjects wasn't their research question, Jackson and Griggs failed to follow up on this finding, merely commenting that it is "probable that mathematics subjects are more familiar with the material conditional and the other rules in propositional logic" (p.329).

The lack of an education effect was questioned somewhat by Stanovich and West (2000) who found that there is a correlation between performance on the abstract Selection Task and SAT scores. They claimed that this showed that performance on the Selection Task was linked with 'cognitive abilities'. On an abstract problem, they found that those who found the correct answer (12% of their sample) had an average SAT score of 1159, compared to an average SAT score of 1098 for those answering incorrectly. This difference is significant ($p < 0.05$) and gives a surprisingly large effect size ($d = 0.558$) (Stanovich & West, 1998, p.210). Others, however, have noted that SAT scores are not a particularly good indicator of cognitive ability, and that a correlation is not surprising since SATs often contain similar reasoning tasks (Sternberg, 2000).

Although no research has set out to look at the differences in performance on the Selection Task between mathematicians and non-mathematicians, there have been several studies that did just this for scientists. None has found that they perform significantly better than the general population on an abstract version of the task (Kern, Mirels, & Hinshaw, 1983; Griggs & Randell, 1986).

Other researchers have investigated whether exposure to thematic versions of the Selection Task can facilitate performance on the abstract version. Very little transfer between the task types has been found (Evans et al., 1996; Johnson-Laird et al., 1972; Wason & Shapiro, 1971). However, it has been found that giving participants feedback regarding their answers can improve performance. Klaczynski, Gelfand, and Reese (1989) found that explaining the task to participants improved performance on the abstract version. Intriguingly, explaining the task to participants appeared to actually decrease their performance on thematic versions.

4.3.4 Changes in the wording of the task.

There is some evidence that performance on thematic versions of the Selection Task can be affected considerably by small changes on the wording of the task. For example, Griggs and Cox (1982) found that when participants were given

thematic materials and were asked to pick the cards that might be "violating the rule" – as opposed to picking the cards that would help determine "where the rule is true or false" – performance increased. Interestingly, the same change in wording on the abstract task had no effect (Manktelow, 1999). Jackson and Griggs (1990) had similar findings. On thematic versions of the task, instructions to pick out violators facilitated performance when compared to "true or false" instructions. However, the same was not the case on abstract versions of the task. As mentioned above, Wason (1969) found that changing the rule on the abstract task from "if a card has a D on one side then it has a 3 on the other side" to "every card with a D on one side has a 3 on the other" had no effect on performance.

4.3.5 Summary of §4.3.

Although there are many highly controversial results from Selection Task research, several robust findings stand out. Any theory that attempts to account for people's performance on the task must provide an account for these findings.

- On the standard task typically less than 10% of the well educated population select the normatively correct answer.
- Matching bias – the tendency for people to select the cards mentioned in the rule, regardless of the normatively correct answer – is a widespread phenomena on the Selection Task.
- The thematic effect: participants tend to select the normatively correct answer in significantly higher numbers when the task is phrased in realistic contexts, with thematic content.
- The training/education non-effect: there appears to be no correlation between level of education and performance on the Selection Task. However, it has been found that there is a correlation between solving the task correctly and achieving high SATs scores.

In the next section seven theoretical frameworks that try to account for these results are discussed.

4.4 Theories of reasoning.

There have been many theories proposed to explain the Selection Task. Most have been discredited since they were first proposed. However, there are several that still attract supporters. In this section the most important theories are described, in rough chronological order. It is important to note that all of these

theories are attempts to explain general reasoning, they do not simply apply to the Selection Task. However, since the task has become so ubiquitous in the literature the review particularly concentrates on explaining how each theory accounts for the surprising results the task has uncovered.

4.4.1 Mental models theory.

Mental models theory (e.g. Johnson-Laird & Byrne, 1991; Johnson-Laird, 2001) proposes that instead of following logical rules during reasoning, people construct models in their minds which they modify and reason from. When a new situation is encountered, the reasoner goes through three stages:

- They look at the premises ("world knowledge") and create a mental model of the possible situation they find themselves in.
- They form a non-trivial conclusion that is based upon the premises of their model.
- They look for counterexamples to their model and conclusion. If they cannot find any, then they accept the conclusion.

To explain the different models reasoners may construct, Johnson-Laird and Byrne (1991) use the so-called 'mental models notation'. In this notation, each line represents a different modelled case, and each item is represented in a column. The absence of an item simply means that it does not feature in that particular model. For example, the statement "there is A and B" would probably be modelled as:

$$A \quad B$$

Whereas the statement "there is A, or there is B" might be represented with two alternative models:

$$A$$
$$\quad B$$

To represent that an item has been exhaustively modelled, Johnson-Laird and Byrne's (1991) use square brackets. Thus, a model of "either there is A or there is B, but not both" might be initially modelled:

$$[A]$$
$$\quad [B]$$

But since B is exhaustively represented in the second model, the first model could be "fleshed out" to become:

$$[A] \quad [\neg B]$$

and similarly for the second model.

Johnson-Laird and Byrne (1991) explain that the rule "if P then Q" can result in several different models, and that, in the Selection Task, participants will consider selecting "only those cards that are explicitly represented in their models of the rule" (p.79). They explain that this is a consequence of the so-called principle of truth, a fundamental assumption of the mental models theory:

> "Individuals minimize the load on working memory by tending to construct mental models that represent explicitly only what is true, and not what is false." (Johnson-Laird, 1999, p.116).

So, from cards explicitly represented in the model, only the cards that have a hidden value which might effect the truth/falsity of the rule will be selected.[5]

So, for example, the rule "if P then Q" often results in this model:

$$[P] \quad Q$$
$$\ldots$$

Here the "..." represents a model with no explicit content. Note that this model is, in the terminology adopted in §3.1.9, a version of T1. The mental models theory argues that the material model of the conditional is correct, and that the problems with it discussed earlier can be explained away by which aspects of the reasoner's model have been 'fleshed out' (Over, 2004).

As a consequence of the model above, the reasoner considers the P and Q cards, but picks only the P card. The majority of participants who interpret the rule as a biconditional form the following model, and pick the P and Q cards:

$$[P] \quad [Q]$$
$$\ldots$$

The $\neg Q$ card will only be picked if the reasoner has fleshed out their original model sufficiently so as to represent it explicitly:

$$[P] \quad Q$$
$$\neg Q$$

Since P is represented exhaustively in the first model, it is tacit that the situation must be:

$$[P] \quad Q$$
$$\neg P \quad \neg Q$$

[5]This explanation is an adaptation of the no/partial/complete insight explanation detailed by Wason and Johnson-Laird (1970).

Thus, the mental model theory attributes poor performance on the task to the rarity of participants fleshing out their model of the implication. The frequency of the P and Q selection can be explained by the large proportion of participants who initially model "if P then Q" as the bi-conditional "P if and only if Q". Exactly why so many people misinterpret the conditional like this is left unexplained.

Mental models theory accounts for the matching bias effect by suggesting that when a negated sentence is involved in the rule the model is expanded to include the non-negated sentence. So, for example, $P \Rightarrow \neg Q$ might be represented by:

$$\begin{array}{ll} [P] & \neg Q \\ & Q \\ \ldots & \end{array}$$

Here the second model is incomplete, meaning that the reasoner is thinking about the situation that Q is true, but is not considering whether P or $\neg P$ is true in this case. Thus, by some creative application of the theory, the matching bias effect can be accounted for. However, the mental models theory has been criticised by Evans and Over (1996a) for its inability to parsimoniously explain biases.

As a consequence of their theory Johnson-Laird and Byrne made several predictions of facilitative changes that could be made to the Selection Task. The most striking was that changing the rule to an "only if" structure should improve performance, as it has a different typical initial representation. However it doesn't. Instead, both Evans et al. (1996) and Evans, Legrenzi, and Girotto (1999) found that it increased the frequency of $\neg P$ and Q selections, and slightly depressed performance overall. The theory has also been criticised for not being specific about how reasoners translate "world knowledge" into mental models.

It should be noted that mental models theory has been applied to a vast array of reasoning tasks, not just the Selection Task. In general it accounts for most of the effects reported on these tasks successfully.

4.4.2 Mental logic (mental rules) theory.

An alternative to Johnson-Laird and Byrne's (1991) mental models theory is the mental logic theory[6] (e.g. Rips, 1989, 1994) which claims that there is an inbuilt logical system that guides human's reasoning. He explained that

> "a person faced with a task involving deduction attempts to carry it out through a series of steps that take him or her from an initial

[6]The mental logic theory is also sometimes referred to as 'natural deduction theory', 'inference rule theory' or 'mental rules theory'.

Given that:
If the letter is a K then the number is a 7.
The number is not a 7.
Therefore?

Figure 4.2: A standard modus tollens inference task.

> description of the problem to its solution. These intermediate steps are licensed by mental inference rules, such as modus ponens, whose output people find intuitively obvious." (Rips, 1994, p.x)

So, a person constructs a sort of mental proof and then verifies it. That many people answer non-normatively to reasoning tasks is because they are not flawless at their proof construction.

Rips developed this theory using the "knights and knaves" puzzle.[7] He argued that the logical system was formed of abstract inference rules and schemas that are used across all domains, for all reasoning problems. Evidence for this theory came from qualitative 'think aloud' interviews that Rips conducted with participants on knights and knaves tasks.[8] By constructing a computer model of the supposed natural deduction system, he was able to accurately predict how long participants would take to solve problems based upon how many steps his computer program required.

The mental logic theory argues that the modus ponens rule is universal – it is one of the rules that can be applied directly. Modus tollens, however, requires a complicated argument to justify it. For example, consider a standard inference task such as that shown in Figure 4.2. Here, according to the theory, the question can only be answered by using a complicated contradiction argument as follows: "Suppose the letter is a K. It follows (by modus ponens) that the number is a 7. The number is not a 7 (by assumption). Therefore the letter cannot be a K." Naturally, the success rate at completing this argument is lower than for the straight forward modus ponens question.

Rips (1994) explained data from the Selection Task by pointing out that his theory suggests people should only pick the P card. As there is no conclusion to test, only forward rules can be used to solve the task. In particular, he claimed,

[7]This puzzle, beloved of recreational mathematics books, involves an island populated solely by knights and knaves. Knights always tell the truth, and knaves always lie. A vast array of problems such as the following can be posed: "We have three inhabitants, A, B, and C, each of whom is a knight or a knave. Two people are said to be of the same type if they are both knights or both knaves. A and B make the following statements: A: 'B is a knave'. B: 'A and C are of the same type.' What is C?" (Rips, 1989, p.86).

[8]Interestingly, Rips is one of the few psychologists to have used a clinical interview methodology in the area of logical reasoning (see also Stenning & van Lambalgen, 2001).

only the fundamental 'if-elimination' rule is used, and only with reference to the P card. The modal selection, that of the P and Q cards, is explained in Rips' scheme by suggesting people who answer this read the "if P then Q" rule as a biconditional. In this respect, the theory is similar to the mental models account. Indeed Oaksford and Chater (1995b) have argued that the two theories are, on a fundamental level, the same.

However there are problems with the mental logic account. Evans et al. (1995), for example, noted that because of the structure of the natural deduction system, mental logic theory suggests that participants in the Selection Task should make the mistakes of denying the antecedent and of affirming the consequent in roughly the same frequency. This does not happen. Participants affirm the consequent (select the 3 card) much more frequently than they deny the antecedent (select the K).

Chao and Cheng (2000) argued against the mental logic theory by showing that, for young children, modus tollens *and* modus ponens inferences were much more likely to be made on a permission based thematic task than on the abstract version. They used this result to argue that pragmatic rules (see §4.4.3) develop before generalised logical rules. This finding would appear to contradict Rips' claim that his natural deduction system is innate.

4.4.3 Pragmatic reasoning schemas theory.

According to the pragmatic reasoning schemas theory (Cheng et al., 1986), rather than reason according to logic rules, individuals use pragmatic reasoning schemas: abstract structures of knowledge derived from day-to-day life experiences. Examples of important experiences that give rise to prominent schemas would be permissions, obligations and causations.

Cheng and Holyoak (1989) propose four rules that they claim participants have as part of a conditional permission pragmatic reasoning schema that may be used when tackling a thematic version of the Selection Task, such as the drinking age problem:

1. If the action is to be taken, then the precondition must be satisfied.

2. If the action is not to be taken, then the precondition need not be satisfied.

3. If the precondition is satisfied, then the action may be taken.

4. If the precondition is not satisfied, the the action must not be taken. (Cheng & Holyoak, 1989, p.287)

They note that this conditional permission schema maps successfully onto the material conditional, but write:

> "When an 'if-then' statement evokes a schema that does not map onto the material conditional, or when no schema is evoked at all, then performance will be less likely to conform to the specification of formal logic." (Cheng & Holyoak, 1989, p.287)

Thus the reason why participants fare so poorly on the traditional abstract version of the task is that they do not evoke a pragmatic reasoning schema; the question is simply too far removed from their everyday lives. Conversely, however, if the task involves a permission or obligation conditional then performance will be facilitated. One of the predictions of the theory, that as long as participants have had experience of permission and obligation rules, performance will be facilitated, was tested by Chao and Cheng (2000) (mentioned above). They found that young children, who have had experience with these rules had their performance facilitated.

Another important prediction of the theory is that facilitation should result from permission and obligation rules rather than just from cost-benefit rules (see §4.4.5). Cheng and Holyoak (1989) tested this prediction by using so-called precautionary rules of the form "If one is to engage in hazardous activity P, then one must have protection Q". Despite having no obvious cost-benefit structure, these rules resulted in significant facilitation.

Despite this finding, Cosmides (1989) and Gigerenzer and Hug (1992) both argue the opposite case, that cost-benefit (or deontic) materials are necessary for facilitation. For proponents of this idea, precautionary rules are a special form of a more general cost-benefit rule structure (Fiddick, Cosmides, & Tooby, 2000). This so-called social contract theory will be described in more detail in §4.4.5.

The other two main criticisms that are levelled against the pragmatic reasoning schemas theory can also be deployed against all domain specific theories (including social contract theory, §4.4.5). Firstly, by concentrating entirely on the context the task is set in, the theory cannot hope to be a complete account of reasoning. When the task is given in an abstract context, reasoning is still going on, but no domain specific theory can explain it. Furthermore, no such theory can account for matching bias in an abstract context. Given the dramatic level of facilitation that results from rotating the negatives in the rule, this is a serious omission for any theory that attempts to explain the Selection Task.

Secondly, there are methodological problems with comparing thematic and abstract versions of the task. Jackson and Griggs (1990) looked at differences in wording between permission rules and abstract problems. They found that when wording changes (instructions to look for violators) were eliminated, per-

formance on the permission rules was just as poor as for the abstract version.

In addition, it can be argued that Wason's original question and (for example) the drinking age task are in fact two subtly different questions. In the original, the participants are being asked to test *the rule* (Manktelow & Over, 1990). In most thematic versions the rule is a given and participants are being asked to test *the cards*. For example, the rule "if there is a D on one side then there is a 3 on the other" is in doubt, the participants are being asked to test whether it is the case or not. However, the rule "If a person is drinking alcohol then they must be aged 18 or over" is not up for dispute, here the task is to evaluate whether the cards satisfy the rule, not to test whether the rule satisfies the cards (see also the categorisation task given by Sperber & Girotto, 2002).

4.4.4 Information value theory.

First proposed by Oaksford and Chater (1994), information value theory[9] starts from the idea that the Selection Task is a problem about decision making, not about logical reasoning. According to the theory, the key aim of participants is to increase the amount of information they have about the situation by reducing uncertainty. Indeed, for Oaksford and Chater, extra information is defined to be less uncertainty. Oaksford and Chater argue that it is unreasonable to label participants who fail to make the logically correct selection as 'irrational'. They suggest that participants will only turn over cards that have a high probability of resulting in a substantial information gain, by reducing uncertainty.[10] This, for them, is an entirely rational strategy.

Uncertainty is measured using information theory, and the expected information gain associated with each card is measured with Bayes' theorem, allowing for errors through a so-called 'noise' factor. Ordering cards by expected information gain reveals the same order ($P > Q > \neg Q > \neg P$) as the frequency of selection by samples in most studies.

Importantly for this theory, the amount of information a card can be expected to reveal is heavily dependent upon whether or not it is part of a large or small sample. The theory only makes sense if participants interpret the conditional as describing the situation of an entire population of cards, of which the four in front of them are merely a sample.

In this respect the theory has a close connection with the notorious ravens paradox (Hempel, 1945). This is the observation that observing a non-black non-raven – a yellow bus, for example – provides evidence to support the claim

[9]Also known as 'utility theory', 'rational analysis', 'information gain theory' or 'optimal data selection theory'.

[10]This idea has strong echoes of relevance theory's notion of cognitive effect (see §4.4.6). In fact, Oaksford and Chater (1995a) argue that expected information gain is a quantified measure of relevance.

"all ravens are black". No one, however, would go round looking for yellow buses if they were asked to evaluate a claim about black ravens, finding one wouldn't result in a sufficient information gain to make the exercise worthwhile. Note that the materials used by the experimenter in the Selection Task do not always clarify whether or not the cards are samples from a population. Despite this, Oaksford & Chater claim that this is the interpretation that the participants always make.

These considerations can be used to explain the thematic facilitation effect. For example, on the drinking age task it is not reasonable to assume that the probability of Q is low. This increases the information value of $\neg Q$, and Oaksford and Chater's (1994) calculations do not apply.

One of the major predictions of information value theory then, is that the experimenter can alter card selections by altering the population size from which the antecedent and consequent conditions in the rule are drawn. This was experimentally demonstrated to be an accurate prediction by Kirby (1994). No other current theory has accounted for this. Kirby gave a rule where the antecedent came from either a small, medium or large population. He found that selections of $\neg Q$ increased as the size of the P set increased.

However, these results are very controversial and have not been successfully replicated by, amongst others, Hattori (2002). Hattori found that increasing the P set made subjects more likely to choose P and had no effect upon the rate of $\neg Q$ selection, contrary to the predictions of information value theory. There are also other difficulties with the theory. Laming (1996) pointed out that since the P card will *always* provide the greatest information gain, it is peculiar that not every participant makes this selection. Only a large percentage do.

Evans and Over (1996b) criticised the theory by wondering whether uncertainty is a useful measure of information gain. They pointed out that if one's belief about a hypothesis' probability changed from 0.2 to 0.8 then the amount of uncertainty remains constant (at 0.2) despite a dramatically altered belief. Despite minor alterations in the theory to account for this criticism, Evans and Over (1996a) reiterated it, saying that new data leads to 'epistemic utility', an altogether more complicated construct to measure than information gain.

It is important to realise that the information value theory is does not attempt to explain *how* participants reach their conclusions on the Selection Task, it is merely an attempt to justify why these are not irrational responses. In this respect the theory is quite different to many of the others discussed in this chapter.

4.4.5 Social contract theory.

As mentioned previously, social contract theory attempts to explain why performance on the Selection Task is facilitated by using certain thematic materials. Proposed by Cosmides (1989), the theory suggests that humans have evolved a domain specific 'cheater detection' mechanism that allows humans to easily detect those who 'cheat'. For Cosmides, a social contract is a situation where two individuals agree to exchange a benefit for a cost. A cheater is someone who fails to fulfil their part of the bargain, i.e. someone who tries to take the benefit without paying the cost. Pinker (1997) explains that an advanced cheater detection unit is required if altruism was to have evolved by means of natural selection. In a survey of research, Cummins (1996) found that Selection Task studies that used deontic rules had significantly higher percentages of participants finding the correct answer than those who had used other types of rules.

Cosmides and Tooby (1992) argue that their social contract theory can explain both the abstract and the thematic versions of the Selection Task. Since the abstract version has no relation to any kind of real world situation, they argue that the evolved mechanisms the brain has to deal with this kind of implication aren't activated. (This explanation for poor performance on the original task leaves open the question as to how the people that *do* make the correct selection succeed. This omission is a common problem with Selection Task theories which concentrate on thematic versions of the task). However, when the rule is phrased in terms of a cost-benefit structure, the theory suggests that the brain's cheater detection unit is activated and performance is facilitated. Although, as mentioned previously, in response to Cheng and Holyoak (1989), Fiddick et al. (2000) modified their theory to include an evolved 'hazard management' system that allows for Cheng & Holyoak's results.

In the drinking age task, for example, they explain that the law "expresses a social contract in which one is entitled to a benefit (beer) only if one has satisfied a requirement (being a certain age)" (Cosmides & Tooby, 1992, p.183). As Stenning and van Lambalgen (2004) point out however, being over 18 doesn't automatically entitle one to beer. Presumably Cosmides and Tooby mean that the benefit is *the right to purchase* beer rather than the beer itself. When phrased like this the cost-benefit structure is somewhat complex, and it could be argued that it requires familiarity with a very particular Westernised type of social contract (in the sense of Rousseau, 1762/1997) that emphasises individual rights.

Social contract theory makes a couple of important predictions regarding the Selection Task. Firstly it claims that even an abstract Selection Task rule

can result in facilitation, if it is phrased in a cost-benefit structure; and secondly, that the familiarity of the material used in the rule should have no effect on performance. If it is phrased in a cost-benefit structure it will facilitate performance, if it isn't then it won't.

It is important to stress that social contract theory doesn't simply entail facilitated logical performance. It merely states that participants attempt to detect cheaters. For example, given the rule "if you buy a cappuccino, then you must pay for it", the correct selection would be "bought cappuccino" and "didn't pay for it" which agrees with the social contract theory prediction. However, switching the rule to "if you have paid for the cappuccino, then you may drink it" changes the logical answer (which is now "paid for the cappuccino" and "didn't drink it"), but the social contract cheater detection answer remains the same: "bought cappuccino" and "didn't pay for it". Cosmides (1989) found that participants do indeed perform as predicted on these so-called switched rule versions. She also claimed that only social contract problems exhibited evidence of facilitation, regardless of whether the participants were familiar or not with the thematic content. This finding is highly disputed.

Gigerenzer and Hug (1992) argued that although deontic rules are necessary for facilitation, they are not sufficient. They suggest that a key issue is whether or not "a person is cued into the perspective of a party who can be cheated" (p.127). Using the same social contract as Cosmides, they put the participants into the frame of mind of two groups: one group was involved in the social contract (they could be cheated), and one group was an impartial onlooker (they could watch either of the other parties be cheated). They found that performance was significantly higher in the group that was a party to the contract.

Despite all this support for social contract theory, it is much criticised and highly controversial. Firstly, and most importantly, it is not supported by some of the empirical evidence. Many studies have found evidence of facilitation without cost-benefit or deontic materials (e.g. Almor & Sloman, 1996; Cheng & Holyoak, 1989; Sperber, Cara, & Girotto, 1995; Sperber & Girotto, 2002; Wason & Green, 1984), and several of these papers were published *before* Cosmides' own study. Rather ironically, given that the Selection Task illustrates that a preference for confirmation can lead to logical errors, Cosmides and Tooby (1992) appear to have concentrated on confirming their hypothesis rather than looking for data that falsifies it.

Others have argued that facilitation on the Selection Task in cost-benefit rules is down to differences in task understanding due to the linguistic features of each version. Liberman and Klar (1996) wrote that performance on the Selection Task

> "largely depends on three aspects related to how people understand the task: (1) the clarity of the rule in terms of determination and direction; (2) the nature of the alternative to the tested rule and the falsifying instance it entails; (3) the perceived relevance of looking for violation strategy." (p.127)

They modified the original cheater story by removing linguistic features that they claimed would facilitate performance. Instead they added these features to a non-cheater story that was used by Cosmides. They found that participants in the non-modified cheater version performed to the same level as those in the modified non-cheater version. In other words, the key factor was the understanding of the purpose of the task rather than social contract rule. Similar results were found by Ahn and Graham (1999).

As noted above (see p.44), there are problems with any domain specific theory. In particularly, social contract theory (and the pragmatic reasoning schema theory) analyses the thematic and abstract Selection Tasks as if they were the same question. They are not. Sperber and Girotto (2002) illustrate that the cost-benefit cheater detection version is merely asking participants to categorise cards rather than conduct logical reasoning:

> "The cheating question is not a conditional reasoning question but a categorization question. As explained in detail by Cosmides and her collaborators, cheating is commonly understood as the co-occurrence of the taking of a benefit and the failure to fulfil a requirement, in particular of paying a cost. It is, in other terms, characterized by the conjunction of these two features. In order to answer the cheating question, then, all that participants have to do is select the cards that exhibit one of these two features (and that might have the other characteristic feature on the other side)." (Sperber & Girotto, 2002, p.282)

Note that finding the conjunction of these two features is not the same as finding the conjunction of P and $\neg Q$. 'Switching' the rule changes the logical answer, but leaves the categorisation answer the same.

Sperber and Girotto illustrate their point by giving participants an abstract categorisation and a cost-benefit version of the task. (Participants were asked to select cards that might represent food items that were not Italian, giving them the options: 'food item', 'non-food item', 'Italian item', 'non-Italian item'; 91% made the correct selections). They found that performance was facilitated most for the abstract categorisation version. This led them to argue that instead of having found an inbuilt cheater detection unit, Cosmides had merely found that

humans are very good at looking for cheaters when they are asked to look for them.

A final blow to social contract theory came from Stenning and van Lambalgen (2004), who criticised the theory by enquiring why the supposed cheater detection unit fails to catch liars. They argued that lying is form of cheating that is significantly more useful to be able to detect (from an evolutionary standpoint) than, for example, drinking age cheaters. However, Wason's (1968) original experiment demonstrated that explicit instructions to check whether the person putting forward the rule was "lying" or not, fails to improve performance. Given this, Stenning and van Lambalgen suggest that social contract theory "cannot plausibly" explain the difference in performance between the original and the thematic versions of the task.

These criticisms of social contract theory, and of thematic versions of the task in general, are highly persuasive. It is methodologically flawed to regard thematic versions of the Selection Task as being isomorphic to abstract versions. Whilst studying how people respond to these thematic versions may be interesting in its own right, it should not be of interest for those who want to concentrate on abstract reasoning. Consequently, no thematic Selection Tasks are used in the research presented in this thesis.

4.4.6 Relevance theory.

Relevance theory is an inferential theory of general communication (Sperber & Wilson, 1986; Wilson & Sperber, 2004). Such theories claim that communication takes place via a process where the audience *infers* the meaning that the communicator intended using the evidence available to them. By defining the notion of relevance, Sperber and Wilson can begin to explain how the inferential process takes place. It wasn't until many years after relevance theory was established in linguistic fields that it was applied to the analysis of the Selection Task.

The key feature of the theory is the concept of relevance. Each attempted communication carries with it a certain level of relevance: "an input is relevant to an individual when its processing in a context of available assumptions yields a positive cognitive effect" (Wilson & Sperber, 2004, p.251). There are two main factors here: when the cognitive effect of an input (be it verbal or written) increases, so does its relevance; however, when the effort needed to process the input increases, its relevance decreases.

The classic example from the literature is of telling a friend what time their train is. Saying "the next train to Dorridge is at 5:30 p.m" is generally more relevant than either

1. "the next train to Dorridge is sometime after 4:00 p.m." or
2. "the next train to Dorridge is scheduled to leave 2 hours and five minutes after 3:25 p.m."

Sentence (1) yields less of a cognitive effect, and sentence (2), whilst yielding the same cognitive effect, requires more processing to reach it (Sperber, Cara, & Girotto, 1995, p.49).

Suggesting that the maximisation of relevance is one of the main aims of human cognition, the theory attempts to explain communication with the communicative principle of relevance. The principle states that every communication "conveys a presumption of its own optimal relevance" (Wilson & Sperber, 2004, p.256). Optimal relevance here means two things: that every communicator believes that their communication is relevant enough to be worth the audiences' processing effort, and that the interpretation is the most relevant one that is compatible with their abilities and preferences.

When a hearer receives a communication they, subconsciously and following a path of least processing effort, infer a meaning and test it against their expectations of relevance. If the meaning does not meet this expectation, it is enriched and expanded (again, following the path of least effort) until it does.

Sperber, Cara, and Girotto (1995) used this theory to attempt to explain the Selection Task. According to Sperber and his colleagues, there are two sorts of participants. A minority that always get the task right, regardless of version. These select few are using a meta-inferential approach and "know the difference between demonstrative and non-demonstrative truth evaluation" (p.46). The majority of participants, however, attempt to infer meaning from the rule in a relevance theoretic manner.

They suggested that reasoners attempt to infer the consequences and conclusions from the rule, but do this from easiest to hardest in an attempt to minimise cognitive effort and therefore maximise relevance. Once they've reached a consequence of the rule that they deem relevant they are satisfied and select the cards that directly test the consequence that they've derived. There are three cases, in order of increasing complexity.

1. $P \Rightarrow Q$ achieves relevance by allowing you to deduce Q from P. This leads you to select the P card.
2. $P \Rightarrow Q$ achieves relevance by being interpreted as '$\exists\, x$ such that $P(x) \wedge Q(x)$' (there are cases of P and Q). This leads you to select the P and Q cards.
3. $P \Rightarrow Q$ achieves relevance by being interpreted as '$\neg(\exists\, x$ such that $P(x) \wedge$

$\neg Q(x)$)' (there are no cases of P and $\neg Q$). This leads you to select the P and $\neg Q$ cards.

By following this relevance theoretic analysis, Sperber, Cara, and Girotto (1995) were able to predict and test three specific ways of facilitating performance on the Selection Task. Firstly they suggested picking P and Q such that $P \wedge \neg Q$ is easier to represent than $P \wedge Q$. This has the effect of reducing the relevance of interpretation 2 and increasing the relevance of interpretation 3. Secondly, they suggested placing the task in a context where knowing that there are x that satisfy $P(x) \wedge \neg Q(x)$ will have at least as much of a cognitive effect as knowing that there are x that satisfy $P(x) \wedge Q(x)$.[11] The last method they suggest is to present the rule 'if P, then Q' in such a way as to reduce arbitrariness in the rule.[12]

By concocting versions of the Selection Task according to their 'recipe', Sperber, Cara, and Girotto were able to substantially facilitate performance. They also found that reducing effort was a more effective manner of increasing performance than increasing cognitive effect, thus proving that the human brain rewards laziness.

In short, Sperber, Cara, and Girotto argue that the Selection Task is not merely a test of formal reasoning, it is heavily dependent upon discourse comprehension, a feature of human cognition that is guided by relevance.

Several researchers have argued against the relevance theory explanation. Ahn and Graham (1999) found that simplifying the expression of $\neg Q$ (by using words such as 'unmarried' rather than 'not married') had no effect upon performance, contrary to the predictions of relevance theory. However, the cognitive effort saved (and therefore the level of facilitation gained) by transposing these terms must surely be minimal. Fiddick et al. (2000) also produced evidence that contradicted the predictions of relevance theory but, as described earlier, their methodology was convincingly refuted by Sperber and Girotto (2002).

A more reasonable criticism of relevance theory is that it does not offer sufficient explanatory power. Why are interpretations 2 and 3 less relevant than 1? Because they require more cognitive processing effort. But why do they? Sperber, Cara, and Girotto (1995) write

> "The conditional form 'if P, then Q' is more likely to be a felicitous way of conveying that there are no *P-and-(not-Q)* cases, when the fact that an item has the feature Q is inferable from the fact that it

[11] One way of doing this, for example, would be to make the $P \wedge Q$ cases trivial, thus lessoning their cognitive effect.

[12] They actually refer to this as "presenting the rule in a pragmatically felicitous manner" (p.61).

has the feature P (otherwise, why not just say 'There are no *P-and-(not-Q)*s'?)" (p.61)

But why should saying "there are no *P-and-(not-Q)*s" be more relevant than saying "if P then Q"? Why does it require less cognitive effort? Relevance theory fails to answer this question in any great depth.

4.4.7 Dual process theories of reasoning.

Recently, the idea that there are two distinct cognitive units that deal with reasoning has become fashionable amongst psychologists. Roughly speaking the first system corresponds with intuitive thought, and the second with abstract reasoning. Although there are many different versions of similar theories (e.g. Evans & Over, 1996a; Sloman, 1996; Stanovich & West, 2000) the generic terminology – System 1 and System 2 – adopted by Stanovich and West, has become commonplace.[13]

The key idea of (most versions of) dual process theory is that System 1 heuristically selects representations that are relevant to the situation you find yourself in. System 2, then, slowly operates on these representations to generate inferences and form judgements. System 1 filters irrelevant features of the environment so that the effort-intensive System 2 does not have to waste time on them. In short System 1 helps to form conscious thinking.

Note that the language used here to describe and discuss System 1 and System 2 is problematic. For ease of communication it is helpful to metaphorically talk as if each System were almost human like in its behaviours and attitudes. It is vital to emphasise that this is *only a metaphor*, used to aid to communication. It is clear that conceptualising System 2 as if it were a person making decisions for the brain is not at all helpful; it merely relegates the question from how the person thinks to how their System 2 thinks. No progress has been made. This difficulty, the so-called homunculus problem, becomes an issue only if the language used when discussing the two Systems is taken as literal rather than metaphorical (for an extended discussion of the homunculus problem see Pinker, 1997, or Stanovich, 2004).

System 1 is characterised by processes that are quick, operate in parallel and are highly context specific. These processes are preconscious in nature, only the end product is deposited in the conscious brain. The system is independent of language, and is old in evolutionary terms. System 1 is believed to be a large collection of subsystems that operate autonomously.[14] Most of these subsystems

[13]Stanovich (2004) now refers to System 1 as TASS (The Autonomous Set of Systems) and System 2 as the Analytic System. This thesis uses the original names.

[14]Hence Stanovich's (2004) recent adoption of the name TASS.

System 1	System 2
Associative	Rule based
Holistic	Analytic
Parallel	Serial
Automatic	Controlled
Relatively undemanding of cognitive capacity	Relatively demanding of cognitive capacity
Relatively fast	Relatively slow
Highly contextualised	Decontextualised

Table 4.2: Some properties of System 1 and System 2.

are innate. However some processes from System 2, if "habitually invoked" can, over time, become part of System 1 as well (Stanovich, 2004, p.66).

System 2, on the other hand, is slow, operates in serial and allows for non-contextualised hypothetical reasoning. It is controllable and conscious, has evolved relatively recently and it has been argued that it is unique to humans. It is this part of the brain that allows humans to construct complex abstract simulations that are context independent and depersonalised. Fluency with System 2 is often measured using reasoning tests, and tends to be correlated with measures of general intelligence (although it is perhaps not surprising that one form of reasoning test correlates with another). System 2 is also involved in expressing the output of System 1, and it has the ability to monitor and, potentially, override these intuitive responses, although, as we shall see, this does not always happen. Table 4.2 summarises some of the key differences between System 1 and System 2 (Stanovich, 2004, p.35).

Sloman (1996, p.11) argues that the notorious Müller-Lyer illusion (Figure 4.3) illustrates that perception and knowledge are located in different systems. Despite *knowing* that the two lines are the same length, the perception that the higher line is longer is undiminished. Stanovich (2004, p.41) concludes that, like System 1, "the perceptual input systems are another important part of your brain that ignores you".

A good example of how human cognition is shaped by the two cognitive Systems in the brain comes from examining how chess players decide which moves to make. Consider Figure 4.4. Black has just played ...N×a2, what should be white's next move?

Many studies on the psychology of chess have noted that when chess players look at a position they intuitively see which moves should be considered (e.g. de Groot, 1978; Hartson & Wason, 1983; Kotov, 1971). For example, in this position it is clear to chess players that White should consider moving the threatened queenside rook, and possibly ought to investigate going on the attack

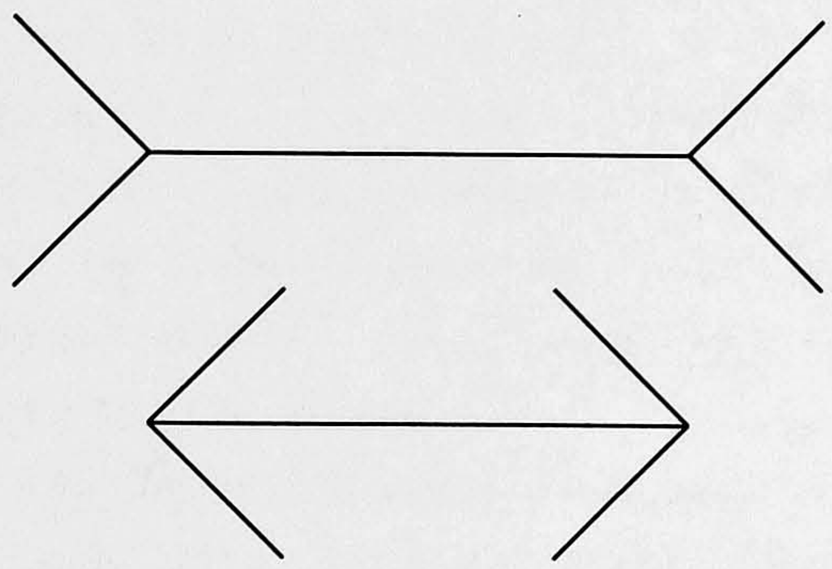


Figure 4.3: The Müller-Lyer illusion.

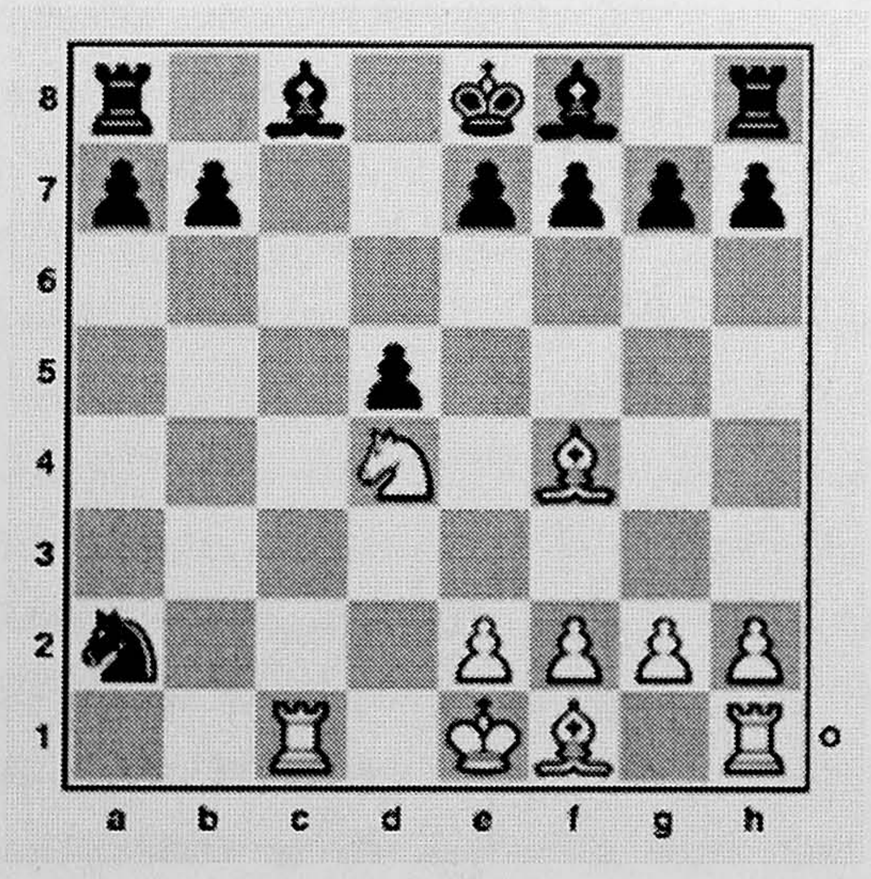


Figure 4.4: Black has just played 12 . . . N×a2.

by placing the knight on b5, threatening a fork on c7. Each of these candidate moves can then be analysed in detail before finally deciding which to play.

It is clear that there are dual processes present here. System 1 heuristically cues which moves appear relevant, and then System 2 takes over by performing the slow sequential task of an in depth analysis of the different possibilities. No chess player when faced with this situation would for an instant analyse the consequences of playing their kingside rook to g1. This is not because it is a bad move (although it is), but rather because they do not even consider it. System 1 does not preconsciously deem it to be relevant. Thus it does not get rejected by System 2; it is not even contemplated.

Without such a relevance based System 1 heuristic, chess (and indeed, life in general) would be impossible. There are simply too many possible moves. If every move had to be analysed by System 2 there would soon be a combinatorial explosion, that would make the processing effort required impractical (Evans, 1995). System 1, in this context, filters the workload of the less efficient System 2.[15]

Dual process theory suggests that there are two distinct parts of cognition when a participant tackles the (abstract) Selection Task. Firstly, various System 1 heuristics preconsciously direct the participant's conscious attention to certain apparently relevant parts of the problem (see also Sperber & Wilson, 1986). It is only after attention has been directed that the slow conscious System 2 processes take over and analyse the problem. The nature of the System 1 heuristics and the extent of the post-hoc involvement of System 2 allow the theory to account for the different results in different types of the Selection Task experiment.

This thesis will refer to situations such as this – where individuals are led to respond to a task in one manner by System 1, but then may come to realise (possibly with prompting) that there is an alternative more normative response (consistent with System 2 reasoning) – as satisfying *Criterion T*. This terminology is adapted from Sloman (1996), who uses the term *Criterion S* to refer to similar, yet subtly different, situations: "A reasoning problem satisfies *Criterion S* if it causes people to simultaneously believe two contradictory responses" (p.11). Note that Criterion S and T are not necessarily identical. Few participants in the Selection Task, for example, simultaneously believe both normative and non-normative responses, so the Selection Task does not satisfy Sloman's Criterion S; but it is clear that the Selection Task does set up a Criterion T

[15]The position in Figure 4.4 is from a real game. White played 13 Nb5 and went on to checkmate black after a further 22 moves.

situation for many people.[16]

Evans and Over (2004) explain that there are two fundamental System 1 heuristics that appear to influence behaviour:

The if-heuristic directs your attention, when interpreting hypothetical statements of the form "if P then Q", towards situations where P is true.

The matching-heuristic directs your attention towards the surface content of a statement, irrespective of the sense of that content. For example, the statement "I'm not a girl" is seen by this heuristic as being about girls, not about boys. Thus in both statements "if P then Q" and "if P then $\neg Q$", the matching-heuristic directs attention to P and Q.

So, System 1, as a consequence of these two heuristics, directs the participants attention towards certain apparently relevant parts of the question, the P and the Q cards. It is only then that System 2 takes over and analyses the relevant parts.

Note that the dual process account suggests that the if-heuristic and the matching-heuristic have different influences in different people. So, for example, those who answer P may be under the influence of the if-heuristic to a greater degree than than the matching-heuristic and vice-versa for those who answer P, Q.

Usually this System 2 reasoning does not affect the output. Most participants tend to merely *check* and *rationalise* their System 1 answer and output it. The evidence for this is twofold.

First, using semi-structured clinical interviews reveals that participants tend to construct post-hoc explanations for their selections (Wason & Evans, 1975). When considering statements such as "if P then Q" participants typically claim that they are looking to confirm the rule; whereas *the same* participants will explain that they are looking for falsifying evidence when confronted with the rule "if P then $\neg Q$" (Nisbett & Wilson, 1977). Dual process theorists account for this apparently strange finding by suggesting that the participants are subconsciously directed towards the P and Q cards for both rules, and then use System 2 to construct post-hoc rationalisations for these selections (Lucas & Ball, 2005).

Second, it has been found that participants spend significantly more time inspecting the cards that they eventually select (Ball, Lucas, Miles, & Gale, 2003; Evans, 1996; Roberts, 1998b), whilst the non-selected tasks have very low

[16]Indeed, Sloman's (1996) notion of Criterion S suggests that his form of dual process theory is one where System 1 and 2 compete for control of behaviour. In contrast, following Evans (1996, 2006), this thesis adopts a sequential version, where System 1 shapes the reasoning that takes place in System 2.

inspection times (although see the in depth methodological discussion on this paradigm in §6.5.1). This appears to confirm that participants rapidly attend to only those cards they eventually select, suggesting a central guiding role for the System 1 heuristics. Again, however, there are alternative accounts for this effect. Indeed the mental models theory also makes the prediction that many people will inspect the cards that they end up selecting, since these are the cards in their initial model. In this respect, the inspection time paradigm's prediction is perhaps not a good one. This matter is discussed at length in later sections (§6.5.1).

To summarise, when tackling the Selection Task participants' attention is directed at certain cards by their System 1 heuristics. It is only if System 2, when prompted to check and rationalise the output of System 1, overrides this output though a slow complex analysis of the situation, that the correct cards can be selected. This is clearly a Criterion T situation. The lack of success on the task suggests this overriding process is very difficult.

Although not specifically focussed on thematic versions of the Selection Task, it should be noted that the dual process theory does provide an account for the increased performance on these variants. It suggests that these versions of the task cue quite different System 1 heuristics – possibly even ones evolved to detect cheaters – and thus System 2 is cued to attend to different aspects of the problem (Evans & Over, 2004).

The dual process account of the Selection Task has several significant advantages over other theories. It can explain the matching bias result. It can explain the inspection time finding. By suggesting that the Selection Task is a complex mixture of two sorts of processing, it begins to suggest why there is apparently no training or education effect; perhaps education, initially at least, only affects System 2 processing (c.f. Pinker, 2002).

However, it also has weaknesses. There is no explanation as to where the System 1 heuristics come from, whether they are innate or learnt. It seems difficult to grasp how one could measure this. The theory is also lacking in details about how System 2 operates. It could be that this analytical stage works along the lines described by either mental models or mental logic theories.

So, although not without its critics and problems, the dual process account of the abstract Selection Task is on stronger theoretical ground than many of the other theories discussed in this chapter. For a longer review of dual process theories see Chapter 7 (or Evans, 2003, 2004b; Stanovich, 2004).

4.5 A note about rationality, and a defence of the Selection Task.

The issue of whether human beings are 'rational' or not is a recurring feature of intellectual debate for the last two millennia or more. Indeed Aristotle reported that he believed the distinction between rationality and irrationality was the defining feature of humanity: man is a "rational animal". Recently, however, the experimental results on psychological tasks such as the Selection Task have cast doubt upon this claim: if people do not reason successfully on such 'simple' tasks then they are surely irrational.

This sort of argument infuriates some critics who seem to have adopted a brief to speak on behalf of mankind's intellectual abilities. The classic response is of the form: "OK, people behave irrationally on these tasks, but this has nothing to do with the real world" (for a typical exchange along these lines see Reid & Inglis, 2005).[17] Evans (2004b) described this argument as "ridiculous":

> "The idea that psychologists [have] somehow contrived by incompetence or malevolence to consistently provide evidence of bias in normally bias free people in many hundreds of independent experiments [is] frankly ridiculous." (Evans, 2004b, p.257)

In a classic paper L. J. Cohen (1981) suggested that experimenters were, in some sense, tricking participants. He argued that if participants were prompted to reflect upon their response to the Selection Task then they would respond normatively. This argument was brutally dispatched by Wason (1983):

> "[Cohen] is obviously wrong to claim that "a few moments' prompted reflection" would enable subjects to admit that their reasoning had been invalid. ...errors are often systematic and resistant to correction" (p.59).

Wason also pointed out that L. J. Cohen's (1981) theory of human rational competence was unfalsifiable, and thus unscientific: any new experiment that upset the defenders of human rationality could be explained away by alleging 'tricks' on the part of the experimenter.

Piattelli-Palmarini (1994) noted that although modifying the Selection Task slightly (by using a thematic context) can remove the apparently 'irrationality', this is irrelevant to the rationality debate. He pointed out that this argument is analogous to suggesting that the Müller-Lyer illusion (Figure 4.3, page 55) is

[17] Criticising tasks' realism is not the only way that people attempt to rescue human rationality. Some even attribute 'poor' performance on such tasks to incompetent teaching! (Bringsjord, Noel, & Bringsjord, 1998).

not really an optical illusion because if you remove the arrow heads at the end of each line then the lines appear to be identical in length. Such a claim would of course be absurd, and yet the analogous argument in relation to reasoning tasks is regularly put forward by the defenders of human rationality (e.g. L. J. Cohen, 1981; Lopes, 1991).[18]

Commenting on the sort of responses to his work that L. J. Cohen typified, Wason (1981) wrote:

> "There is something interesting, however, about the reaction of some people to my work. Jonathan Cohen is not the first academic to criticise it, but such criticisms have sometimes been rather affective in tone. In an earlier draft of the present paper [L. J. Cohen, 1981] Cohen referred to my experiments, not as cognitive illusions, which is splendid, but as conjuring tricks, which is a little derogatory. Others have been more impolite. Why? Those who are most concerned to vindicate the basic rationality of man seem to me a little worried by what might be construed as evidence to the contrary." (p.356)

Wason (1981) went on to say that he himself would not subscribe to such a construal. His substantive point, however, on the improper nature of derogatory commentaries, is well made.

The goal of research instruments such as the Selection Task is not to belittle human intelligence, it is to deepen understanding of human intelligence. When participants respond to the Selection Task or to other similar experimental tasks, *they are behaving in some fashion.* Our job as researchers is to understand this behaviour. Our job is emphatically *not* to criticise the situation in which they are behaving because the behaviour we observe does not fit with our pre-existing quasi-moral beliefs.[19]

The experimental results discussed in this chapter clearly demonstrate that, if rationality is defined as reasoning in accordance with the formal logical calculus, then all humans, intelligent and unintelligent alike, are irrational. However, it is reasonable to ask whether formal logic is an appropriate normative standard for 'rationality'. Very few theorists would suggest it is. Simon (1983), for example, points out that human rationality is necessarily bounded by combinatorial explosion considerations. He introduced the notion of 'satisficing', the idea that

[18]Although even if this argument was taken at face value it would not suggest that humans are 'rational'. It would suggest that they are sometimes 'rational' and sometimes 'irrational'. But no researcher has ever claimed that reasoning differs from normative standards in *every* conceivable situation.

[19]Indeed, Stanovich (2003) has even argued that, in fact, it is the defenders of human rationality who are being immoral, by underplaying the importance of reasoning expertise: "For intellectuals to use their abstract reasoning skills to argue that the 'person in the street' is in no need of such skills of abstraction is like a rich person telling someone in poverty that money is not really all that important." (p.55).

human reasoning is a trade off between 'satisfying' and 'sufficing'. Instead of analysing the situation you are faced with until you obtain the perfect (logical) answer it is more 'rational' to find a sufficient answer that requires less processing effort (see also the arguments put forward by evolutionary psychologists; e.g. Cosmides & Tooby, 1992; Gigerenzer, 1991).

Evans and Over (1996a) discuss human rationality in detail, and suggest that the construct, as traditionally understood, does not make sense. Instead, they argue, the word rationality has been used to mean two quite different things, and much of the so-called rationality debate can be attributed to people confusing them. Evans and Over distinguish between rationality$_1$ and rationality$_2$:

Rationality$_1$ "Thinking, speaking, reasoning, making a decision, or acting in a way that is generally reliable and efficient for achieving one's goals."

Rationality$_2$ "Thinking, speaking, reasoning, making a decision, or acting when one has reason for what one does sanctioned by a normative theory" (Evans & Over, 1996a, p.8).

Note that whilst rationality$_1$ and rationality$_2$ are clearly linked in some fashion to System 1 and System 2 respectively, there is no direct mapping. For example, it is unreasonable to suggest that rational$_2$ responses come only from System 2; but it is clear that System 1 heuristics are, in most situations, rational$_1$ (otherwise they would not have survived the evolutionary process). Note that the idea of dual rationality does not solve the problem of which theory to use as a normative system with which to judge whether something is rational$_2$ (Anderson, 1991; Evans, Over, & Manktelow, 1993).

It is possible to be rational$_1$ and irrational$_2$ at the same time, or vice-versa. Indeed, irrational$_2$ arguments that are rational$_1$ are often seen in day-to-day life. When politicians in parliament respond to a difficult question by joking about the inadequacies of the opposition they are clearly being irrational$_2$. By any normative standard of political debate, *ad hominem* attacks are not satisfactory answers to difficult questions. However such attacks, by mocking and belittling the questioner are entirely rational$_1$ as they can potentially distract attention from a difficult question by challenging the credibility of the question's origin. Witness, for example, Tony Blair's habit, during the 2005 general election campaign, of responding to questions from Michael Howard by reminding him of his role in implementing the Poll Tax. Such a response is surely irrational$_2$, yet

entirely rational$_1$.[20]

According to the idea of dual rationality, irrational$_2$ responses to tasks such as the Selection Task may in fact be entirely rational$_1$. The corollary to this analysis, therefore, is that the critics who instinctively feel that they need to defend human intelligence are missing the point. When they criticise tasks that call into question human rationality, they need to ask themselves whether their brief of defending human rationality has the aim of defending human rationality$_1$ or defending human rationality$_2$. It is impossible to do both.

4.6 Summary of Chapter 4.

- Wason (1968) found that on a "deceptive not complex" reasoning task very few of his sample performed in accordance with formal logic.
- From nearly five decades of work with the Selection Task four relatively non-controversial findings have been established:
  - Matching bias. Participants tend to select cards that are mentioned in the rule, regardless of the presence of negatives.
  - The thematic effect. Phrasing the task in day-to-day contexts tends to facilitate performance.
  - The training/education non-effect. There appears to be no relation between high levels of education and performance on the task. Similarly, standard training methods (such as a term's course on logic) appear to have no effect.
  - Changes of wording. Thematic versions of the task seem to be very vulnerable to wording changes. Even the most minute change in instruction can have dramatic effects upon performance.
- There have been many theories that attempt to explain the results of the Selection Task. None has gained acceptance and the task remains highly controversial.
- Dual process theory explains the Selection Task by positing the existence of two quite distinct methods of reasoning. It is argued that the standard mistake arises out of System 1's automated response, and the failure of System 2 to adequately monitor and override it.

[20]The issue of rationality is one where different versions of dual process theory differ. Stanovich and West (2000) and Stanovich (2004) talk of normative rationality and evolutionary rationality. The former refers to maximisation of the goals of the individual organism whereas the latter is defined in terms of the interests of the genes (see §7.1 for a longer discussion of this point). Contrary to Evans and Over's (1996a) account, Stanovich and West propose a direct one-to-one mapping between System 1 and evolutionary rationality, and System 2 and normative rationality.

Chapter 5

Methods and Methodologies

5.1 The research question.

The main aim of this thesis is to elaborate on how mathematicians use logic when reasoning mathematically. In a sense, this research can be seen as a direct response to Rav's (1999) assessment that

> "As things stand now, we have remarkable mathematical theories of formal logic, but inadequate logical theories of informal mathematics" (p.14).

The goal then, is to begin to develop a theory of how logic is used in informal mathematics, and, in particular, develop a theory of how mathematicians *evaluate* mathematical conditionals. Informal mathematics in this thesis means (following Rav) the type of mathematics that is done everyday by mathematicians, not the formal type of logic analysed by logicians and proof theorists. Given this research aim, two stages are necessary:

- Firstly to critically and empirically assess the numerous theories of reasoning discussed in Chapter 4, with the aim of determining which is best suited to analysing mathematical reasoning.

- Secondly, to conduct a study of how successful mathematicians evaluate mathematical conditionals in 'realistic' mathematical situations, and to analyse this data using the adopted theoretical framework.

This chapter discusses the methodological options available, both in terms of methods and analytical approaches. Issues associated with validity and reli-

ability are also assessed. The precise detail of the methods adopted, however, are discussed in later chapters.

5.2 Methods of data collection.

The options available for data collection in this study are twofold: standardised reasoning tasks, such as the Selection Task, or interviewing. Both of these methods have their strengths and weaknesses, and it is important to consider them in detail.

5.2.1 Standardised tasks.

The psychology literature is full of attempts to 'measure' participants' thinking and reasoning processes using standardised tests administered through questionnaires (e.g. Manktelow, 1999). The examples of the maze task and the inference task were given in Chapter 2; and the Wason Selection Task was discussed at length in Chapter 4.

The primary advantage of using such methods is that they are, almost by default, highly reliable. Each participant is given exactly the same task and other researchers are able to repeat (and thereby confirm the reliability of) the experiment directly. There have been very few examples of psychology experiments that used standardised reasoning tasks being found to be unreliable.[1]

However, a significant disadvantage of the use of standardised reasoning tasks is the question of validity. What exactly do these tasks measure? And is it what we are interested in? This question has been largely absent from Selection Task research for many years. Since Wason's (1968) original paper, what exactly the Selection Task measures has become rather unclear. Instead the goal has shifted to explaining *why* people respond to it as they do, and in particular to explaining differences in performance between task versions. In a sense then, many researchers have avoided the question of validity by merely (and uncontroversially) assuming that the Selection Task measures how people respond to the Selection Task.

This approach was criticised by Sperber and Girotto (2002) who argued that the Selection Task is not a reasoning task, and therefore shouldn't be used to analyse reasoning. However, dual process theorists argue that Sperber and Girotto would be wrong to say that the Selection Task does not involve reasoning. As Evans and Over (2004) pointed out, according to the dual process

[1] One example of this that comes from Selection Task research was that of Wason and Shapiro's (1971) thematic version. Contradictory results with much larger samples were found by Griggs and Cox (1982) and Manktelow and Evans (1979). This example indicates how reliability can be linked to sample size.

theory, it does not *only* involve reasoning. As we have seen, dual process theory posits that both relevance effects (from System 1) and reasoning effects (from System 2) need to be considered when analysing the Selection Task. Thus, if one adopts a dual process (or mental logic, mental rules or information value) framework, the Selection Task can be considered a useful tool in researching reasoning. However, in order to ensure validity, extreme care must be taken when interpreting research results. It is not as simple as saying that if a person fails to correctly solve the Selection Task then they are 'bad' at reasoning.

5.2.2 Clinical task-based interviews.

Using interviews to discern thinking processes has a longer tradition in mathematics education than it does in reasoning research. It has quite distinct advantages and disadvantages over the standardised-reasoning-task approach.

The clinical interview method was adopted by Piaget (1929) in an attempt to understand the development of childrens' minds. A clinical interview, in Piaget's sense, begins with the experimenter giving the participant an open ended task to complete, whilst 'thinking aloud'. The experimenter then asks further questions contingent on the participant's response. This ability to extend and develop themes in the interview through the use of contingent questions allows for a much more detailed exploration of what thinking processes may be happening. It can be described as a 'semi-structured' interview methodology (L. Cohen, Manion, & Morrison, 2001).

Ginsburg (1981, p.5) suggested that clinical interviews can be used to address three distinct aims: the *discovery* of cognitive processes, the *identification* of what is behind these cognitive processes and the evaluation of the participants' *competence.* Ginsburg went on to describe a distinct clinical method for each of these aims, although also admitting that a study may have as its own purpose a combination of these three. The nature of this methodology naturally raises important validity and reliability issues (Swanson, Schwartz, Ginsburg, & Kossan, 1981).

Since the manner of questioning adopted whilst clinically interviewing is contingent upon the participant's response, it is clear that each interview cannot be properly replicated. It is, of course, possible to standardise the initial task used, and even some of the follow up questions; however, the possibility of contingent follow-up questions when the participant raises an interesting issue is what gives this methodology its strength. Reducing the flexibility of the interviewer's tools in the name of reliability will, paradoxically, impact upon the interview's validity.

Validity is, perhaps, an even more serious concern with the clinical interview

method than reliability. Several factors need to be considered. Piaget (1929) noted that it is very difficult to avoid asking leading questions that are based upon preconceived ideas of what you are looking for. This is especially problematic given the instantaneous response needed when coming up with contingent questions. He also pointed out that it is easy to miss opportunities for further in depth questioning that might have led to useful and insightful data. Recognising that there is no real way to deal with these problems other than through practice and training, Piaget (1929, p.9) noted that the clinical interview method was "difficult", and that it requires training lasting upwards of a year.

One of the biggest problems with the clinical interview approach is that it can rely, in some situations, upon introspection: the self-reporting of mental activity. Introspection has been highly controversial in psychological research, and there has been much discussion as to its validity and reliability (Nisbett & Wilson, 1977). The view adopted by early psychologists was the workings of a person's mind were transparent to them, and thus their reports could be trusted. Sir Francis Galton, one of the pioneers of psychology, wrote:

> "I do not see why the report of a person upon his own mind should not be as intelligible and trustworthy as that of a traveller upon a new country, whose landscapes and inhabitants are of a different type to any which we ourselves have seen." (Galton, 1880, p.256)

This blind trust in introspective reports came under attack in the early 20th century. The argument was not that, as Galton seemed to suggest, participants might be dishonestly reporting things they had observed, but that the things they were reporting on were not accessible to them.

As we have seen, dual process theory suggests that all System 1 processes are preconscious. Any introspective reports that people relay about these processes are, by definition, post-event rationalisations lacking in validity.

Wason and Evans (1975) adopted a pseudo-clinical interview approach during a Selection Task study. They found that when participants were given a standard rule ($P \Rightarrow Q$) they justified their answer by explaining that they were trying to prove the sentence true. But when the same participants were given a matching bias version ($P \Rightarrow \neg Q$) they used the exact opposite explanation, that they were looking to prove the rule false. Wason and Evans concluded that the subjects were providing a System 2 post-event rationalisation for their choice rather than a genuine report of their (System 1) reasoning processes.[2]

It is clear, then, that not all mental processes are available to introspective descriptions. Proponents of the clinical interview method accept this:

[2]Although, of course, Wason and Evans didn't phrase their explanation using the modern dual process theory terminology.

> "Once it is admitted that the mind is not transparent, ... questions concerning the reliability and significance of introspective reports begin to loom large." (Swanson et al., 1981, p.31)

However, Swanson et al. claimed that as long as certain safeguards are taken, the issues affecting clinical interviewing are not significantly greater than those affecting any other methodology:

> "Put simply, our position is that: 1) Verbal data do have a place in cognitive research; 2) there are important limits and constraints on their use; 3) effective use of verbal data requires paying careful attention to these limits and constraints; 4) provided this is done, any of the remaining qualms about using verbal reflections are also those which apply to other sorts of data collected by more standard research methods." (Swanson et al., 1981, p.31-32)

Swanson et al. (1981) went on to explain that it is unlikely, for example, that a person has access to the source of creative insight; but that they may well be able to report how they tackled a mathematical problem. They give the example of doing a column addition "starting at the bottom". Thus, it is fair to say that a person may be able to report on *what* they did, but perhaps not always *why* or *how*. Furthermore, Smagorinsky (1989) argued that 'thinking aloud' whilst tackling a problem does not alter internal thought processes, as long as the verbalisations do not require the reporting of information that wouldn't normally be used whilst performing the task. Thus, with careful task design and interpretation, it seems that valid data can be gathered using a clinical interview approach. Similar arguments were made by Ericsson and Simon (1980).

5.3 The quasi-judicial method of analysis.

Most of the research on the Wason Selection Task described in Chapter 4 was empirical experimental work. This sort of work, with its emphasis on repeatability and hypothesis testing is sometimes referred to as being within the 'scientific paradigm'. The somewhat cliched view of this approach to research is that scientists form hypotheses, design experiments to test them, go out into the 'real world', collect quantitative measurements, and then return to the office to confirm or disconfirm the original hypothesis. There is some doubt as to whether this is actually how research is conducted, and there is even evidence that scientists are peculiarly bad at rejecting falsified hypotheses (some of this evidence, coincidentally, was collected by Wason, 1960).

Despite these doubts, there are several key criteria that distinguish the scientific method. Cuff and Payne (1979), for example, wrote:

> "A scientific approach necessarily involves standards and procedures for demonstrating the 'empirical warrant' of its findings, showing the match or fit between its statements and what is happening or has happened in the world" (p.4)

These standards and procedures typically involve the controlled manipulation of variables to see if the results match those predicted by various hypotheses. Data are then analysed with statistical tests to establish whether correlations and differences are either present, or if they are present, significant. The notion of repeatability is also important to the scientific method: experiments should be described in such detail as to permit other researchers to repeat the experiment and verify its results.

However, for certain research questions standard scientific methods may not be the most appropriate. In particular, the traditional experimental view of science runs into difficulties when it comes to qualitative data of the form that is collected by clinical task-based interviews. In this section an approach to analysing qualitative data that stresses close connections and parallels with the scientific method is discussed.

The quasi-judicial method of analysing qualitative data was developed by Bromley (1986) in the context of psychological case-studies. It is an attempt to recognise that whilst scientific experimental methods cannot be used to effectively study all real-life situations, they need not be disposed of entirely. Bromley's basic claim is that it is possible to investigate situations using a case-study methodology and retain both scientific rigour and reliability. He wrote:

> "In advocating a case-study approach to psychological problems we are not abandoning scientific method. [...] One can generalise from individual cases, and many important real-life human problems cannot be studied as effectively, or at all, by experimental methods. [...] Case-studies are not an inferior form of scientific method. On the contrary, they are possibly *the* basic method of science." (p.286, 289).

Bromley (1986) named his method of analysing case studies the 'quasi-judicial method'. For Bromley, a psychological case-study is simply an account of how and why a person behaved in a given, presumably interesting, situation. The account is adequate if it "contains enough empirical evidence, marshalled by a sufficiently cogent and comprehensive argument, to convince competent investigators that they understand something that previously puzzled them" (p.37). The quasi-judicial method is a systematic sequence of steps through which to describe and interpret empirical evidence.

The name 'quasi-judicial' comes from an analogy with practice in jurisprudence, although the analogy is with events that take place before a trial rather than those that happen during trial:

> "A quasi-judicial case-study, by contrast [to a well rehearsed argument of the form found in legal trials], is more an exercise in problem-solving. The aim is to understand scientifically what is going on and to manage the affair in a professional and businesslike manner. It is perhaps more akin to the French 'inquisitorial' system of judicial enquiry." (Bromley, 1986, p.30).

In short, Bromley recognised that the methods of analysing qualitative case-study data had often been inadequate in psychology and social science research, and prescribed a normative structure to follow, based on notions from the legal profession.

Bromley's (1986) normative structure for analysing a case, then, contains ten steps:

1. State initial problems and issues clearly.
2. Collect and state background context to the case.
3. Propose prima facie explanations.
4. Through examination of the prima facie explanations and solutions, search for additional evidence.
5. Search for sufficient evidence to eliminate as many of the suggested explanations and solutions as possible.
6. Closely examine the evidence, and sources of evidence, to check for consistency and accuracy (analogous to a cross-examination).
7. Conduct a critical inquiry into the internal coherence, logic and external validity of the arguments in the favoured explanations.
8. Adopt the 'most likely' explanation given these steps.
9. Formulate, if appropriate, what implications there are for action (this would be appropriate, for example, in the context of the evaluation of a course syllabus, or in the case of an educational psychologist attempting to resolve children's behavioural difficulties).
10. Prepare a coherent case-report. (Adapted from Bromley, 1986, p.26).

It is clear to see from these steps that, both philosophically and practically, the quasi-judicial approach is fundamentally at odds with other common methods of qualitative data analysis such as grounded theory (Glaser & Strauss, 1967), phenomenography (Marton, 1981), or the discipline of noticing (Mason, 2002). Rather than allowing theories to emerge from the data, the quasi-judicial method uses the data to test pre-existing theories in a manner akin to 'standard' scientific experimental methods.

Although Bromley (1986) was primarily concerned with case-study analyses, he introduces the notion of "case law" to deal with the weighing up of evidence from multiple case-studies:

> "Within a given field of inquiry there may be family resemblances between different cases. By comparing and contrasting cases, a kind of 'case-law' can be developed. Case-law provides rules, generalisations, and categories which gradually systematise the knowledge (facts and theories) gained from the intensive study of individual cases." (p.2).

Here, again, the analogy is with the legal profession. Case law, or common law, in the context of jurisprudence is a collection of prior legal judgements which can be cited as precedent in order to influence future legal decisions. Similarly, Bromley sees case law as a series of case-studies organised in such a way as to form theories useful for applications to future cases. Finding such case law is dependent upon "finding close structural similarities or identities between case-studies" (p.298).

Bromley (1986) recommended Toulmin's (1958) representational structure for informal argumentation as a method of analysing different case-studies to develop case law (see Bromley, 1986, chpt 9). By "shearing" an argument contained in a case-study "of its particular identity it can then represent a *class* of similar cases" (p.229). That is to say, that by representing arguments from different case-studies using Toulmin's scheme, structural similarities and differences between cases can be identified and explored, with the aim of constructing case law.

As with all qualitative methods of data analysis, Bromley's (1986) quasi-judicial method naturally raises questions of validity and reliability. Having analysed a case-study, and even having developed a system of coherent case law, how can a researcher using the quasi-judicial method generalise to other comparable cases? Bromley argues that the issue of generalisation from single case-studies (or case law based on several case-studies) should be seen in terms of the validity of the analysis, not in terms of the representativeness of the case. J. C. Mitchell (1983) put forward a similar argument, writing that "the validity

of extrapolation depends not on the typicality or representativeness of the case but upon the cogency of the theoretical reasoning" (p.207).

The quasi-judicial method should not be compared to statistical inference, instead it is a "strong form of hypothetico-deductive theorising". Bromley (1986) concluded:

> "We do not infer things 'from' a case-study, we impose a construction, a pattern of meaning, 'onto' the case. Ideally, the individual case puts our theories to the test." (p.290)

The quasi-judicial approach to qualitative data analysis is clearly at odds with other well known methods of data analysis. Rather than building theory from qualitative data in the manner of grounded-theorists or phenomenographers, the quasi-judicial researcher seeks to use qualitative case-studies to test various different theoretical frameworks: theory is applied *to* the qualitative data, not built *from* it. Theories are then accepted, refined or rejected, and case law is built up from comparing and contrasting several comparable case-studies. The reliability of the method comes from the validity of the manner in which the analyst applies the theory to the case, not from the representativeness of the case itself.

According to this view of qualitative data analysis, threats to the validity and reliability of a research project come from the deficiencies of the researcher: if theory is incorrectly applied or evidence is marginalised or overlooked than the quasi-judicial approach may not result in either valid or reliable conclusions. Sadler (1981) listed several possible deficiencies of human analysts, including a tendency to rely upon first impressions, a tendency towards overemphasising confirmatory data and downplaying disconfirmatory data, and a tendency to compare to a fictional base line. Some of these defects are related to findings from the heuristics and biases research programme discussed in Chapter 7. A conscious awareness of these possible deficiencies and a concerted effort to overcome them is necessary if the validity and reliability of a quasi-judicial study is not to be threatened.

5.4 Overview.

The empirical research reported in this thesis falls into two parts. Firstly, in Chapter 6, the various theories of reasoning discussed in Chapter 4 are critically evaluated by comparing the performance of mathematics students with the general well-educated population on the Wason Selection Task (Experiments 1 and 2) Using an inspection time eye-tracker based methodology, it is argued

that only the heuristic-analytic dual process theory of reasoning can successfully account for the behaviour of successful mathematicians on the Selection Task (Experiment 3).

Having adopted the heuristic-analytic dual process account, before being applied in the second stage of the empirical research, the theory is reviewed in greater detail in Chapter 7.

The second stage to the empirical research (Experiment 4) is reported in Chapter 8. It consists of a qualitative interview study which attempts to apply the dual process framework to the specific research question that this thesis set out to answer: how do mathematicians evaluate conditional statements? There are two parts to this study. Firstly, the role of preconscious heuristics in realistic mathematical contexts is examined; and secondly, the conscious processes involved in the evaluation of mathematical conditionals are discussed with reference to Toulmin's (1958) argumentation scheme. Bromley's (1986) quasi-judicial method is adopted throughout.

The layout of the empirical research reported in this thesis is summarised in Figure 5.1.

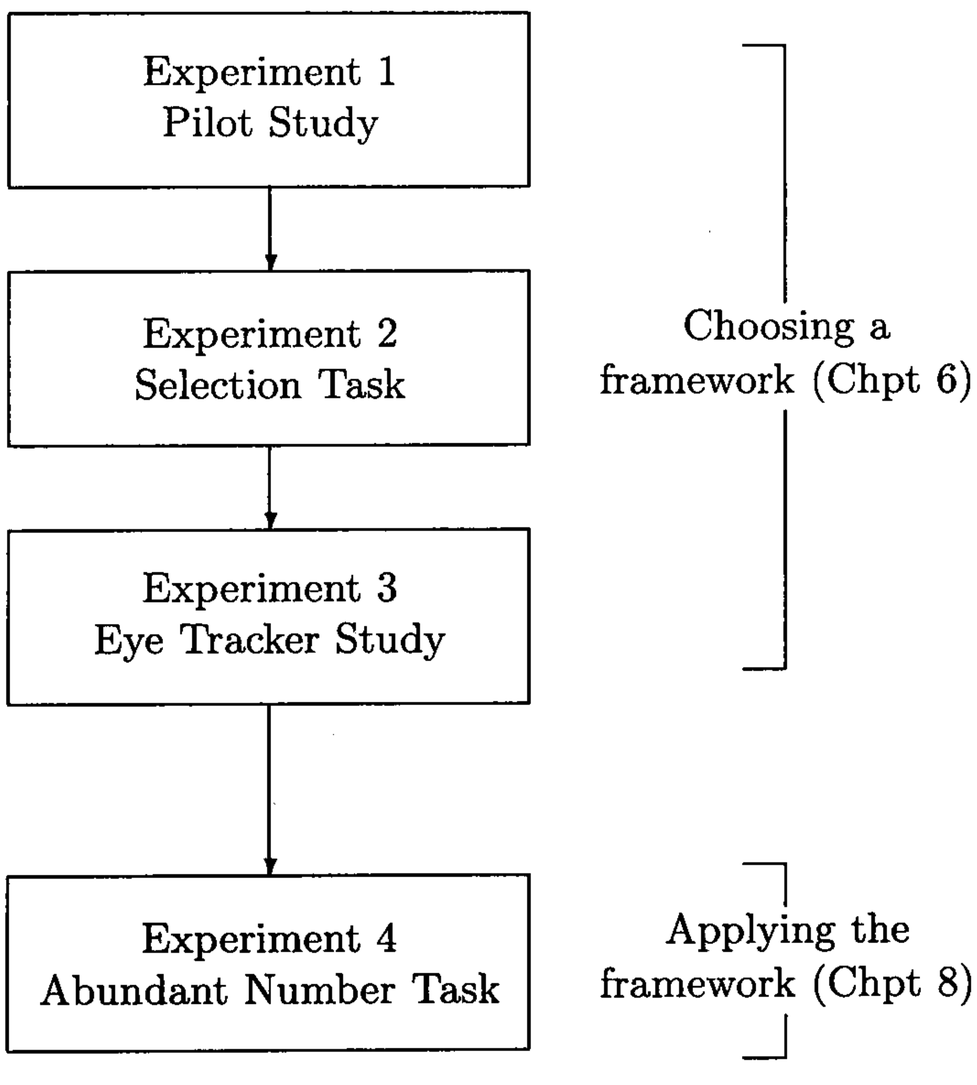


Figure 5.1: The organisation of the experimental section of the thesis.

Chapter 6

Adopting a framework: Mathematicians, Dual Processes and the Selection Task

This chapter reports the first half of the empirical work contained in the thesis. The primary goals of the experiments reported here are to distinguish between the various theories of reasoning discussed in Chapter 4: which of these frameworks is best placed to be applied to the study of mathematical reasoning? To this end, the experiments reported in this chapter consider mathematics students' responses to the Wason Selection Task. This task has historically been the most insightful instrument available to reasoning researchers, and so it seemed a natural choice to be used in the current context.

Experiment 1 reports the outcome of a straightforward experiment which investigated the differences (if any) in response between mathematics students and the general well educated population. The finding that there is indeed a difference is then followed-up in Experiments 2 and 3. The results of these experiments are used to argue that only Evans's (1996) heuristic-analytic dual process theory can successfully account for the responses of successful mathematics students to the Wason Selection Task.

6.1 Experiment 1: The pilot study.

"All mathematicians *can* solve the four cards problem

if they put their minds to it." (Devlin, 2001, p.120)

As noted in Chapter 4, the literature on the Wason Selection Task is vast. However, the key finding – that the general population struggle to select the normatively correct cards – has remained unchallenged. In view of the supposed importance of logic in mathematics discussed in Chapter 2, the question of how successful mathematicians perform on the Selection Task is of considerable interest. Is Devlin correct to claim that mathematicians *can* solve the task? And more importantly, *do* mathematicians solve it?

6.1.1 Method.

The aim of the pilot study was to speculatively administer the Selection Task to large mathematical and non-mathematical populations; and to compare the results. To this end, an internet based methodology was adopted.

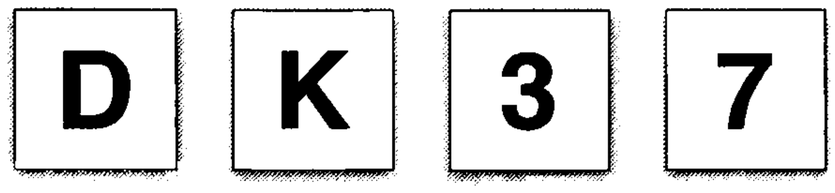
There were four categories of participants in the study, all from the University of Warwick; mathematics undergraduates, mathematics (academic) staff, chemistry undergraduates and history undergraduates. History undergraduates were selected to represent the general population. Clearly such a highly education population isn't very general; but, for practical purposes, the control group needed to be contained within the university, and so would inherently be unrepresentative of the population at large. This is a problem that affects virtually all psychological research; Wason (1968), for example, used psychology students as his population. History is a subject that contains little or no overt mathematical content, and so it seemed a good choice of department to act on behalf of the non-mathematical population. Chemistry undergraduates were selected as a 'half-way house'. It was assumed that they would have a relatively strong background in mathematics, but without the emphasis on proof and logic.[1] The mathematics sample was particularly highly qualified. A typical offer from the Warwick Mathematics Institute is an A and B in two mathematics A-Levels. The department's international research reputation attracts some of the best students in the country.

The undergraduate course at the University of Warwick contains a 30 hour first year module on the 'Foundations of Mathematics'. Approximately three/four hours of this module are devoted to introducing basic logical structures such as the formal definition of the conditional using truth tables. This is used to justify using proof by contradiction. Barring these few lectures, there is no explicit logic taught to undergraduate students (and neither is there any explicit logic on GCSE or A Level mathematics syllabi).

[1]Part of the first year chemistry core at Warwick is a 30 hour course in mathematical methods which covers material to A-level standard. A-level mathematics is not a prerequisite for studying Chemistry at Warwick.

Four cards are placed on a table in front of you. Each card has a letter on one side and a number on the other.

You can see:



Here is a rule: *"every card that has a D on one side has a 3 on the other."*

Your task is to select all those cards, but only those cards, which you would have to turn over in order to discover whether or not the rule has been violated.

Figure 6.1: The task used in Experiment 1.

The participants first received an email that explained the task and asked them to participate in the study. They were assured that the experiment would be entirely anonymous. If they agreed to participate, they would click on a link which directed them to the experimental website (Figure 6.1). Having submitted their answer, the participants were directed to a post-test thank-you page and were invited to request a digest of the results. Thirty-seven people asked to be notified of the results.

The phrasing of the task was taken from Wason (1969), Griggs and Cox (1982) and Jackson and Griggs (1988). All of these studies found that their changes of wording from Wason's (1968) original had no significant effect on the results.

Forty-eight hours after the original e-mails were sent the results were downloaded and imported into a spreadsheet for analysis. Those who had seen the task before were deleted from the results – very few people fell into this category. In total 562 people participated in the study (compared to 62 in Wason's (1968) original study). The breakdown of numbers by group is given in Table 6.1.

6.1.2 Web based experimenting.

Using the internet in psychology experiments is a relatively new phenomena and it offers both opportunities and challenges. Clearly, it would have been impractical to obtain the same level of response using a traditional survey method. This can easily be seen by the limited numbers that Wason (1968) and others

	Maths u/g	Maths Staff	Chem. u/g	Hist. u/g
n	260	21	67	123
%	34	24	27	23

Table 6.1: The breakdown of response numbers, n gives the raw figure and % gives n as a percentage of those who were sent the e-mail. Here, u/g refers to undergraduate students.

were able to recruit as participants – some studies reviewed in Chapter 4 had sample sizes of less than 20. However, having adopted a web based strategy, the amount of control that the experimenter has is more limited. For example, one cannot be sure that the participants didn't perform the task in consultation with others.

Perhaps the most worrying problem that web based experiments face is that of multiple submission. There is no foolproof way of preventing participants submitting their answer more than once. Several options are available. It would be possible to write a website which places a 'cookie' on the user's computer when they first load the page. If they tried to resubmit their results then the website would detect the presence of the cookie and the result wouldn't be recorded. Of course, if the subject was determined to resubmit, they could delete the cookie and take the task again again. The use of cookies in website design causes concern amongst some users, as they are a potential security risk. Due to this, most browsers can be set to refuse them.

Instead of using cookies, advice was taken from Reips (2000) who suggested logging the IP address of the subject. Again this isn't a foolproof method; often users have dynamic addresses – each time they go online they are assigned a different IP address. Reips also suggests that a combination of e-mail and IP addresses is a better solution, but this option was rejected in order to preserve the anonymity of replies.

By logging the time and the IP address of participants, it was possible to catch those who resubmitted in quick succession. There seemed to be only one case of this – one mathematics undergraduate submitted 34 times within 5 minutes. His or her answers were deleted. Of course, if someone with a dynamic IP address had gone to the trouble of logging out, logging back in and then resubmitting, there would have been no way of catching them. It is highly unlikely that this occurred enough to adversely affect the results. In the end the main defence against people resubmitting is simply that they have no incentive to do so. Unless incentives to take part are being offered, it provides them with no benefit. Indeed, the research that has been conducted in this area suggests that the number of internet resubmissions is very low indeed. One experiment

	Maths u/g	Maths Staff	Chem. u/g	Hist. u/g
$P, \neg Q$	29	43	8	10

Table 6.2: The percentage of each group selecting the correct answer in Experiment 1.

(in the days where dynamic IP addresses were rare) put the resubmission figure at 0.5% of total submissions (Reips, 2000, p.105).

The question of how valid web based psychological research is has been studied intensively. One study compared the results of twenty internet based surveys with their laboratory counterparts and found a remarkable degree of congruence between the two methodologies (Krantz & Dalal, 2000). It seems clear that the benefits of using the web for this pilot study substantially outweigh the disadvantages. The literature suggests that problems such as lack of experimental control and multiple submissions do not affect the validity of web experiments findings; furthermore the opportunity to survey larger sample sizes is a great help in ensuring both validity and reliability.

6.1.3 Results.

The main results of the pilot study – the percentage of each group correctly selecting cards D and 7 (P and $\neg Q$) – are shown in Table 6.2. In Table 6.2, and the tables that follow, only percentages are shown.

The first thing to highlight is that the maths undergraduates did indeed find the normatively correct answer significantly more often than the history undergraduates, $\chi^2 = 20.8, \mathrm{df} = 1, p < 0.001$, who performed in a similar manner to the Chemistry students. However, the mathematics students' range of answers is far from consistently normative. Less than a third of students, and less than half of staff identified the normatively correct answer. Interestingly a χ^2 test reveals that the responses of the mathematics staff were not significantly different to the students, $\chi^2 = 1.21, \mathrm{df} = 1, \mathrm{NS}$, although perhaps this can be put down to the small sample of staff. Interestingly, there was no significant difference between the responses of the chemistry and the history undergraduates ($\chi^2 = 9.22, \mathrm{df} = 6, \mathrm{NS}$), suggesting that increased exposure to non-proof based mathematics is insufficient to cause the different range of responses detected between the mathematical and history samples.

Looking at the detailed results (Table 6.3) reveals that not only are the mathematics undergraduates more successful at finding the normatively correct answer, but across the whole range of selections they perform differently, $\chi^2 = 95.9, \mathrm{df} = 8, p < 0.001$. In particular, they seem to make *different* mistakes.

	Maths u/g	Maths Staff	Chem. u/g	Hist. u/g
P	35	24	22	27
$P, \neg P$	0	0	0	0
P, Q	6	5	33	30
$P, \neg Q$*	29	43	8	10
$P, \neg P, Q$	0	5	2	0
$P, \neg P, \neg Q$	13	14	1	3
$P, Q, \neg Q$	3	10	7	4
all	8	0	19	13
non-P	5	0	9	12
n	260	21	67	123

Table 6.3: The percentage of each group selecting each answer in Experiment 1; *logically correct answer.

	Maths u/g	Maths Staff	Chem. u/g	Hist. u/g
P	50	42	24	30
$P, \neg P$	1	0	0	0
P, Q	8	8	36	33
$P, \neg Q$*	–	–	–	–
$P, \neg P, Q$	0	8	2	0
$P, \neg P, \neg Q$	18	25	1	3
$P, Q, \neg Q$	4	17	7	5
all	11	0	20	15
non-P	7	0	10	13

Table 6.4: The frequency that each group selected each answer as a percentage of total mistakes; *logically correct answer.

The modal non-normative answer from non-mathematicians was to select the D and the 3. A third of history undergraduates, and nearly as many chemistry undergraduates selected these two cards. Only 6% of maths students made this mistake, an extremely significant difference, $\chi^2 = 46.4, \mathrm{df} = 1, p < 0.001$.

The frequency of each mistake is highlighted when the figures from Table 6.3 are quoted as percentages of the *non-normative answers only.* Using the maths students as an example; 35% selected D only, but this selection amounted to 50% of all non-normative answers from the group. This is shown in Table 6.4.

As well as looking at the frequency of each selection of cards, the results can be analysed in terms of the numbers selecting each card collapsed across selections. This is shown in Table 6.5.

Despite the clearly higher success rate of mathematicians, it must be re-emphasised that one in five of the mathematics students in the sample affirmed the consequent (see Chapter 3). Furthermore, the academic staff were just

	Maths u/g	Maths Staff	Chem. u/g	Hist. u/g
D	95	100	91	88
K	25	19	26	25
3	20	19	62	52
7	57	67	40	37

Table 6.5: The percentage of each group selecting each card, collapsed across all combinations of selections.

as prone to this as the undergraduates. Of the students, only 57% correctly applied the modus tollens argument, and the academic staff, at 67%, did not have a significantly higher figure.

This then, is a very interesting and surprising result. It appears that mathematicians, both undergraduates and academic staff, apparently go about solving the Selection Task in a different way to the general population, but, despite this, their method doesn't lead to an overwhelmingly successful result. To summarise:

- Mathematicians are significantly more likely to find the normative answer to the Selection Task than non-mathematicians (both the history and the chemistry groups).

- Despite this, the mathematicians do not perform flawlessly. Less than a third of the mathematics students, and less than half of the mathematics staff made the normatively correct selection.

- Those mathematicians who fail to answer normatively seem to make different mistakes than the non-mathematicians. In particular the 'standard mistake' – that of selecting the D and 3 cards – was very rarely made by mathematicians.

- Instead of the 'standard mistake' mathematicians seem to be more likely to select either the D card only, or the D, K and 7 cards. This latter response might be attributed to a misunderstanding of the question.

It is important to re-emphasise that all the participants in Experiment 1 were very able mathematicians; and the staff involved are amongst the top thousand or so research mathematicians in the world. It is clearly untenable to suggest that the range of answers detected can be attributed to a lack of mathematical knowledge or ability. A more sophisticated explanation for these range of results is required.

6.2 Experiment 2: Mathematicians' performance on the Selection Task.

The results of Experiment 1 are a serious challenge to all the existing theories of reasoning described in Chapter 4. Several obvious questions can be raised: Why should the mathematicians respond differently to the general population? Why don't they make the same mistakes as the general population? And can any existing theories of reasoning be adapted to answer these questions? Each theory needs to be evaluated as to how it can be adapted to explain these new results.

The theories discussed in Chapter 4, then, need to be re-evaluated in light of the results of Experiment 1. However, before moving on to this discussion the results of additional research are first reported. Experiment 2 served two purposes. Firstly to attempt to repeat the surprising results for the mathematicians that were obtained in Experiment 1, and secondly to explore in further detail the background of the participants and how they went about solving the task.

6.2.1 Method

The methodology used in Experiment 2 was broadly the same as Experiment 1, except in this instance participants were directed to a preliminary page which collected data on their year group (as a measure of their mathematical *experience*) and the final classification from their previous year (as a measure of their mathematical *attainment*). If there was a significant relationship between the responses to the Selection Task and experience or attainment, it should be expected they will be detected by these measures.

The participants in the second experiment were mathematics undergraduates from two further top ranked UK universities.[2] A total of 408 people took part.

In the second experiment the time in seconds from when participants submitted their preliminary information and progressed to the question to the point when they submitted their final answer was also recorded. Recording the time that each participant took to complete the task was designed to help throw further light upon which of the competing accounts of reasoning described in Chapter 4 can best account for the results.

For practical reasons, in this design the recorded time *included* time spent reading the instructions. Clearly this introduces an inaccuracy into the timing. Slow readers might appear to take longer than fast readers, regardless of the

[2]The two mathematics departments who took part in the second experiment both recruit undergraduates with A levels of AAB or higher. One department achieved an RAE grade 5*, the other a grade 5.

actual thinking time. However, as there is no reason to believe there is a correlation between reading speed and Selection Task performance, this factor will not affect the centre of the results, merely the spread.

Whilst the timing could have been made more accurate by putting the instructions on a separate page and only starting the timer when the page with the cards had been loaded, this would have changed the nature of the task from a pure reasoning question to a reasoning and memory question. Paradoxically, the more precisely the time taken to complete the Selection Task is measured, the less faithful a version of the task can be presented. And, as many studies have demonstrated, small changes in the task can lead to large changes in results and the meanings which should be attributed to them.

Although other experimenters have used a version of the Selection Task where the instructions and cards are not visible simultaneously (e.g. Ball et al., 2003; Evans, 1996; Roberts, 1998b), they had comparatively small samples, and were more interested in the time spent inspecting each card rather than the overall time taken. Clearly small samples increase the seriousness of the inaccuracy introduced by this timing error. Given the large sample size expected from the internet method, it was felt that the introduction of this timing error was not as important as preserving the structure of the task.

So, in addition to the answer, the following additional pieces of information were collected about each participant:

- The date and time they took the task.
- The IP address they used.
- Their year group.
- Their classification from their previous year's exams (if applicable).
- The time they took to complete the task.
- Whether or not they claimed to have seen the task before.

Naturally those participants who had seen the task before were deleted from the analysis.

6.2.2 Results.

The overall results were distributed in a similar fashion to the mathematics undergraduates in Experiment 1, 24% selected the normatively correct answer, and only 11% selected the modal response of the history undergraduates from Experiment 1 (see the last column of Table 6.6). However, some surprising findings emerged from a more detailed analysis.

	Year 1	Year 2	Year 3/4	All
P	33	28	38	33
$P, \neg P$	1	1	1	1
P, Q	13	9	11	11
$P, \neg Q$*	28	23	21	24
$P, Q, \neg P$	0	1	1	1
$P, \neg P, \neg Q$	7	11	8	9
$P, Q, \neg Q$	2	5	4	3
$P, Q, \neg P, \neg Q$	8	14	6	9
non-P	9	8	10	9
n	151	115	142	408

Table 6.6: The % of each undergraduate mathematics year group selecting each answer (for the rule $P \Rightarrow Q$). *logically correct answer.

Lack of an experience effect

Table 6.6 shows the answers selected within each year group. Given the results of the Experiment 1, superficially one might expect that the more experience of mathematics one has, the more likely one is to respond correctly to the task – that is, there may be some *experience effect*. Perhaps surprisingly, this table indicates that there seems to be no such effect. That is to say, 28% of first year undergraduates solved the task correctly, compared with 23% of second years and 21% of third years. There is no significant relationship between solving the task correctly and the year of study, $\chi^2 = 1.97, \mathrm{df} = 2, \mathrm{NS}$.

Lack of an attainment effect

Again, one might suppose that there is a *attainment effect* - a link between ability in mathematics and performance on the task. Table 6.7 shows the answers selected by participants against the classification they received in their last exams. It should be noted that of the 408 participants, 183 answered 'not applicable'. Clearly, the majority of these would have been first year undergraduates, though it may also include those who preferred not to answer this question.

From the 225 participants who did answer the question, the data reveals that there appears to be no attainment effect. That is to say, there is no significant relationship between solving the task correctly and performance in examinations, $\chi^2 = 3.12, \mathrm{df} = 4, \mathrm{NS}$.

To re-emphasise, these measurements of mathematical experience and mathematical attainment are crude. However, if a suitably significant correlation existed between either experience and attainment and ability to solve the Se-

	First	2:1	2:2	Third
P	33	30	40	31
$P, \neg P$	2	0	1	0
P, Q	11	13	6	8
$P, \neg Q$*	24	18	23	39
$P, Q, \neg P$	0	1	0	0
$P, \neg P, \neg Q$	15	12	6	0
$P, Q, \neg Q$	5	5	1	15
$P, Q, \neg P, \neg Q$	8	7	11	0
non-P	3	12	11	8
n	66	76	70	13

Table 6.7: The % of each classification's answer (for the rule $P \Rightarrow Q$). Some participants, including all first years, did not provide their classification from last year. *logically correct answer.

	mean time (s)	Std. Dev.	n
P	66.9	34.0	133
P, Q	61.5	38.9	45
$P, \neg Q$*	91.7	54.0	97
All	77.9	45.5	402

Table 6.8: The mean time taken by participants who selected the three most common answers. 6 outliers were ignored. *logically correct answer.

lection Task correctly it would be expected that it would have been detected by this instrument.

The implications of the lack of an experience or attainment effect are discussed in greater detail in §6.6.

The timing effect

Table 6.8 gives the mean time taken by participants who selected the three most common answers. Six participants took over 305 seconds (more than 5 standard deviations from the mean) to answer, and these cases were deleted from the analysis. It was assumed that their attention had been diverted elsewhere during the task.

Recall that this is the time spent reading *and* thinking about the question. From the table it is clear to see that those who selected the correct answer $(P, \neg Q)$ spent longer than those who selected either P or P, Q.

These data were analysed using a one-way ANOVA between the various groups given in Table 6.8. There was a significant difference between these groups, $F(3) = 8.983, p < 0.001)$. A Scheffe post-hoc comparison test indicated

that the time differences between $P, \neg Q$ and P; and $P, \neg Q$ and P, Q – despite the apparently high standard deviations – are highly significant (both at the $p < 0.01$ level). There was, however, no significant difference between the time spent by those who selected P and P, Q.

So, participants who selected the correct answer took significantly longer than those who selected the next two most frequent answers, P and P, Q. But the time difference between these two selections was not significant.

6.3 Discussion of Experiments 1 and 2.

These results, on the face of it are surprising. To summarise:

- The results from Experiment 2 replicated the mathematics undergraduates' results from Experiment 1.
- 24% of mathematics undergraduates selected the normatively correct answer $(P, \neg Q)$, 33% selected P. Very few (11%) selected P, Q – by far the most common mistake in the general population.
- No relationship between year of study or classification and answer to the task was detected.
- Those who answered correctly took significantly longer than those who answered either P, Q or P. There was no time difference between these two selections.

The question now is: how to interpret these results? Of the theoretical frameworks mentioned in section 4.4, which can best be applied to analysing the reasoning of the mathematicians? In the following sections each framework, and how they can be applied to these results, are considered in turn.

6.3.1 Mental models theory.

Recall that the mental models theory of reasoning (§4.4.1) suggests that humans reason by creating a mental model of the situation, but that often the model is incomplete as a result of the so-called 'principle of truth'; the idea that it is generally more efficient (in terms of cognitive effort and working memory load) to construct models only of things we know to be true. The correct answer is only reached if participants successfully 'flesh out' their model. Thus we need to consider three potential models:

- $[P] \quad Q$
  $\ldots$

- $[P]$ $[Q]$
  $\ldots$
- $[P]$ Q
  $\neg Q$

The first is a standard initial model, and if not fleshed out, results in the participant considering P and Q but picking only P. The second model is the result of the participant confusing the conditional with the biconditional and results in them considering P and Q *and* picking them. The third model is has been fleshed out from the first, and results in the participant picking the correct answer: P and $\neg Q$.

Adopting a mental models framework, then, the results from Experiments 1 and 2 would seem to imply that the undergraduate and professional mathematicians are:

1. Much less likely than the general population to construct a biconditional initial model, that is to say that mathematicians are less likely to interpret "if P then Q" as $P \Leftrightarrow Q$.

2. More likely than the general population to flesh out their initial model, and thus find the correct answer.

The first of these claims seems plausible. However, as we saw in Chapter 2, there is evidence from the mathematics education literature which suggests that some students have considerable difficulty distinguishing between '$P \Rightarrow Q$' and '$P \Leftrightarrow Q$' (Hazzan & Leron, 1996; O'Brien, 1973). There are several points to note here. Firstly, Hazzan and Leron's participants were computer science majors, not mathematics majors. We cannot claim that at that institution they had a lower mathematical ability based on this fact,[3] however it may be reasonable to argue that they are not as socialised into mathematical culture as the participants in Experiments 1 and 2. Certainly it seems reasonable to suggest that the two groups have different departmental affiliations with all that that entails (Bingolbali & Monaghan, 2004). Also, of course, the mistake that Hazzan and Leron report was only made by a minority of their sample. If the mental models explanation were adopted to explain this data, a small minority of participants *would* confuse the conditional with the biconditional. Unfortunately O'Brien does not report the departmental affiliation of his participants, so we cannot assume that they were mathematics majors: the same objection regarding departmental affiliation could apply.

[3]Indeed Leron reports that computer science majors enter the course with *higher* mathematics marks than the mathematics majors (private communication).

Suppose, then, that we were to accept the first of these claims. What could explain this? Given that there is no relationship between either mathematical experience or mathematical attainment, it would seem that the factor we are looking for must manifest itself before successful students reach university. Possible explanations are discussed in §6.6.

The second implication of adopting a mental models framework is less easy to explain. Why should mathematicians be more likely to flesh out their initial model than the general population? The mental models theory suggests that fleshing out an initial model takes time, effort and puts a load on working memory (Johnson-Laird & Byrne, 1991; Johnson-Laird, 2001). The results of Experiments 1 and 2 would, therefore, seem to suggest that mathematics undergraduates are significantly more willing to spend time and effort fleshing out their initial mental model. The timing data provide strong support to this hypotheses. Only those people who answered $P, \neg Q$ – those who explicitly fleshed out their initial model – took a significantly longer time. The two other main groups P and P, Q – those who did not flesh out their model – took significantly less time. This is perhaps not surprising. Mathematics undergraduates are, on the whole, people who enjoy logical problems, it is not unreasonable to assume that they would consider the Selection Task to be a logical problem, and that they would therefore be willing to devote more cognitive effort to solving it than a non-mathematician.

So, with the assumption that mathematics undergraduates very rarely interpret conditionals as biconditionals, the mental models framework successfully provides a way of interpreting the results of Experiments 1 and 2.

6.3.2 Mental logic theory.

As with the mental models account, the mental logic theory (§4.4.2) of Rips (1994) explains the results of the Selection Task in three categories:

- Those who answer P are applying modus ponens, a standard part of everybody's reasoning armoury.
- Those who answer P and Q are also applying modus ponens, but have interpreted the conditional rule as a biconditional.
- Those who answer correctly, P and $\neg Q$, have successfully constructed the complicated modus tollens argument with a contradiction proof (as discussed in §4.4.2).

The results of Experiments 1 and 2, then, would seem to suggest two fairly similar conclusions to the mental models account. Undergraduate and professional mathematicians are:

1. Much less likely than the general population to interpret the conditional statement in the rule ("if P then Q") as a biconditional ($P \Leftrightarrow Q$).

2. More likely than the general population to be able to successfully construct a contradiction proof of the modus tollens deduction.

The first of these claims is identical to that discussed in the mental models account.

The second is subtly different, however. It is entirely plausible, and indeed probable, that successful mathematicians would be more fluent at constructing contradiction proofs than the general population. In fact, it is perhaps more of a surprise that this fluency appears to high enough in *only* around 25% of mathematics undergraduates. Given that proof, and proof by contradiction, are commonplace in advanced mathematics it would seem surprising that such limited numbers are successful in this fashion.

Furthermore, it seems astonishing to claim that only 43% of professional mathematicians are able to successfully construct a straightforward modus tollens argument. It is also surprising that this ability does not seem to be related to either mathematical experience or mathematical ability. This issue can be seen as a weakness of the mental logic account of reasoning. It is clear that all professional mathematicians are capable of constructing a modus tollens contradiction type argument, if the mental logic theory explanation was adopted it would need to explain why they *don't* in this situation.

Having said this, it is clear that Rips' account can be adapted to explain the results of Experiments 1 and 2, even if the explanation is somewhat counter-intuitive.

6.3.3 Information value theory.

Oaksford and Chater's (1994) information value theory is different to most other theories of Selection Task performance. It offers a justification of the most common response, rather than an explanation of how the response is reached. Using a mathematical model of the situation, Oaksford and Chater deduce that selecting P and Q is actually the most 'rational' response to the task, since it is these two cards (with various sensible assumptions) that provide the highest information gain.

The results from Experiments 1 and 2 showed mathematicians very rarely making this selection. Instead they overwhelmingly selected P on its own, or the correct answer P and $\neg Q$. It seems very difficult to use information value theory to account for these results.

Perhaps of more value is to see information value theory as a part of a dual process framework. That is to say, to use Oaksford and Chater's (1994) analysis

as a plausible explanation as to *why* System 1 biases participants in the way that it is; but to also accept that there is a System 2 analytical stage to reasoning as well. The dual process theory explanation for the results of Experiments 1 and 2 is discussed in §6.3.5. A move in this direction was made by Oaksford and Chater (2003) who admitted that "an elite band of very high IQ participants" may have an analytical stage in their work on the Selection Task. However, they suggested, contrary to Evans (1996), that only this very small subsection of the population used analytic processes.

6.3.4 Relevance theory.

The relevance theoretic (Sperber, Cara, & Girotto, 1995) explanation of the Selection Task (§4.4.6) attributes correct answers to two different sources. A minority of participants "know the difference between demonstrative and non-demonstrative truth evaluation" (p.46) and therefore successfully solve the task. The majority, however, attempt to make the rule relevant through successively more complex interpretations.

In order to make the relevance theoretic account fit the data from Experiments 1 and 2, we would need to accept the following:

1. More of the mathematical group than the general group fall into Sperber, Cara, and Girotto's minority who successfully solve the task; however this group is still a small minority.

2. Very few mathematicians interpret "if P then Q" as '$\exists\, x$ such that $P(x) \wedge Q(x)$' (since this is the interpretation that leads to the P and Q cards being selected).

The second of these claims is difficult to accept. The interpretation that leads to the P and Q selection is, according to Sperber, Cara, and Girotto (1995), a more complex interpretation than the interpretation that leads to the P selection. It seems peculiar to suggest that, on the one hand, a larger percentage of mathematicians were sophisticated enough to remove relevance based calculations from their decision making process; but, on the other hand, that the rest of the mathematicians gave relevance to the rule by interpreting it in the least complex way. The relevance theoretic interpretation almost seems to be that a minority of the mathematical sample was substantially more sophisticated than the general population, but the rest of them were less sophisticated. This conclusion is surely not tenable.

6.3.5 Dual process theory.

Recall that dual process theory (e.g. Evans, 1996; Stanovich, 2004), discussed in §4.4.7, suggests that the standard mistake originates from preconscious heuristics in System 1, and is a result of System 2 failing to adequately monitor and override its partner system. If a dual process framework is adopted, there would appear to be two alternative hypotheses that explain the increased frequency of the P selection, and decreased frequency of the P, Q selection, by mathematicians in Experiments 1 and 2.

Hypothesis 1. Exposure to mathematics on a daily basis modifies System 1 heuristics so as to reduce the chances of the standard mistake – that of selecting P and Q – being made; but also to increase the chances of selecting the P card alone.

Hypothesis 2. Mathematicians' System 1 heuristics tend to operate in the same way as in the general population. But exposure to mathematics on a daily basis results in an increased tendency to use System 2 for monitoring and possibly modifying the output of System 1.

The psychology literature that discusses dual process theory would seem to indicate that hypothesis 2 is more likely, since the migration of rules and heuristics from System 2 to System 1 is not well understood (Evans, 2004a; Stanovich, 2004).

Furthermore, there would appear to be some support for the second hypothesis from the mathematics education literature. Jacques Hadamard, in his celebrated essay on mathematical thinking, emphasised the role that checking for errors plays in mathematics:

> "Good mathematicians, when they make [errors], which is not infrequent, soon perceive and correct them. As for me (and mine is the case of many mathematicians), I make many more of them than my students do; only I always correct them so no trace of them remains in the final result." (Hadamard, 1945, p.49)

Could this expertise in error checking be a consequence of a continual monitoring process by System 2?[4] If hypothesis 2 were the case, however, it would raise the question: Why don't the mathematicians *always* get the answer right? Every successful mathematician is surely aware of the contrapositive of $P \Rightarrow Q$, if System 2 is being used by the majority of the sample, why doesn't the majority of the sample select the normatively correct answer?

[4]Note that Hadamard (1945) is referring to all error checking, not just of conscious System 2 checking of preconscious System 1 processes.

The answer to this question is also suggested by dual process theory. Recall that the theory proposes that attention is preconsciously directed to the matching cards by System 1. It is *only* these cards that System 2 attends to. Thus the reason why even highly proficient mathematicians do not always notice that the $\neg Q$ card is needed may be because *they do not even consider it,* their attention has been preconsciously biased towards the matching cards.

However, it is clear that many mathematics undergraduates *do* consider the $\neg Q$ card, as around one in four of them select the correct answer. In a dual process framework this explanation can be accounted for by an increased tendency amongst the mathematicians to use System 2 to slowly and analytically consider all the cards, not only the ones to which their attention has been filtered.

This explanation would fit well with the timing data. Those who selected P or P, Q took roughly the same amount of time. According to a dual process framework, they were inspecting the cards that System 1 preconsciously directed System 2's attention towards, and either rationalising this selection (in the case of the P, Q selection) or correcting the mistake contained within it (in the case of the P selection). In contrast, however, those who selected $P, \neg Q$ took significantly longer, which may indicate they were slowly and analytically analysing the entire situation.

The second hypothesis associated with the dual process explanation, then, rests on three assumptions:

1. Mathematicians have similar preconscious System 1 biases as the general population.

2. Whilst rationalising their selection, mathematicians are much more likely to detect the Q error.

3. A large subset of mathematicians spend time and energy analysing the entire task using System 2.

The first of these assumptions fits with the psychological literature better than Hypothesis 1, but it is important to note that, as yet, there is no empirical data from this study to support it (but see Experiment 3). The second assumption seems very reasonable. As discussed earlier, it is entirely probable that mathematicians are less likely to confuse conditionals with biconditionals. This may well account for this effect. The third assumption is also reasonable. Once you have noticed that your System 1 biases are inaccurate, as most of the mathematicians do, it seems sensible to believe that you are much more likely to slowly analyse the entire situation using System 2. Perhaps, then, the fact

that the rationalisation process finds an error in the original bias, causes a subset of mathematicians to examine the whole problem again in a slow analytical fashion, and therefore select the $\neg Q$ card. This explanation for the results of Experiments 1 and 2 is summarised in Figure 6.2.

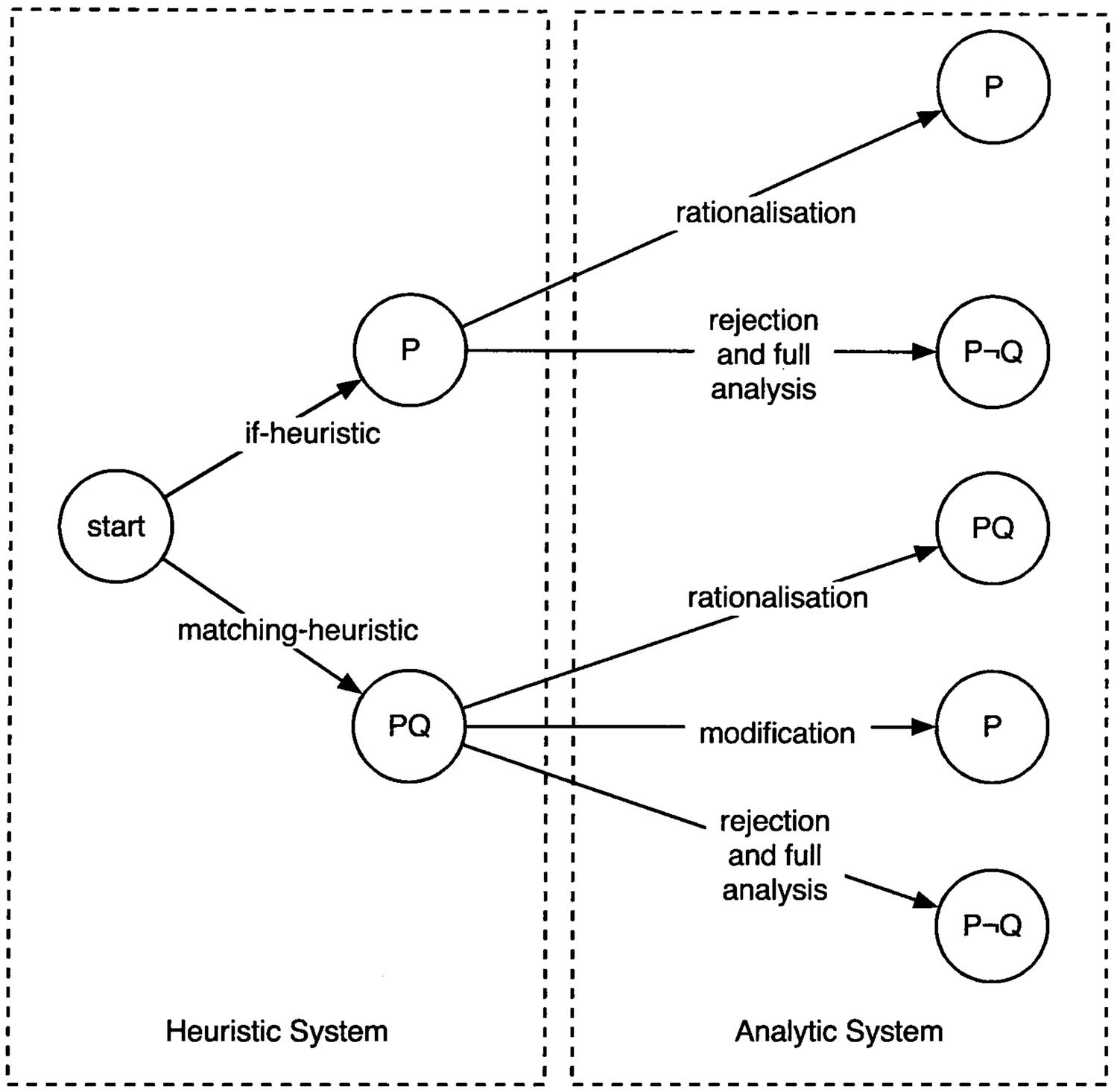


Figure 6.2: The dual process interpretation of Experiments 1 and 2.

Note that the explanation shown in Figure 6.2 is not intended to be interpreted as a universal account. Clearly, as with all psychological processes there will be a large degree of individual differences between participants. Rather, this explanation should be taken to be an account of the modal response, in an attempt to explain the trends in the data from Experiments 1 and 2.

6.4 Summary of Experiments 1 and 2.

Of all the theoretical frameworks that have been proposed to explain the Selection Task results, section 6.3 argued that there are three that can be successfully

adapted to explain the results of Experiments 1 and 2:

- Mental models theory (§4.4.1, §6.3.1).
- Mental logic theory (§4.4.2, §6.3.2).
- Heuristic-analytic dual process theory (§4.4.7, §6.3.5).

Before embarking on a detailed study on mathematical conditionals, it was decided to try to investigate further which of these three theoretical frameworks provide the best model for investigating how people reason with conditional statements.

6.5 Experiment 3: The eye tracker study.

6.5.1 Inspection time studies.

The strongest empirical support for Evans's heuristic-analytic dual process theory account of the Selection Task has come from inspection time studies. These were briefly discussed in section 4.4.7, but will be elaborated upon in this section.

Evans (1996) provided strong evidence for the dual process account of the Selection Task by measuring how long each participant spent inspecting each card. Participants were presented with Selection Tasks on a computer screen and were asked to hover their mouse pointer over the card that they were 'thinking about'. Participants selected each card by clicking on it with the mouse. As a consequence of this methodology, Evans argued that he had managed to measure how long each participant thought about each card.

Evans (1996) suggested that if System 1 heuristics cued the cards that System 2 considered slowly and analytically before selecting, there should be higher inspection times for the cards that ended up being selected. Thus, argued Evans, the dual process account of the Selection Task yields two predictions:

P1 Cards associated with higher selection rates will also be associated with longer inspection times.

P2 Within a given card, those subjects who choose it will have longer inspection times than those who do not. (Evans, 1996, p.226).

Evans's (1996) data supported these two predictions. There were large differences in inspection time between selected and non-selected cards: typically, selected cards were inspected for in the region of 4 seconds and non-selected cards for under 2 seconds. Evans wrote:

> "Our findings suggest that we may not think at all about options that we fail to choose." (p.236).
>
> "It appears that many [subjects] decide first and think afterwards" (p.238).

Roberts (1998b) later pointed out that these two hypotheses can be combined, and analysed with statistical techniques that have greater power than Evans' methods:

P3 For each individual, the mean inspection time should be longer for selected than for nonselected cards (Roberts, 1998b).

Roberts repeated Evans's (1996) experiment and found similar results.

However, Roberts argued that the methodological technique used by Evans was unsound. He pointed out that Evans's version required participants to undertake two activities simultaneously: solving the Selection Task, and using the mouse to show what they are thinking about. There was a great danger, suggested Roberts, that there could be sensory leakage; that participants could consider and reject cards before they managed to hover the mouse over them.

Furthermore, since participants were asked to click on cards they wanted to select, it might be the case that they would inevitably have to spend longer 'inspecting' the cards they wished to select as an artificial consequence of the task set-up. Thus, Roberts (1998b) argued, the inspection time effect could be an artefact of the way in which Evans (1996) administered his version of the Selection Task.

In order to investigate this possibility, Roberts (1998b) administered several variations of Evans's (1996) experiment:

- In the first modification, cards were only visible when the mouse hovered over them. This was an attempt to eliminate sensory leakage. Although the inspection time effect was still present, the size of the effect was considerably reduced on this version.

- In the second modification, participants had to either select "yes" or "no" for each card. This was an attempt to reduce the possibility that the inspection time effect was a consequence of spending time whilst physically selecting the cards. This modification again considerably reduced the inspection time effect.

- The third modification was a combination of the first and the second modifications. It completely eliminated the inspection time effect.

- In the last modification, the task was changed so that all the cards were initially selected and participants had to *deselect* cards appropriately (as

opposed to selecting them). In this version the inspection time effect was reversed.

As a consequence of these experiments, Roberts (1998b) argued that it was highly probable that the inspection time effect found by Evans (1996) was artificial.

Evans (1998a) replied to these criticisms by partially accepting criticisms of his methodology, but rejecting the idea that Roberts (1998b) had falsified the entire heuristic-analytic account. Evans pointed out that all of Roberts' methodological variations had the effect of forcing participants to attend to cards that they would not normally attend to. For example, in Roberts' first modification, cards were not visible unless the participant hovered their mouse over them. This clearly has the effect of forcing participants to attend to all cards in a fashion that they might not otherwise have done. Evans' objections were reanalysed by Roberts and Newton (2001) who conducted further experiments designed to remove any possibility of an artificial inspection time effect. Roberts and Newton used a rapid response methodology which insisted that participants responded to each card (presented individually) within two seconds of its display. It was argued that this would increase the frequency of responses attributed to preconscious biases. Strong supporting evidence was found for the dual process account.

Although claiming that Roberts (1998b) had not damaged the theory, Evans (1998a) did accept that factors other than relevance assessments from System 1 can affect what features of the task people attend to. He wrote:

> "For this reason, I would now wish to place less emphasis on inspection times than on other forms of evidence that support my proposals about relevance effects in reasoning" (Evans, 1998a, p.814).

It is also clear that a mouse hovering technique is unlikely to be sensitive enough to give a highly accurate measure of attention. Roberts (1998a), in a response to Evans' reply, suggested that "gaze-tracking studies" would resolve the issue conclusively.

Such a study was conducted by Ball et al. (2003). They used an eye-tracking device to record details of where the participants were looking during the time they tackled the task. Note that this experimental design assumes a strong positive correlation between eyeball fixation location and cognitive attention. This assumption is highly reasonable, and is discussed at length in §6.5.2. Ball et al. gave each participant four different Selection Tasks, with rotated negatives, in a random order. This approach, whilst commonplace, does run the risk of introducing learning effects into the data.

Ball et al. (2003) found convincing support for P1, P2 and the stronger P3. In view of the methodological criticisms of Roberts (1998b), two different experiments were conducted, with each version having the participants indicating their card selections in different ways:

- In the first experiment participants briefly pointed at the cards they wished to turn over using a 20cm metal pointer.
- In the second experiment participants pressed a button to indicate that they were ready to make their selection. Upon the button press, eye movement data ceased to be recorded.

Both of these variations found strong support for P3. A similar finding was reported by Ball, Lucas, and Phillips (2005) who conducted an eye-movement experiment using deontic thematic versions of the Selection Task.

Ball et al. (2003) argued that their two experiments rendered redundant any methodological concerns raised by Roberts (1998b). Despite the resolutions of these issues, there are several problems with the interpreting the inspection time paradigm as providing convincing evidence for the heuristic-analytic dual process account when compared to the mental models or mental logic account of Selection Task behaviour.

Firstly, the mental models theory also makes a prediction about which cards participants will inspect. Johnson-Laird and Byrne (1991, p.79) suggest that participants will "consider only those cards that are explicitly represented in their models of the rule". Thus, the mental models prediction, labelled P5 by Ball et al. (2003) is:

P5 The cards represented in mental models associated with the rule should be inspected for longer than cards not represented in the models. (Ball et al., 2003).

Recall that on the original "if P then Q" version of the task the initial mental model proposed by Johnson-Laird and Byrne (1991) is one of:

- $[P] \quad Q$
  $\ldots$
- $[P] \quad [Q]$
  $\ldots$

In both cases the theory predicts that participants should consider the P and Q cards. However, on task rules with rotated negatives the situation changes. For example, Johnson-Laird and Byrne (1991) suggest that the rule "if P then $\neg Q$" would have an initial mental model of:

$$
\begin{array}{ccc}
 & [P] & \neg Q \\
\bullet & & Q \\
 & \ldots &
\end{array}
$$

Thus for this version, the mental models theory predicts that the participant will consider the P, Q and $\neg Q$ cards, whereas the heuristic-analytic account suggests that the $\neg Q$ card will not be heuristically cued.

As Ball et al. (2003) used a rotated negatives paradigm, they were able to test P5 and found strong support for it. They pointed out that this result was not hugely surprising as there is a "lack of orthogonality" between P3 and P5. Thus the inspection time paradigm, if restricted to comparing the time between selected and non-selected cards, appears not to be able to distinguish between the mental models account and the dual process theory account. Despite this, Ball et al. (2003) suggested that their work provided strong evidence for the dual process theory account, since the inspection time predictions were bold. That is to say, the theory made a novel prediction that could easily have been falsified.

The second problem with the inspection time paradigm as it stands is that it could be argued that, contrary to the beliefs of Ball et al. (2003), P3 is not in fact a particularly novel prediction at all. Is it really a surprise that participants spend longer looking at cards that they select than those they do not select? A sceptic could argue that *any* of the theories of reasoning can account for this result, as long as it were assumed that participants spend a short amount of time checking or confirming their selections, *regardless* of how the selections were made. After all, the difference in inspection times between selected and non-selected cards detected by Ball et al. was not that great (typically the difference was around 1 second).

As a result of these two concerns, it does not seem reasonable to believe that comparing the inspection times of selected and non-selected cards provides sufficient evidence to determine which of the theories of reasoning discussed is the most useful when considering conditional reasoning.

However, the results from Experiments 1 and 2 provide an additional testable hypothesis. As discussed earlier, the dual process account of these results (hypothesis 2) relies upon several points:

1. Mathematicians have similar preconscious System 1 biases as the general population.

2. Whilst rationalising their selection, mathematicians are much more likely to detect the Q error.

3. A large subset of mathematicians spend time and energy analysing the

entire task using System 2.

From this analysis we can derive P6:

P6 Mathematicians will spend longer inspecting non-selected matching cards than non-selected mismatching cards. This will not be the case for the general population.

To begin with, in view of the concerns regarding the inspection time paradigm discussed above, only non-selected cards are being considered, thus the problem of artificial 'checking' time does not apply to this prediction. The claim being expressed by P6 is that mathematicians will reject matching cards in a different way to mismatching cards, and that the general population will not do this.

If mathematicians are being cued towards matching cards in a similar manner to the general population they will, presumably, spend a considerably longer amount of time rejecting these cards than the ones that they have not been biased towards. However, most of the general population who are cued towards the matching cards end up selecting them after a rationalisation process. Those in the general population who do not select matching cards have, for the most part, not been biased towards them.

In short, the unusual results from Experiments 1 and 2 suggest P6, a novel prediction with which to test the dual process theory account of the Selection Task.

6.5.2 The eye-mind assumption.

The validity of the eye tracking methodology used by Ball et al. (2003) and Ball et al. (2005) relies upon there being a strong correlation between eye movements and cognitive attention. This claim is reasonable, but to examine it we need to introduce some terminology. There are, roughly speaking, two sorts of eye activities: fixations and saccades. Fixations occur when the eye is resting on a particular point in the field of view, and saccades are the movements that eyes make between fixations.

Typically people look at something when they wish to acquire information from it, and therefore it is natural to want to deduce cognitive attention and cognitive processing duration from eye movement data. This, the so-called *eye-mind assumption*, was suggested by Just and Carpenter (1980) in the context of reading:

> "the eye remains fixated on a word as long as the word is being processed. So the time it takes to process a newly fixated word is directly indicated by the gaze direction" (p.330).

However, in a recent review of the situation Irwin (2004) pointed out four problems with assuming a 1-1 correspondence between attention and fixation location:

- The view that is afforded by a single fixation is quite large. It is unreasonable to assume that *only* the fixation location is being processed during a fixation.
- The locus of cognitive processing moves from the current fixation location to the next one *before* the eyes themselves shift.
- Sometimes the eyes are captured in an involuntary fashion by unexpected stimuli. Thus the position of the eyes is not always under cognitive control.
- Some cognitive processing takes place during saccades.

Despite these concerns, Irwin argued that fixation duration and fixation location measures can be very useful indices of cognitive processing, but that the 100% correlation suggested by Just and Carpenter (1980) was not justified. In another review of the field, Rayner (1998) agreed with this:

> "Although we can easily decouple the locus of attention and eye location in simple discrimination tasks, [...] in complex information processing tasks such as reading, the link between the two is probably quite tight" (p.375).

Most researchers seem to agree that it is generally more efficient to move the eyes than it is to move attention alone (e.g. He & Kowler, 1992; Liversedge, Paterson, & Pickering, 1998; Sclingensiepen, Campbell, Legge, & Walker, 1986).

The assumption underlying the work of Ball et al. (2003), Ball et al. (2005) and Experiment 3 – that fixation durations and fixation locations correlate sufficiently with cognitive processing to provide a useful measure of Selection Task card inspection times – seems reasonable, and is supported by the currently available eye-movement research. Furthermore, fixation locations have been used to measure attention during problem solving across a wide variety of domains in addition to the Selection Task (e.g. Charness, Reingold, Pomplun, & Stampe, 2001; Knoblich, Ohlsson, & Raney, 2001; Pan et al., 2004; Rayner, 1998).

6.5.3 Design and method.

In order to investigate P6 it was decided to adopt an eye-tracker methodology along the lines used by Ball et al. (2003).

Participants.

The mathematics sample consisted of 30 undergraduate and postgraduate volunteers from the Mathematics and Statistics Departments at the University of Warwick. The control sample consisted of 28 undergraduate or postgraduate volunteers from the Arts Faculty at the same institution. None of the participants had seen the task before or received any tuition on the psychology of reasoning. All participants were paid volunteers, and were recruited on the basis that they did not wear eye-glasses.

Apparatus.

Stimuli were presented on a 19" (445mm [33.05°] visible diagonal) Sony SVGA monitor at a resolution of 800×600 pixels (120 Hz refresh rate) which was driven by a 1 GHz Pentium based PC. Eye position was determined using an SR Research EyeLink I system. This is a head mounted infrared-based system that automatically compensates for head movements and achieves an average gaze accuracy of approximately $0.5 - 1.0°$. Eye position was recorded from the right eye at a rate of 250 Hz. The eye tracker was calibrated for each participant using a 9-point display directly before the presentation of the task. No mechanical means were used to restrict head movements and viewing distance was approximately 75cm. Participants were tested individually in a dimly illuminated sound attenuated room.

Materials.

A standard Selection Task was used, displayed in two parts. As discussed above, splitting the instructions and the cards onto two different screens is unfortunate, but unavoidable when the inspection time for each card is the variable being measured. If, for example, the instructions and cards were on the same screen it would be difficult to know how to interpret time spent looking at the P and Q in the rule. Should this time be counted as part of P and Q's inspection times or not? There is no satisfactory resolution to this question, other than splitting the instructions and cards onto different screens.

The task instructions used are shown in Figure 6.3, and the layout of the cards used is shown in Figure 6.4. Contrary to Evans (1996), Roberts (1998b) and Ball et al. (2003), only one version of the task (a standard "if P then Q" version) was given to each participant. This was designed to eliminate any learning effect that might adversely affect the results. There were $4! = 24$ different configurations for the cards, and they were fully counterbalanced across participants.

On the next screen you will see four cards. Each of the cards has a letter on one side and a number on the other. You will only be able to see one side. Here is a conjectured rule about the cards:

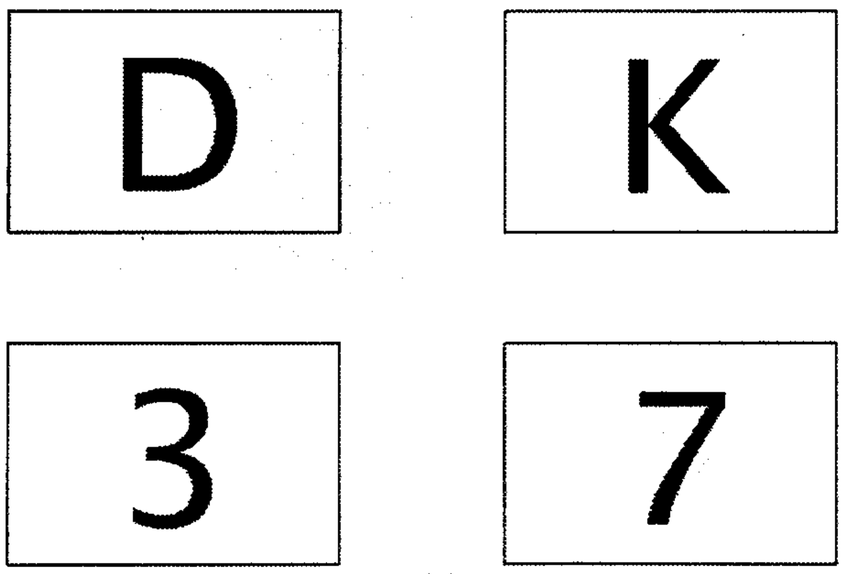
if a card has a D on one side, then
it has a 3 on the other

Your task is to select all those cards, but only those cards, which you would need to turn over in order to find out whether the rule is true or false.

When you are sure you understand these instructions press the button.

Figure 6.3: The instructions seen by participants in Experiment 3.

You can see:



when you are happy with your answer, press the button

Figure 6.4: One version of the cards seen by participants in Experiment 3.

Procedure.

Participants were tested individually. Once they had been fitted with the eye-tracker and given a response box to hold, a series of calibration and drift correction tasks were undertaken. After successful calibration, a screen of instructions was displayed (see Figure 6.3).

When participants had pressed a response button they were asked to confirm that they understood the instructions. Upon confirmation of this the experimenter spoke aloud:

> Thank you. You will now see the next part of the task. When you are ready to answer the question, please press one of the buttons and say your answer out loud. Take as long as you want.

They then fixated (for drift correction purposes) on a dot at the centre of the screen, afterwards the cards were displayed.

Once the participants has pressed the button the screen blanked, the eye-tracker stopped recording and they were asked to state their selection aloud which was recorded by the experimenter. This method of stopping eye-tracker reasoning experiments was identical to that used by Ball et al. (2003, Experiment 2) and Charness et al. (2001). Once the experiment had been completed the data was imported into the EyeLink Data Viewer software for analysis.

Analysis.

To analyse the eye tracking data the on-screen display was divided into areas of interest. Two different analyses were conducted, one where fixations outside the cards were ignored, and one where fixations inside a region that included each card and an area surrounding it (giving a total area of 39245 pixels) were included. The dwell times for the two types of region were highly correlated ($rs > 0.98$). In all the analyses reported below the larger areas have been used.

It is possible that the amount of time required for a participant to press a response button after deciding on their answer (which blanked the display and halted the eye movement recording) could artificially inflate the inspection time of the last card to be fixated. In order to prevent this possibility we analysed and report the data with the last two fixations deleted (although an additional analysis in which all data were included produced an essentially identical pattern of data and levels of significance).

6.5.4 Results and discussion.

Inspection times, P3 and P6.

The distributions of selections made by each group is shown in Table 6.9. Despite the comparatively small sample size, the difference between the two groups' *range* of responses was significant, Fisher-Freeman-Halton Exact Test, $p < 0.05$, with the direction of the key differences between the groups as in Experiment 1a. In addition, the mathematical group, as in Experiment 1a, made more 7 selections than the arts group (43% vs. 14%) and fewer 3 selections (36% vs. 43%), although this latter difference was of a lower magnitude than that found in Experiment 1a.

	Maths u/g		Arts u/g	
	raw	%	raw	%
D	12	40	10	36
D3	3	10	10	36
D7*	4	13	0	0
DK37	4	13	0	0
other	7	23	8	29
N	30		28	

Table 6.9: The percentage of each group selecting each answer in Experiment 3. No single selection in the 'other' category was made by more than 2 participants. *logically correct answer.

Recall that Roberts (1998b) translated Evans' (1996) predictions P1 and P2 into P3:

P3 For each individual, the mean inspection time should be longer for selected than for nonselected cards (Roberts, 1998b).

In order to test P3, the mean dwell time for each card that each participant selected was calculated, along with the mean dwell time for each card that they did not select. These data are shown in Table 6.10. Evidence for P3 was found in both groups. The overall mean inspection time, across both groups, for selected cards was 3.9s, compared to a figure of 2.6s for non-selected cards.

A between-groups within-subjects analysis of variance (ANOVA) provided strong support for P3, $F(1, 52) = 9.66, p < 0.01$. However there was no significant interaction between group and card-type (selected/non-selected) dwell-time, $F(1, 52) = 0.99, \text{NS}$.

However, the primary purpose of this experiment was to investigate a stronger prediction that P3, namely P6:

P6 Mathematicians will spend longer inspecting non-selected matching cards

than non-selected mismatching cards. This will not be the case for the general population.

In order to test P6, mean dwell times were calculated for participants' non-selected matching cards and non-selected mismatching cards.[5] These data are given in Table 6.11, and have been graphed in Figure 6.5. Both groups had similar inspection time means for non-selected mismatching cards (2.2s for the mathematical group and 2.3s for the arts group). However there was a difference between the groups on the non-selected matching card mean inspection times. The mathematical group's figure was 4.4s, substantially higher than the arts group's figure of 1.7s.

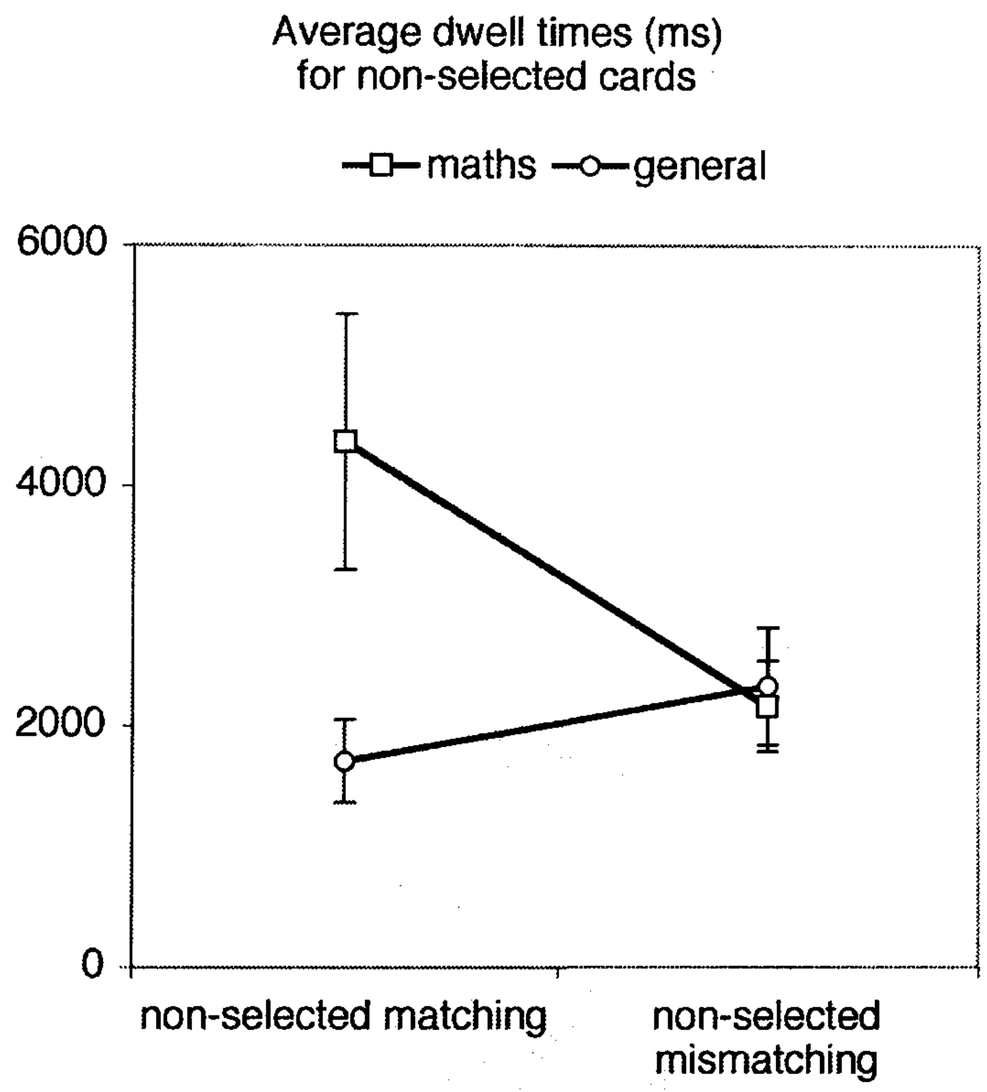


Figure 6.5: The mean dwell times (ms) for non-selected matching, and non-selected mismatching cards. Error bars represent $+/-1$ standard error around the mean.

A univariate ANOVA was conducted to compare dwell-times on non-selected matching and non-selected mismatching cards. The group×card-type interaction effect was significant, $F(1,86) = 4.51, p < 0.05$, with the mathematical group spending significantly longer inspecting the non-selected matching cards than the arts group, $F(1,39) = 4.53, p < 0.05$.

[5]Note that the four participants who selected all the cards have been deleted from this analysis.

Card type	Group	N	Mean	Std Error
Selected cards	Maths	26	4716.2	1009.4
	Arts	28	3091.1	377.8
	Total	47	3873.6	530.6
Non-selected cards	Maths	26	3069.7	636.6
	Arts	28	2243.1	505.8
	Total	47	2641.1	381.3

Table 6.10: The mean dwell times (ms) for selected and non-selected cards.

Card type	Group	N	Mean	Std Error
Non-selected matching cards	Maths	23	4363.2	1066.3
	Arts	18	1709.0	342.7
Non-selected mismatching cards	Maths	23	2165.2	379.1
	Arts	26	2331.7	487.8

Table 6.11: The mean dwell times (ms) for non-selected matching, and non-selected mismatching cards.

Card type	Group	N	Mean	Std Error
Non-selected 3s	Maths	19	4237.8	1246.0
	Arts	16	1558.5	329.9
Non-selected 7s	Maths	17	1668.7	405.4
	Arts	24	2226.3	572.1
Non-selected Ks	Maths	22	2030.0	363.9
	Arts	24	2341.8	542.5

Table 6.12: The mean dwell times (ms) for non-selected cards.

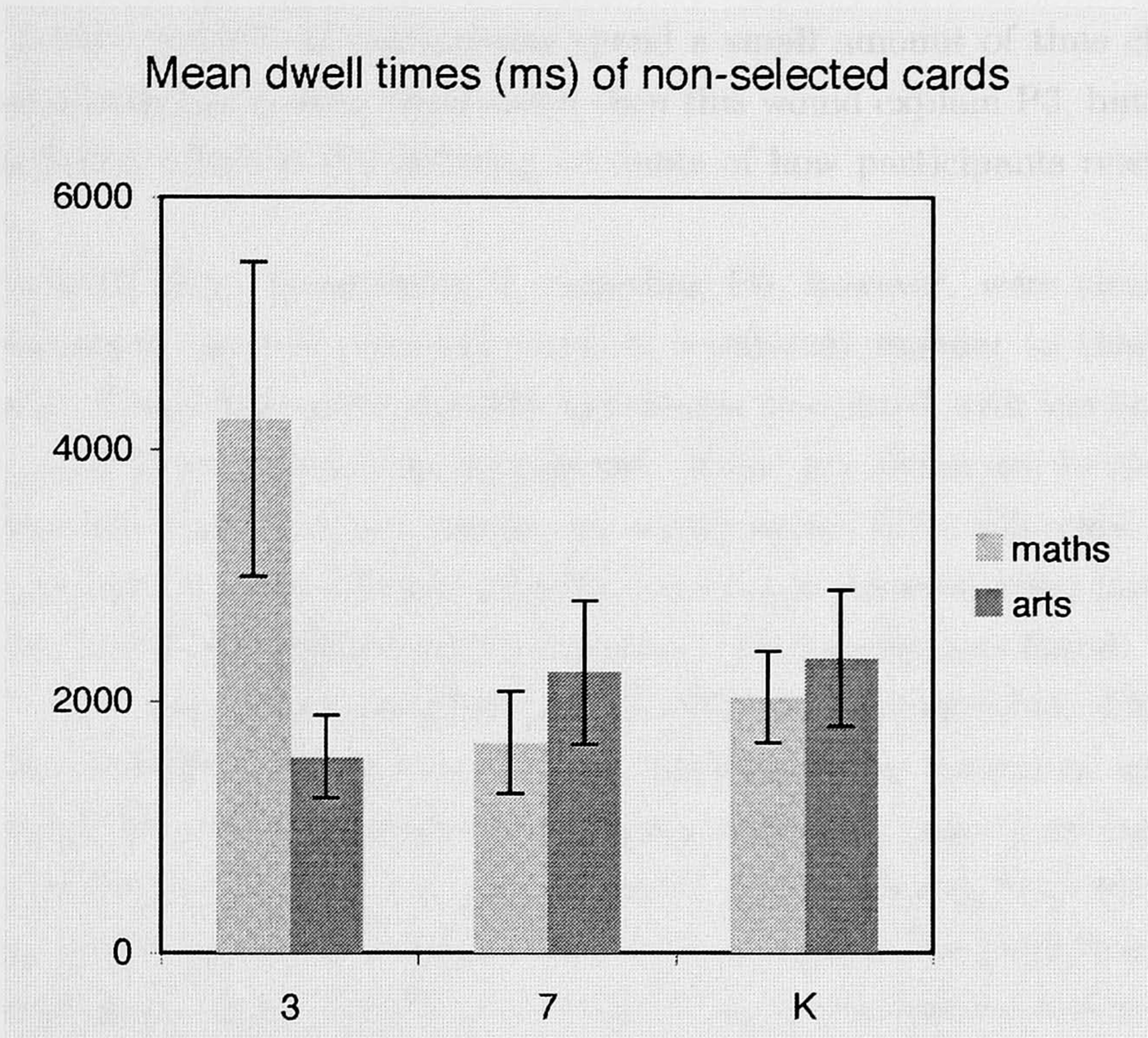


Figure 6.6: The mean dwell times (ms) for the non-selected 3, 7 and K cards. Error bars represent +/ − 1 standard error around the mean.

In addition, an analysis on individual nonselected card's mean dwell times was conducted, excluding the D card on account of the high proportion of participants who selected this card. These data are shown in Table 6.12, and graphed in Figure 6.6. Both groups had similar inspection times for rejected K and rejected 7 cards (ranging from 1.7s to 2.3s), but the mathematics group's mean inspection time for the rejected 3 card was 4.2s compared to 1.6s for the arts group. A univariate ANOVA was conducted to compare mean dwell times on these non-selected cards, the group×card interaction effect was significant, $F(2, 116) = 3.46, p < 0.05$.

These data suggest that the mathematical group rejected matching cards in a significantly different manner to the arts group; and furthermore that the mathematical sample rejected matching cards in a significantly different manner to how they rejected non-matching cards.

These strongly support the second hypothesis associated with the heuristic-analytic dual process theory account. The data replicates Ball et al.'s (2003) work with regards to P3; it was found that both groups spent longer inspecting cards they selected than those they did not select. However, the confidence Ball et al. have in this prediction to reject certain theories of Selection Task perfor-

mance is not justified. If participants spend a small amount of time checking their answers before stating them aloud then this would explain P3, but it does not distinguish between the differing accounts of how participants reach their answers.

The results from Experiment 3, regarding P6, however, were clear. The mathematicians rejected matching cards in a different manner to that of the arts group. These data allow the first hypothesis associated with the heuristic-analytic dual process theory to be rejected. If the key difference between the two groups was their System 1 biases, one would expect little difference in their inspection times between the non-selected matching and non-selected mismatching cards. In fact a significant group×card-type interaction was found.

Note that these data also shed light on the issue of individual differences discussed in §6.3.5. The relatively large standard error associated with the mathematicians' mean non-selected matching inspection time is an indication that not all mathematicians are preconsciously biased towards both the P and Q cards: this is merely the modal case. A large number, as predicted by the heuristic-analytic theory, will be more effected by the if-heuristic and be biased towards the P card alone. This factor accounts for the difference in standard errors shown in Figure 6.5.

Mean fixation duration measures.

The Eyelink system records a substantial amount of data which can be explored. Alongside the predictions P3 and P6 some of the other data recorded by the equipment in Experiment 3 was also investigated. In particular, some researchers have suggested a correlation between the mean fixation duration on an area and the cognitive processing difficulty associated with that area (for a full discussion see Irwin, 2004; Rayner, 1998). Although it should be noted that this correlation is not well understood, and may well only be justified when considering *visual* cognitive processing difficulties. Nevertheless, several researchers have used mean fixation durations as direct measure of task difficulty and information complexity (e.g. Ikehara & Crosby, 2005; Pan et al., 2004).

The Eyelink equipment records fixation durations in ms for each fixation. The mean fixation duration for each card, f_X, was baselined according to the formula

$$f_X = \frac{100 \cdot d_X}{f_S \cdot n_X}$$

where d_X is the dwell time for card X, f_S is the mean fixation duration for the entire screen and n_X is the number of fixations on card X. Thus for each card the mean fixation duration as a percentage of the mean fixation duration for the entire screen was calculated. This baselining was designed to control for

Card type	Group	N	Mean	Std Error
Matching cards	Maths	29	1.051	0.0382
	Arts	28	1.075	0.0227
	Total	57	1.062	0.0270
Mismatching cards	Maths	28	0.962	0.0557
	Arts	27	0.989	0.0252
	Total	55	0.975	0.0307

Table 6.13: The mean baselined fixation durations for matching and mismatching cards.

any general speed variations between participants.

The mean fixation durations for various types of cards were then calculated. Table 6.13 shows the mean fixation duration times for matching versus mismatching cards.

A univariate ANOVA showed that the mean baselined fixation durations for matching cards were significantly higher than those for mismatching cards, $F(1, 108) = 5.29, p < 0.05$, but that there was no significant group×card-type interaction, $F(1, 108) = 0.002$, NS.

So, the participants' mean fixation times were significantly longer for matching cards than they were for mismatching cards. The mean fixation duration analyses provide some support to the heuristic-analytic dual process account of Selection Task performance. Participants had significantly longer mean fixation durations on matching cards than they did on mismatching cards. Mean fixation durations have been used by some as a measure of cognitive processing difficulty. Evidence from a study by Ballard, Hayhoe, Pook, and Rao (1997) suggests that whereas short fixations are primarily concerned with encoding, or reencoding, information into working memory, long fixations are associated with deeper processing. As a consequence of such work several researchers have used mean fixation durations to give an indication of task difficulty and information complexity (e.g. Knoblich et al., 2001; Pan et al., 2004). Given this, these data could be interpreted as indicating that the manner in which participants processed matching cards is different to how they processed mismatching cards. This might suggest that the purpose of fixations on mismatching cards was to reencode the identity of the card into memory, whereas some of the fixations on matching cards were correlated with more advanced processing. Such an account fits with the heuristic-analytic dual process theory.

Ordering data.

In order to investigate the role of heuristic biases in this task further, an analysis of the orders in which participants looked at different cards was conducted. Each

participant's trial was split into fiftieths, and each fiftieth-segment was coded based on which card was fixated on for the longest amount during this fiftieth. If the participant was not fixating on any card during the segment it was coded as "missing". Unfortunately this process needed to be done by hand using printouts and acetate, as the EyeLink Data Viewer does not have the facilities to conduct such an analysis.

After each participant's trial had been coded, each fiftieth was given a score based on whether the card was matching (1) or mismatching (−1). Thus a string of fifty values was associated with each participant. These values were averaged for the two groups and a five point moving average taken. The graph of this data is shown in Figure 6.7.

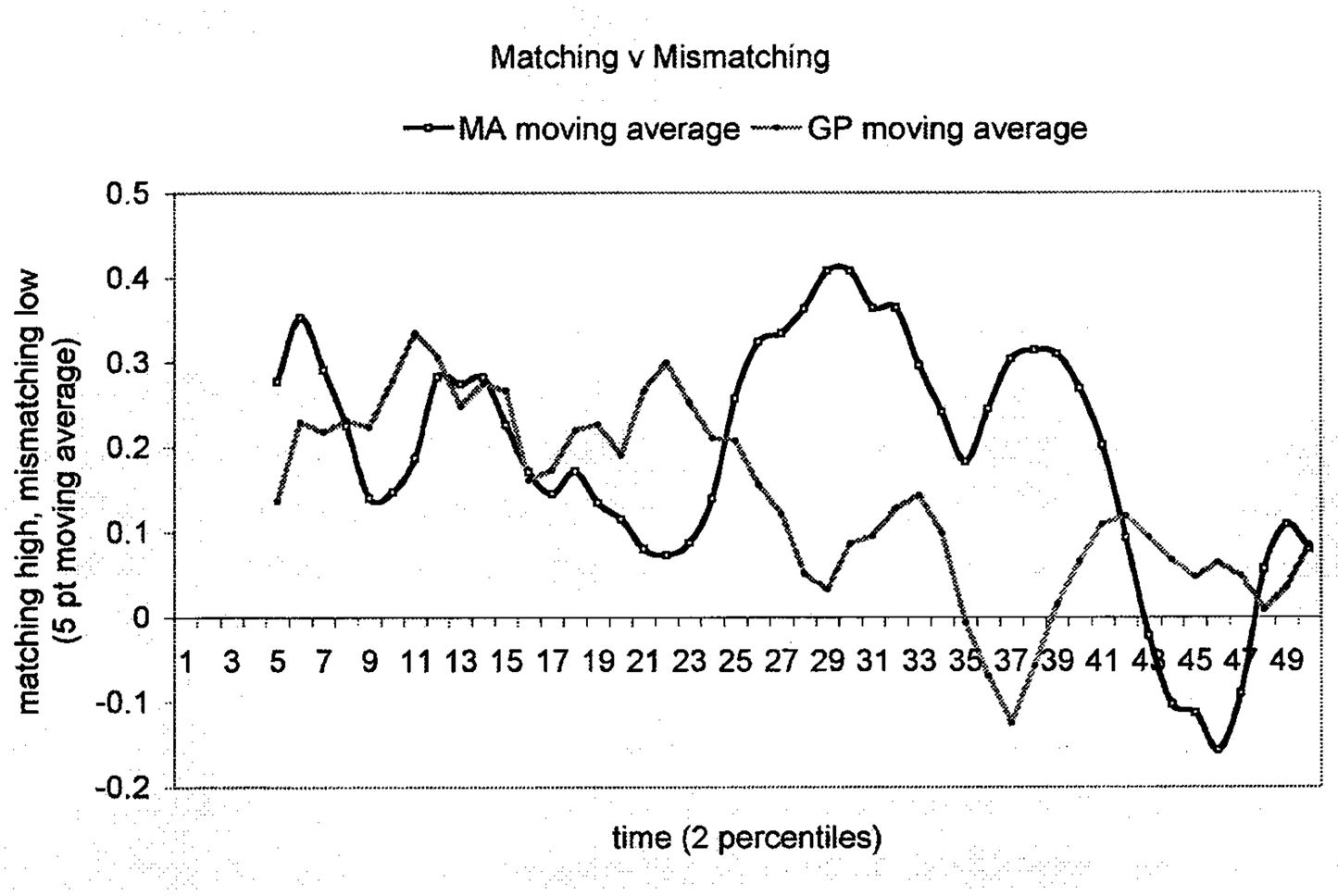


Figure 6.7: The order in which participants looked at matching and mismatching cards.

It can be seen that both the mathematical and the control group were, initially, biased towards looking at the matching cards. The control group's line moves in a generally negative direction, indicating a slow reassignment towards looking at the mismatching cards as well as the matching cards. However, the mathematical group's line initially starts this process before steeply returning to the matching cards, and entering into another cycle.

Extreme care needs to be taken with these data, as the standard errors involved are large. However, the data is consistent with the heuristic-analytic account outlined above. The mathematical group's sudden lurch up the graph about half way through the time period could be interpreted as being the point

when they notice that the 3 card is not necessary and begin a slow analysis of the matching cards. However, it is important to emphasise that this interpretation of these data is speculative. Notwithstanding this caveat, it is fair to say that the ordering data provides weak support to the heuristic-analytic account of Selection Task performance.

6.5.5 General discussion.

The main aim of Experiment 3 was to attempt to distinguish between the various theoretical accounts of Selection Task performance from Experiments 1 and 2.

Using eye-movement data the behaviour of mathematics students and arts students whilst solving the task was compared. Although the main purpose of this experiment was to investigate the results from Experiments 1 and 2 further, as a byproduct, Ball et al.'s (2003) findings were successfully replicated: It was found that participants' inspection times associated with selected cards are longer than those for non-selected cards.

However this result cannot, as Ball et al. (2003) suggested, be used to successfully distinguish between the various accounts of Selection Task performance. Instead P6 was derived and tested, a prediction based on the heuristic-analytic dual process theory. P6 suggested that the mathematical group would have higher inspection times for non-selected matching cards than for non-selected mismatching cards, and that this would not be the case for the general population. Strong support was found for this prediction.

The results of Experiment 3 raise doubts about both the mental logic and mental models accounts of the Selection Task. Both these theories can explain the differences in the range of responses from mathematicians and the general population in the same way: by suggesting that fewer mathematics undergraduates interpret conditional ($P \Rightarrow Q$) as a biconditional ($P \Leftrightarrow Q$) than the general population. This is a plausible explanation for the different range of responses, but not the difference in inspection times.

Mental logic theories would suggest that because fewer mathematicians misinterpret the rule in this fashion they make the P selection more frequently than the non-mathematicians. However, the theory cannot account for why the inspection times for mathematicians and non-mathematicians should be different for the non-selected matching cards.

The mental models account implies a difference between the two group's distribution of initial models. That is to say that there would be proportionately fewer mathematics undergraduates who represent the problem with the biconditional model:

$$[P] \quad [Q]$$
$$\ldots$$

and proportionately more who represent the problem with the conditional model:

$$[P] \quad Q$$
$$\ldots$$

Crucially, according to the mental models theory, the proportions of each group who have each initial model has no bearing on how each individual participant reasons from this model. The theory argues that participants who select the P card start with this model and reason from here, so no difference in behaviour would be expected between the mathematicians who start with this model and the non-mathematicians who start with this model. That is to say that all those non-mathematicians who represent the problem with a conditional model should inspect exactly the same cards (P and Q), and for similar amounts of time, as the mathematicians who represent the problem with a conditional model. But a significant interaction between group and card-type was found, showing that the mathematicians spent considerably longer inspecting non-selected matching cards than they spent inspecting non-selected mismatching cards. This result runs contrary to the predictions of the mental models theory. As the theory currently stands, it cannot successfully account for these data.

As well as calling into question the mental models and mental logic theories, these data cast doubt upon Oaksford and Chater's (1994) information value model as an explanation of the Selection Task as a whole. Whereas the results do not impact on Oaksford and Chater's (1994) mathematical model, they do provide strong evidence for an analytical stage of reasoning, as suggested by Evans (1996). It is preferable to see Oaksford and Chater's (1994) analysis as an explanation for why the matching- and if-heuristics have survived the evolutionary process, despite, from a logical standpoint, being suboptimal (see §4.4.4).

These data from Experiment 3 do not comfortably fit any of the mental models, mental logic or information value accounts of Selection Task performance. But Evans's (1996) heuristic-analytic dual process theory can account for them. The mathematics undergraduates in the sample appeared to be affected by the same preconscious biases as the general population, but during the analytical stage of reasoning they seemed to behave differently. This lead them to being considerably more successful at detecting that the Q card (in the rule $P \Rightarrow Q$) is unnecessary. As discussed above, the findings of Stanovich and West (1998) do not seem to explain this difference. It is clearly untenable to suggest that mathematics students have, de facto, higher cognitive abilities than arts students.

Although Stanovich and West's (1998) work cannot explain the difference in the System 2 behaviour of the mathematical and non-mathematical samples, it is less clear what can. The range of responses detected from the mathematics sample appears to be different to that found in any other homogeneous grouping, including research scientists (Griggs & Randell, 1986; Kern et al., 1983). Notwithstanding several largely introspective reports of mathematicians on the psychological aspect of their work (e.g. Hadamard, 1945; Poincaré, 1905; Tall, 1980), there has been little empirical research on the nature of higher-level mathematical cognition with reference to conditional statements. Further research on this subject is reported in Chapter 8.

Evans's (1996) heuristic-analytic dual process theory account of choices on the abstract Selection Task suggests that preconscious System 1 heuristics direct attention to apparently relevant parts of the task. For most participants, conscious System 2 processes only serve to rationalise these biases. This chapter has argued that the reason for behavioural differences between mathematicians and the general population on the task is that a large proportion of mathematics undergraduates do more than just rationalise these biases. They actively analyse and override them. This adaptation of the heuristic-analytic dual process theory can account for the surprising difference in the range of responses to the task given by mathematicians and the general population.

The evidence here suggests, in line with the heuristic-analytic dual process theory account, that there are two stages to reasoning on the Selection Task. Evans's (1996) theory, however, makes no strong claims about *how* these stages operate. So while the theory suggests that there is an analytical stage to reasoning where System 1 biases are either modified or rationalised, there is no clear framework with which to describe and explain how this analytical phase operates. It may well be that, modulo System 1 heuristics, mental logic theories or the mental models theory could form the basis of such a theory. Furthermore, it could well be that the information value theory can account for the existence and evolutionary survival of the matching- and if-heuristics that operate within System 1.

6.6 Aside: The effect of an undergraduate education on reasoning skills.

Although the main purpose of this chapter has been to distinguish between the various theories of conditional reasoning that could be used to analyse mathematical reasoning behaviour, the results from Experiment 2 also reveal some surprising findings with regard to the effect of an undergraduate education on

reasoning. Recall that the data from Experiment 2 showed no correlation between finding the normatively correct answer to the Selection Task and either year of study (experience) or degree classification (attainment). These data are shown in Tables 6.6 and 6.7 (p.82).

In the literature on conditional reasoning there have been few studies that speak directly to the issue of what effect an undergraduate education has on reasoning skills. Lehman and Nisbett (1990) conducted a longitudinal study that looked at this question. They constructed a battery of tests and tasks to produce an instrument that scored participants on 'statistical and methodological', 'conditional' and 'verbal' reasoning abilities. Participants were tested twice: once at the beginning of their undergraduate course, and once in their fourth year of study. Students from four different disciplines took part: natural science, humanities, social science and psychology. Of particular note for the current purposes is that Lehman and Nisbett's 'conditional reasoning' instrument included a version of the abstract Selection Task, alongside three other tasks.

Lehman and Nisbett (1990) found that, over the course of their degree, students from natural science and humanities backgrounds improved their conditional reasoning scores by substantial percentages ($> 55\%$), whereas students from social science and psychology backgrounds made no significant improvements. The data also revealed a significant correlation between the number of mathematics courses taken by the students and their conditional reasoning score improvement, $r = 0.31, p < .0.002$. Restricting this analysis to only the natural science students (who took most of the mathematics courses) increased the correlation to $r = 0.66, p < 0.001$.

At first glance the results of Experiment 2 appear to contradict the conclusions of Lehman and Nisbett's (1990) study. The data from Experiment 2 showed *no* correlation between experience of an undergraduate mathematics education and responses to the abstract Selection Task. However, some care is needed in interpreting these results in this fashion. Lehman and Nisbett used four tasks to compile their participants' 'conditional reasoning' scores. Their paper has no details on how each component of their instrument contributed to this score, or what the non-Selection Task questions were. Unfortunately the raw data regarding participants' responses to the Selection Task component of the instrument are no longer available, so no reanalysis is possible (Lehman and Nisbett, private communication). Given this, it is impossible to tell whether the improvements detected by Lehman and Nisbett on their overall conditional reasoning instrument are representative of an increased tendency to respond to the Selection Task normatively. However, there are further reasons to be cautious of direct comparisons between the studies. Lehman and Nisbett found

improvements in natural science and humanities students. No mathematics undergraduates took part in the experiment. Perhaps the first year students in Experiment 2 had already been exposed to as much mathematics as the fourth year natural science students in Lehman and Nisbett's work. Without access to the original materials and raw data it is hard to speculate further.

Putting Lehman and Nisbett's (1990) work to one side, it is still perhaps surprising that there appears to be no correlation between increased mathematical experience or attainment and responding normatively to the Selection Task. However, once the results of Experiment 3 have also been digested, the lack of correlations are not so unexpected. The heuristic-analytic dual process account of the data from Experiments 1 and 2 posits that in fact the vast majority of mathematical participants in Experiment 2 responded normatively, *modulo their preconscious biases.* The parts of the problem that the mathematics students considered *were* analysed normatively by almost all participants.

The following is a plausible account of where each of the answers detected by Experiments 1 and 2 come from:[6]

- P, Q – Very small numbers of mathematics students made this selection.
- P – This answer could result from those mathematics students who were biased towards the P and Q cards, and normatively analysed this part of the problem and correctly detected that the Q card was unnecessary. It is probable that an awareness of the distinction between '$P \Rightarrow Q$' and '$P \Leftrightarrow Q$' could be responsible for this.
- $P, \neg Q$ – This answer could result from those mathematics students who were biased towards P and Q (or P alone), but were sufficiently engaged with the task to begin a slow analysis of the *entire* problem. That is to say that they ignored their preconscious biases and conducted a protracted analysis of the $\neg P$ and $\neg Q$ cards in addition to those that they were biased towards. It seems unlikely that any knowledge or skills gained during a mathematics degree could predispose anybody to become more inclined to conduct this analysis; instead it could be hypothesised that an enthusiasm for logic puzzles might account for the increased engagement with the task that would be a prerequisite for such an analysis.[7]

[6]Once again these descriptions are intended to describe trends and tendencies only, as with all psychological research large individual differences between participants are to be expected.

[7]Some evidence for the idea that the mathematics students were more familiar and comfortable with dealing with questions such as the Selection Task comes from the behaviour of participants in Experiment 3. The mathematics group's inspection times for the instruction page were lower than the control group's. For the third (most complicated, see Figure 6.3) paragraph the mean inspection time for the control and mathematics groups was 12.9s and 8.8s respectively; this difference is significant, $t(30.5) = 2.49, p = 0.018$. For the first paragraph the mean inspection time for the control and mathematics groups was 13.5s and 10.2s respectively; this difference is approaching significance, $t(49) = 1.97, p = 0.054$.

It is reasonable to assume that an increased mathematical ability or increased mathematical experience might reduce the chances of confusing '$P \Rightarrow Q$' and '$P \Leftrightarrow Q$', but it seems unlikely that it would increase the chances of engaging enthusiastically with logic puzzles such as the Selection Task. Perhaps an enthusiasm for such matters is simply a matter of personal taste; and perhaps more people with such tastes are filtered into mathematics degrees.

This account, then, suggests why no correlation between mathematical experience or attainment was detected. The tendency to respond with the normatively correct answer is not related to either of these factors. Instead it is, perhaps, related to an enthusiasm with logic puzzles which is not developed during the course of a mathematics degree. The one factor that is developed by an increased exposure to mathematics is that which ensured that almost all the mathematics students in Experiments 1 and 2 analysed the parts of the problem that they considered normatively – namely the reduced tendency to conflate '$P \Rightarrow Q$' and '$P \Leftrightarrow Q$'.

6.7 Conclusions and summary of Chapter 6.

The main goals of this chapter were twofold:

- To distinguish which of the theories of reasoning discussed in Chapter 4 provide the best framework to investigate mathematicians' use and understanding of conditional statements.
- To explore how successful mathematics students respond to the Wason Selection Task.

The results from Experiments 1, 2 and 3 satisfy both these aims. Indeed, Experiment 3 has gone further, suggesting that the mental models, information value and mental logic theories of deduction struggle to satisfactorily account for the reasoning behaviour of mathematicians. Only the heuristic-analytic dual process theory put forward by Evans (1996) can account for the data reported here, and this is the framework that is taken forward into the remainder of this thesis. In the next chapter dual process theories are reviewed from a wider perspective, and compared to existing dualities in the mathematics education literature. The framework is then put to use in Chapter 8 to analyse how successful mathematics research students use and understand conditional statements in realistic mathematical contexts.

In summary, the main findings of this chapter were:

- Mathematics students respond differently to the general population on the Wason Selection Task.

- Mathematics students are preconsciously biased to the same parts of the problem as the general population, but are notably better at noticing and overriding the mistakes contained within these biases.

- It is hard to see how the mental logic, information value, or mental models theories can successfully account for these data. Conversely, strong support was found for Evans's (1996) heuristic-analytic dual process theory.

Chapter 7

Dual Processes Revisited

Following the empirical work contained in Chapter 6, it seems clear that the heuristic-analytic dual process theory of reasoning is best placed to account for the behaviour of successful mathematicians on the Wason Selection Task. In view of this finding, it is this theory which will be taken forward to the qualitative interview study reported in Chapter 8. However, before describing this study, the current chapter seeks to review dual process theories in greater detail. Some of the literature that has used versions of dual process theories in domains other than deductive reasoning is reviewed, and the theory's connections with frameworks that are well known in the mathematics education literature are discussed.

7.1 Dual process theories.

As discussed in Chapter 4, dual process theories posit the existence of two quite different systems (or sets of systems) in the brain that affect reasoning behaviour. System 1 is fast, automatic and preconscious. It is seen as being an innate system (although actually comprised of a set of systems), common to both humans and animals, that includes innate instinctive behaviours and (perhaps) domain-specific knowledge and skills.

System 2, on the other hand, is slow, conscious and analytical. It permits hypothetical thinking and is believed to be constrained by the limits of working memory capacity. It has been suggested that System 2 is unique to humans, although this is disputed by some animal behaviourists who see similarities between dual process theory and the theoretical distinction between stimulus-response (S-R) and cognitive processing that has been observed in

many animals[1] (e.g. Toates, 1998).

As mentioned above, although System 1 is seen as innate, some researchers argue that it can be developed over time through experience. For example, it has long been recognised that chess grandmasters, as well as having superior analytical skills, have a different way of 'seeing' the chess board to amateur players (Charness et al., 2001). Their experience of chess playing has altered their System 1 heuristics as well as developed their System 2 analytical skills. Stanovich (2004, p.40) explains that

> "conceptual systems and rules may enter [System 1] with practice. This is one way that humans structure their own cognition — by explicitly practising higher-level skills so that they become an automatized [System 1] process".

It could, however, be argued that the chess players are merely applying existing innate System 1 heuristics to new knowledge gained through their chess experience. Thus the change may be in how innate heuristics are applied rather than a change in the heuristics themselves. It is not at all clear how to distinguish between these two possibilities, or indeed whether a distinction of this sort is meaningful. Exactly how development of System 1 skills happens, and what limits there are on modifying it, remains an open question.

It is important to re-emphasise that many of the previous theories of reasoning discussed in Chapter 4 can comfortably fit within a System 1/2 framework. For example, mental models, mental logic, information value theory and relevance theory could all be seen as attempts to explain the mechanisms behind how either System 1 or System 2 operate. (It is less easy to situate either pragmatic reasoning schemas or social contract theory within a dual process framework, as both appear to dramatically underestimate the role of System 2). In this sense dual process theory can be seen as less of a theoretical framework, and more a framework for theoretical frameworks.

In addition to the empirical data that has been gathered from reasoning research, there is also neuropsychological evidence that supports the dual process account of reasoning. Goel and Dolan (2003) used fMRI brain scans whilst participants took standard reasoning tasks. They found that responses traditionally associated with System 1 were related to activity in the ventral medial prefrontal cortex, whereas the logically correct System 2 responses originated in the right inferior prefrontal cortex, an entirely different part of the brain (see also Parsons & Osherson, 2001).

[1]Indeed, in some ways, the theoretical models associated with this animal behaviourist work are more advanced than the dual process theories of human reasoning discussed here; particularly regarding the question of how experiences can affect the operation of the two Systems.

Stanovich (2004) went further than some dual process theorists by attributing different purposes and origins to the different systems. He argued that System 1 and System 2 have fundamentally different goal structures. That is to say that an individual's System 1 is shaped by evolution to maximise the likelihood of reproducing their genes. It is a "short leash" control mechanism of behaviour that is pre-programmed by your genes,[2] and has the aim of preserving and reproducing those same genes. System 2, in contrast, is formed by cultural memes[3] and is focussed on an individual's personal goals. It "instantiates a flexible goal hierarchy that is oriented toward maximising goal satisfaction at the level of the whole organism" rather than at the genetic level (Stanovich, 2004, p.64).

In short, it is possible for System 2 to override the System 1 response, and thus gain advantage for the individual rather than the genes. Stanovich referred to this as "the robot's rebellion". Using contraception during sex is a clear case of this. It is in the interests of your genes to always have unprotected sex, as it maximises the chances of genetic reproduction, however the analytical System 2 part of your brain realises that it may not be in your best interests to obey this heuristically cued response. System 2 allows you to take the benefits of sexual pleasure without the long term costs that could be associated with it. It is straightforward to list many such examples of where an individual's goals are best served by the overriding the genetic heuristics. However, as we have seen, it is not always the analytical System 2 that prevails when there is conflict between it and System 1.

In terms of the notion of dual rationality discussed in §4.5, Stanovich (2004) was proposing a further type of rationality. Whereas Evans and Over (1996a) suggested that rationality$_1$ is about enhancing *your* goals, and rationality$_2$ is about behaving in accordance to some normative standard, Stanovich suggested we also need to consider evolutionary rationality, which is about acting in the interests of *your genes*.

There are critics of dual process theory. Some, for example, claim that two distinct systems are not necessary to explain the various experimental results. Osman (2004) argued that a single cognitive system which distinguishes between cognition in a three dimensional 'dynamic graded continuum' better accounts for the experimental data than a two system model. In this model System 1 and 2 sit at different sides of the continuum (see also Cleeremans & Jiménez, 2002).

---

[2]Here genes should be understood in the sense of *The Selfish Gene* (Dawkins, 1976).
[3]Again, in the sense of Dawkins (1976).

No police dogs are vicious.
Some highly trained dogs are vicious.
Therefore, some highly trained dogs are not police dogs.

No nutritional things are inexpensive.
Some vitamin tablets are inexpensive.
Therefore, some vitamin tablets are not nutritional.

Both these deductions are valid, but whereas the first is rated as sound by 86% of the population, the second is so rated by only around 62% (Evans et al., 1983).
For logically invalid deductions the percentages accepting believable and unbelievable content were 66% and 13% respectively.

Figure 7.1: Belief bias, a System 1 heuristic.

7.1.1 Other System 1 biases in reasoning.

Alongside the matching- and if-heuristics discussed in Chapter 4, there are numerous other examples of apparently irrational System 1 heuristics. For example, it has been found that, in syllogistic reasoning tasks, valid deductions with unbelievable conclusions are deemed valid far less often than those with believable deductions (Evans, Barston, & Pollard, 1983). This phenomena has become known as *belief bias* (see Figure 7.1). Evans et al. (1983) explained it by suggesting that there is a heuristic within System 1 that leads people to be more willing to accept believable conclusions without any logical analysis.[4] Clearly this sort of heuristic makes evolutionary sense: why waste time evaluating an argument whose conclusion you know to be true? There are, however, alternative accounts. Oakhill, Johnson-Laird, and Garnham (1989) adopted a mental model framework and suggested that a believable conclusion would reduce the chances of the individual fleshing out alternative models.

Support for the dual process interpretation of these findings came from Evans and Curtis-Holmes's (2005) study. They asked participants to determine whether various syllogisms were valid or not under two conditions: one group had unlimited time, and the other had to respond within 10 seconds. They found that the belief bias effect increased with time restrictions. The dual process account would predict this, as, when under severe time restrictions, it is harder to use slow System 2 processes to accurately analyse and override mis-

[4]Although, of course, their work predates the System 1 – System 2 terminology introduced by Stanovich and West (2000).

leading System 1 biases. Further support for the dual process interpretation of this bias was found by Inglis (2006) who demonstrated that successful mathematics students are less influenced by the believability of a conclusion when judging whether it follows from a conditional than the general population.

Another System 1 bias investigated in the reasoning literature is known as *negative conclusion bias.* Evans (1977) found that participants on conditional inference tasks (e.g. Figure 4.2, p.42) are more likely to endorse negative conclusions than positive conclusions. That is to say that participants are more likely to make a modus tollens deduction for a rule of the form '$P \Rightarrow Q$' than they are for a rule of the form '$\neg P \Rightarrow Q$': the conclusion for the former is negative ($\neg P$), but for the latter the correct conclusion is positive (P).

Evans et al. (1995) noted that, along with dual process theories, both the mental models and the mental logic theories of reasoning can account for conclusion bias through a 'double negation process' (see also Schroyens, Schaeken, Fias, & d'Ydewalle, 2000).

7.1.2 Heuristics and biases in decision making.

Dual process theories of reasoning have, historically, emerged from two quite distinct research programmes. The traditional psychology of reasoning literature (reviewed above, and in Chapter 4) and the decision making literature. The latter work, whilst not directly relevant for the purposes of this thesis is important enough to merit a brief discussion and review.

Since the Sixties there has been mounting evidence that participants when making decisions under uncertainty do not always respond in a normative manner. This research programme was primarily conducted in the United States by Kahneman, Tversky and colleagues, and it was for this work that Kahneman was awarded the Nobel Prize for Economics in 2002 (Kahneman, 2003). It is interesting to note that until relatively recently there has been little cross-over between this programme of research and the pure reasoning work conducted by Wason, Evans, Johnson-Laird and colleagues (for a discussion of the historical development of the dual process theories see Evans, 2004b).

Perhaps the most famous instrument from this research tradition, known as the heuristics and biases programme, is the 'Linda' problem (Tversky & Kahneman, 1983). In this task participants were told:

> "Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student she was deeply concerned with issues of discrimination and social justice and also participated in antinuclear demonstrations."

Participants were then given eight possible descriptions of her present employment and activities, and were asked to rank them in order of probability. Intriguingly, 85% ranked "Linda is a bank clerk and active in the feminist movement" as more probable than "Linda is a bank clerk". Clearly, such a ranking is impossible. Tversky and Kahneman named this the conjunction fallacy, and explained it by noting that Linda resembles the prototypical feminist bank clerk more than she resembles the prototypical bank clerk.

Tversky and Kahneman, however, argued that it would be unreasonable to claim that their (highly educated) participants had conscious conceptions of probability that were largely based on resemblance to prototypical examples. Instead, the dual process account argues that the Linda tasks standard response comes from System 1. It is intuitive, automatic and more concerned with social data than formal models of probability. System 2 cues the opposite response, that which notes that $P(A)$ cannot possibly be less than $P(A \cap B)$. Individuals who respond with the conjunction fallacy, then, fail to successfully use System 2 to monitor and correct their intuitive System 1 output.

The heuristics and biases programme has uncovered many other systematic System 1 heuristics that can detract from normative decision making. For example, Tversky and Kahneman (1973) spoke of the 'availability heuristic':

> "A person is said to employ the availability heuristic whenever he estimates frequency or probability by the ease with which instances or associations could be brought to mind" (p.164).

A good example of this effect came by asking experimental participants to quickly estimate one of the following products:

- $1 \times 2 \times 3 \times 4 \times 5 \times 6 \times 7 \times 8$
- $8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1$

Participants in the first group (who estimated the first product) replied with an average estimate of 512, but the figure for those in the second group was 2250. The correct answer is 40230. Tversky and Kahneman argued that because of the order of the digits, in the second version the larger numbers are more available than the smaller numbers, and vice-versa for the first version.

A similar effect was observed by Ross and Sicoly (1979), who asked a group of husbands and wives to independently estimate what proportion of various household tasks they performed. Both husbands and wives estimated that they undertook well over half of the odd-jobs. Clearly they cannot be both correct. The availability heuristic explanation of this finding notes that the examples of chores that come easily to mind are those that you yourself have performed, so you are likely to attach more significance to your own actions that those of

anybody else. This finding has implications in very many domains: there are, for example, clearly lessons for researchers who need to determine the order of authorship on academic papers.

The heuristics and biases programme has come under sustained and fierce criticism from defenders of human rationality (e.g. Gigerenzer, 1991, 1996; for a reply see Kahneman & Tversky, 1996; c.f. §4.5). However, these criticisms and counter-criticisms are not directly relevant for the purposes of this thesis. The key thing, for our purposes, is to note that the dual process theories of reasoning and decision making have been applied across many different domains with much explanatory success.

7.2 Dual processes in mathematics education.

7.2.1 The intuitive/analytical distinction.

The role of intuition in mathematics has been noted by many great mathematicians (e.g. Hadamard, 1945; Hahn, 1933/1960; Poincaré, 1905) and philosophers of mathematics (e.g. Feferman, 2000), but the first major study of intuitions in mathematics education was conducted by Fischbein (1987). Fischbein defined an intuition as a cognition characterised by several properties:

> "An intuition is, then, an *idea* which possesses the two fundamental properties of a concrete, objectively-given reality; *immediacy* – that is to say intrinsic *evidence* – and *certitude* (not formal conventional certitude, but practically meaningful, immanent certitude)" (p.21)

Fischbein saw intuition as, in some sense, the opposite of deductive reasoning, arguing that "no productive mathematical reasoning is possible by resorting only to formal means" (p.16). Instead, he argued well developed mathematical intuitions are necessary for the creative aspects of mathematics, and that the development of such intuitions should be a major goal of instruction. Fischbein's definition of intuition is similar to that given by other researchers:

> "Intuition is that faculty of the mind for which comprehension is spontaneous and immediate as opposed to rational and linear, and very often, though not always, sudden" (Schmalz, 1988, p.34)

Again, in this characterisation, intuition is seen as being in some sense the opposite of rational thought.

Whereas Fischbein (1987) saw intuition as something important to be developed, not all mathematicians and philosophers take the same view. Hahn (1933/1960), for example, argued that mathematical intuition was unreliable

and untrustworthy. He cited several example of apparently paradoxical curves in support of this assertion (see the discussion on p.172). Other mathematicians have disagreed with Hahn's stance, by suggesting that intuition is an important part of mathematical cognition (e.g. Feferman, 2000; Hadamard, 1945; Poincaré, 1905).

Hersh (1998) wrote that intuition was "an essential part of mathematics" (p.61) but, nevertheless recognised that, like Hahn, some mathematicians do not agree:

> "One author takes pride in avoiding the 'merely' intuitive [...] Another takes pride in emphasising the intuitive" (pp.61-62).

Burton (2004) conducted a study where she asked professional research mathematicians to introspect on the nature of their working habits. She found that most of her sample recognised that intuition played an important role in their reasoning; although this was not the case for all. However, there was a feeling amongst many of her sample that "intuition" was not the most appropriate word to use, instead preferring "insight", "instinct" or "gut feeling".

When Burton (2004) asked her interviewees what exactly intuition was for them, she got mixed responses:

> "One mathematician explained: 'seeing the best way forward – that's intuition.' [...] Making connections was important: 'Insight is seeing a connection'; 'If the light switches on when I look at a problem, I have had an insight' [...] 'An enhanced understanding that comes to you suddenly'; 'Pattern matching is the best way of describing it'" (Burton, 2004, pp.76-77).

These responses all appear to share similarities with Fischbein's (1987) characterisation of intuition.[5]

Fischbein's (1987) notion of 'intuition' seems to be partially related to the dual process concept of System 1 cognition. However, there are important differences. Intuition and System 1 preconscious heuristics should not be identified as being identical. For example, Fischbein asserts that "an intuition is a theory,

[5]However, some care may be needed when listening to mathematicians speak about intuitiveness. Kemeny (1964) reported the following anecdote:

> "The mathematician's favourite word is 'trivial,' which is a shorthand way of saying 'intuitively obvious.' There are endless stories about the word 'trivial.' My favourite is the one about the mathematician who, in a lecture, asserted that a result is trivial. One of his colleagues challenged him, and they got into a long argument which was still going on at the end of the class. The class tiptoed out, and the two mathematicians were seen arguing vehemently for over two hours. When they finally showed up outside, students eagerly queried the challenger about the outcome. He replied: 'Oh, he was right. It *is* trivial'." (p.41).

never a mere skill or perception" (p.201), but dual process theorists would wish to include non-theories into System 1 cognition. Reflex actions, for example, are categorised as System 1 responses by Stanovich (2004), clearly these are not 'theories'.

Perhaps the most important distinction between Fischbein's (1987) work and dual process theories is the role of *attention.* Evans (1996) suggests that the primary purpose of System 1 is the allocation of attentional resources. System 1 'decides' what is relevant (what is worth spending conscious attention on) and System 2 then attends to these aspects of the environment. Fischbein attributes a different role to 'intuition' than the allocation of conscious attentional resources. For Fischbein, intuitive errors are the consequence of "incomplete or inappropriate" intuitions, not the consequence of failure of an analytical overriding process. Indeed analytical reasoning may not be present at all.

As a consequence of these differences, the two frameworks attribute quite different purposes for the education system. For dual process theorists such as Stanovich (2003), the role of teaching should be to develop System 2 analytical skills in an attempt to promote the successful analysis and evaluation of System 1 heuristic responses. Fischbein (1987), however, suggested that the responsibility of education should be to develop intuitive responses "by assimilating adequate formal structures" (p.209). Whereas Stanovich believed that System 1 responses need to be overridden using System 2, Fischbein believed that intuition itself needs to be modified.

Further differences in both scope and sophistication can be seen between Fischbein's (1987) work and dual process theories by looking at his analysis of the Selection Task, and contrasting it with the heuristic-analytic account described in Chapters 4 and 6. Fischbein (1987) wrote of the task:

> "Generally, the full correct solution is not found even by people who know the truth table of implication. Our hypothetical explanation is that these people have not assimilated *intuitively* the complete structure of implication. To assimilate intuitively means, according to our conception, to get the respective concept turned into an intrinsically obvious and behaviourally meaningful, efficient cognition." (pp.78-79)

When compared to any of the frameworks described in Chapter 4, this analysis seems to offer little more than a redescription of one part of the data (the low number of normatively correct responses). The notion of lack of intuitive assimilation can offer no explanation for the matching bias effect, the thematic effect or the training/education non-effect. Whereas it is possible to see Evans's (1996) heuristic-analytic account as a more sophisticated version of Fischbein's

(1987) work, it, as demonstrated in Chapter 6, does not attribute either 'success' or 'failure' on the task to not having the "complete structure of implication" internalised to System 1, instead it suggests that the key factor is how System 1 and System 2 interact.

7.2.2 Intuitive rules.

The successor to Fischbein's (1987) work on intuition is the 'Intuitive Rules' programme of research by Stavy and Tirosh (e.g. Stavy & Tirosh, 1996; Tirosh & Stavy, 1999). They propose that some aspects of students' reasoning can be understood as, and predicted by, the application of simple 'intuitive rules'. So, according to Stavy and Tirosh (1996), many incorrect answers to mathematical problems can be explained by the intuitive rule 'more A–more B'. That is to say that if an object has more of property A than another object, intuitively it is expected to have more of property B.

Zazkis (1999) asserted that this intuitive rule can be used to explain results from a study that looked at students' responses to a number theory task. She asked students to decide whether the following statement was true or false:

> If a natural number a is bigger than a natural number b, then the number of factors of a is bigger than the number of factors of b.

According to the more A–more B intuitive rule, students might intuitively believe that this statement is true, whereas in fact it is clearly false. Zazkis found that 4 out of 58 participants did in fact believe it was true. Strangely, however, this extremely low figure (7%) was not enough for Zazkis to question the role of the intuitive rule in this context, instead she wrote:

> "'The more of A, the more of B' appeared to be robust and not readily given up by some students even when they were confronted with new evidence." (p.207)

Similar applications of the intuitive rules theory were published by Tsamir (2003) in the context of a task that asked about the lengths of journeys around polygons.

Another intuitive rule that apparently effects students' reasoning patterns is the 'same A–same B' rule. That is, if an object has the same amount of property A as another object, it is intuitively likely that it has the same amount of property B. For example, Tirosh and Stavy (1999) cite the case of a hexagon and a pentagon (see Figure 7.2). In this task the students were told that the sides of the pentagon were equal to the sides of the hexagon, and asked the students to judge whether angle 1 was greater than, equal to, or less than angle 2.

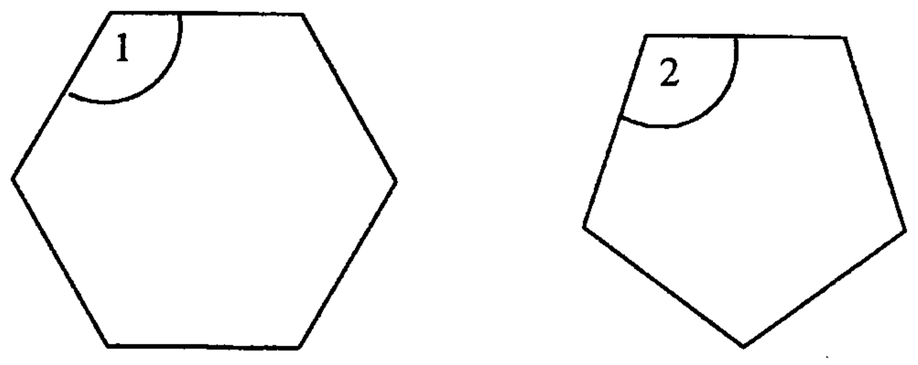


Figure 7.2: The two polygons from Tirosh and Stavy's (1999) paper.

The results from this study showed that large numbers of students incorrectly answered that the angles are equal ($\approx$ 50% in grades 4, 6 and 8). Interestingly, this figure was considerably lower in higher year groups (16% in grade 12). Tirosh and Stavy (1999) attribute the high number of incorrect responses from the lower grades to an application of the 'same A–same B' intuitive rule.

However, there are some potential flaws in Tirosh and Stavy's (1999) analysis. The hexagon in Tirosh and Stavy's task has *more* sides than the pentagon, so perhaps, according to the 'more A–more B' rule, the participants should have judged angle 1 to be greater than angle 2? Van Dooren, de Bock, Weyers, and Verschaffel (2004) published an extensive critique of the intuitive rules theory based on arguments along these lines. They argued that the supposed "predictive power" of the theory does not in fact exist. For example, they produced the following task:

> Is $\sqrt[3]{9}$ larger than / equal to / or smaller than $\sqrt{6}$?

According to the 'same A–same B' rule, students should respond with "equal to" as the ratios $\frac{3}{9}$ and $\frac{2}{6}$ are equal. But, argued Van Dooren et al., according to the 'more A–more B' rule, students should respond with "larger than" as $9 > 6$ and/or $3 > 2$. Across several tasks Van Dooren et al. found little or no evidence of systematic use of intuitive rules in students' responses. The key issue on these tasks appears to be where participants allocate their attentional resources, but Van Dooren et al.'s data suggest that Tirosh and Stavy's intuitive rules do not always accurately predict this.

It is important to note that even if one discards the methodological and conceptual concerns raised by Van Dooren et al. (2004) regarding the intuitive rule theory, its content and breadth are quite different to the dual process theories of Evans (1996), Sloman (1996) and Stanovich and West (2000).

Stavy and Tirosh's (1996) work is only concerned with the 'intuitive rules' of (very young) schoolchildren. Indeed, their data shows that the proportion of answers that are consonant with the 'more A–more B' and 'same A–same B' intuitive rules decline with age. Thus, the status of these intuitive rules, if

one brushes aside concerns over their conceptual coherence, would appear to be of a quite different form to the System 1 heuristics discussed earlier. The if- and matching-heuristics do not decline in importance with age, the evidence suggests that they are innate and manifest themselves in extremely high-ability students of all ages (see Chapter 6).

A second important difference between the theories is the level of detail that they provide. For example, not only do dual process theories set out to describe preconscious System 1 heuristics, they can suggest why they are there. An evolutionary analysis, supported by Oaksford and Chater's (1994) Bayesian model, provides ample evidence for the ecological benefit of the if- and matching-heuristics. No such analysis is present with Stavy and Tirosh's (1996) intuitive rules.

Further differences between the intuitive/analytical distinction of Fischbein (1987) and followers, and the dual process theories discussed earlier were pointed out by Leron and Hazzan (2006). They noted that, in mathematics education circles, there has often been a distinction between the use of the term 'cognition' and 'metacognition' (e.g. Schoenfeld, 1992). Leron and Hazzan pointed out that, within a dual process theory framework, metacognition is clearly a System 2 analytical process. They write

> "The added value of using dual-process theory in mathematics education research is not the distinction between intuitive and analytical thinking; the distinction itself is of course ancient and well-known and much has been written about it. Rather, we see the added value in tightening, refining and operationalising this distinction." (Leron & Hazzan, 2006, p.23)

Leron and Hazzan were right to note that the operationalisation of dual process theory is on substantially stronger theoretical and empirical grounds than the intuitive rules research programme in the mathematics education literature. One obvious benefit of this operationalisation was demonstrated in Chapter 6: by using chronometric analyses such as eye-tracker devices it is possible to directly scrutinise the differing roles of heuristic and analytical processes.

7.2.3 Skemp's Δ_1 and Δ_2.

There have been a number of other dualities discussed in the mathematics education literature. For example, Marton and Saljo (1976) suggested that there are two distinct approaches to learning, which they termed the deep and surface approaches. Along a similar line, Duffin and Simpson (1993) argued that learning experiences could be categorised as being eith 'natural' or 'alien'. Although it may be tempting to see such dualities as being similar to the System

1–System 2 distinction proposed by dual process theorists, these theories are quite different in both scope, aim and scale.[6] There is, however, one further duality from the mathematics education literature that merits further scrutiny to see whether its superficial similarities with dual process theories stand up to examination.

Skemp (1979) proposed that learning can be analysed in terms of two distinct 'director systems' which he labelled delta-one (Δ_1) and delta-two (Δ_2). A director system, for Skemp, is a cognitive system which directs the ways in which effort is applied in order to help meet goals of the organism. Skemp suggested that Δ_1 is a first order director system which operates on features of the physical environment. Δ_2, in contrast, operates on Δ_1. Its role is to manipulate the operation of Δ_1 until it functions optimally.

There are some clear similarities between Skemp's (1979) notions of dual director systems and the System 1 and System 2 of dual process theories. For example, in Evans's (2006) heuristic-analytic theory, it is emphasised that System 2 *acts on* System 1 output, in much the same way as Δ_2 acts on Δ_1. However, when the two theories are scrutinised in greater detail it becomes clear that there are also some important differences.

Firstly, Skemp (1979) emphasised that Δ_1 is *teachable.* Indeed, he saw the primary role of Δ_2 as being the 'teacher' of Δ_1. In contrast, dual process theorists are (at best) ambivalent about whether it is possible to teach System 1. As discussed above, the role education is seen as, primarily, being about the improvement of System 2 operation (e.g. Stanovich, 2004).

Secondly, some clear distinctions between the work of Skemp (1979) and the dual process theorists can be found by examining specific examples of operation. For example, Skemp suggests that translating an unseen prose passage from French to English as being an "intellectual task at the Δ_1 level" (Skemp, 1979, p.82). This is a Δ_1 task in Skemp's terms as it involves operating on objects from the physical environment, namely the passage of text. For dual process theorists, however, this task clearly involves a combination of System 1 and System 2 processing. As has been emphasised throughout this chapter, System 1 first directs attention towards relevant parts of the environment and then System 2 analyses these apparently relevant features.

Thirdly, Skemp (1979) makes no bold claims about the operational characteristics of his two director systems. Whereas the list of features of System 1 and System 2 are quite explicit (see Table 4.2). System 1 operates in a quick and automatic fashion, the same is not true of Δ_1.

[6]Strangely not everybody agrees with this analysis. Bizarrely, one anonymous reviewer suggested to me that dual process theories are merely a redescription of Duffin and Simpson's (1993) natural/alien distinction, but offered no justification for this view.

1. A student wrote in an exam:

"$\mathbb{Z}_3$ is a subgroup of $\mathbb{Z}_6$".

In your opinion is this statement true, partially true, or false? Please explain your answer.

2. A student wrote in an exam:

"S_4 is a subgroup of S_5".

In your opinion is this statement true, partially true, or false? Please explain your answer.

Figure 7.3: The Lagrange's Theorem task.

In summary, whilst the Δ_1/Δ_2 developed by Skemp (1979) shares some similar features to dual process theories of reasoning, there is no direct one-to-one mapping between the theories either in terms of the conceptual relationships they specify, or the empirical observations they seek to account for.

7.2.4 Uses of dual process theory in maths education.

Leron and Hazzan (2006) analysed the so-called Lagrange's Theorem Task (Figure 7.3) using dual process theory. They suggested that research that found a large minority of students are prone to confusing Lagrange's Theorem[7] with its converse (Hazzan & Leron, 1996) can be accounted for by a re-analysis using a dual process framework. In effect, using the language discussed in §4.4.7, Leron and Hazzan claimed that the Lagrange's Theorem Task satisfies Criterion T.

Hazzan and Leron (1996) found that 27% of students invoked Lagrange's Theorem to answer the first part of the task (shown in Figure 7.3). That is to say that they constructed an argument of the form:

> $\mathbb{Z}_3$ is a subgroup of $\mathbb{Z}_6$ because 3 divides 6.

Hazzan and Leron found that, in the second example, 29% of students answered 'false' and justified this response with an argument of the form:

> S_4 is not a subgroup of S_5 because 4 does not divide 5.

[7]Lagrange's Theorem states that if H is a subgroup of a finite group G, then $|H|$ divides $|G|$.

Write an equation for the following statement: "There are six times as many students as professors at this university". Use S for the number of students and P for the number of professors.

Figure 7.4: The Students and Professors problem.

Both these arguments are incorrect, but both are similar in the sense that they are using surface features of the problem in an attempted application of Lagrange's Theorem. Leron and Hazzan's (2006) new interpretation of these results is that the participants' attention was drawn to apparently relevant parts of the problem by System 1 heuristics, and that the lack of System 2 monitoring and criticism was responsible for the incorrect answer. In essence, Leron and Hazzan (2006) argued that the process that produced the incorrect responses to the Lagrange's Theorem task is structurally identical to the heuristic-analytic account of the Selection Task. It would be worthwhile to test this hypothesis further. For example, if this analysis is correct one might expect a rapid-response methodology, of the type used by Evans and Curtis-Holmes (2005), to increase the percentage of these specific incorrect answers.

Leron and Hazzan (2006) also noted that a dual process interpretation can be applied to the 'Students and Professors' problem (Clement, Lockhead, & Monk, 1981). In this task (Figure 7.4) it was found that 37% of university students answered incorrectly, and that two-thirds of these mistakes were of the form '$6S = P$'.

As with the Lagrange's Theorem task, Leron and Hazzan (2006) argued that these responses could be seen as a consequence of linguistic surface relevance features of the problem. Again, it would be interesting to adopt a rapid-response methodology to this task to test specific predictions of the dual process account, as there are other competing explanations (e.g. Crowley, Thomas, & Tall, 1994).

There are many other tasks that are well known to the mathematics education community which can be seen from a dual process perspective. One such example, the 'waiter's profit' task, is given in Figure 7.5. In this task participants are confused by surface relevance features of the problem that there is a 'spare' pound. It is only after a careful, slow and analytical analysis of the situation that the explanation can be revealed.

The waiter's profit?

Three people are in a restaurant. The bill comes to £30 and each person puts in £10. The manager of the restaurant tells the waiter "they are good customers, give them £5 back".

The waiter, being a little dishonest, gives them each back £1 and pockets £2. This means that each has paid £9 for their meal, whilst the waiter has pocketed £2. Now, $3 \times$ £9 is £27. The clever waiter has £2. Where is the other pound?

(Johnson-Wilder, private communication)

Figure 7.5: The 'waiter's profit' task.

7.3 Summary of Chapter 7.

Dual process theories of reasoning are a growing field. This chapter has sought to expand upon the range of literature reviewed in Chapter 4, and to establish links between dual process theories and distinctions which have been noted by earlier mathematics education researchers:

- Along with the matching- and if-heuristics discussed in Chapter 6 with reference to the Selection Task, other System 1 heuristics have been identified from the reasoning and decision making literatures.
- It is possible to see dual process theories as more sophisticated and developed versions of the intuitive/analytical divide that has been discussed in the mathematics education literature, although there are still important differences between the theories.
- Several established results from the mathematics education literature can be seen and analysed in terms of dual process theory.

The application of dual process theorists in mathematics education is a new and growing field, which could have applications in many areas (Ejersbo, Inglis, & Leron, 2006). In the next chapter, the framework is put to use in a study which seeks, amongst other aims, to explore whether the if- and matching-heuristics play a role in advanced mathematical thinking in a realistic environment.

Chapter 8

Experiment 4: Applying the Framework

Chapter 6 argued, using empirical data from three experiments, that the dual process theory of reasoning provides the best framework within which to analyse how mathematicians reason with conditional statements. The purpose of this chapter, then, is to conduct such an analysis within the context of a realistic mathematical setting.

There were several specific goals:

- To investigate the role of preconscious System 1 heuristics in mathematical reasoning with conditionals. Do the matching- and if-heuristics impact on strategies in realistic mathematical contexts in the same way that they do in the Selection Task?

- To investigate the types of processes mathematicians go through when evaluating the validity of conditional statements.

The general discussion in Chapter 5 highlighted the benefits and drawbacks of qualitative and quantitative methods. In view of the main aims of this experiment it was felt that a qualitative method was more appropriate as it would be hard to claim that any standardised test would truly be a "realistic mathematical context". It was also felt that the level of richness of data required to satisfactorily meet the specified goals would be impossible to obtain from a quantitative study.

Naturally a qualitative study takes more space and time to report than the quantitative work of Chapter 6, and so an entire chapter is devoted to describing the task, and reporting the outcome.

This chapter begins by introducing the task, how it was developed, and

describing the methods and participants in the Experiment, it concludes with three lengthy sections:

- The first section (§8.5) analyses the role that preconscious heuristics played in participants' responses to the task. It is argued that, as in the Selection Task, the if- and matching-heuristics play an important role in shaping conscious thought in this mathematical situation.

- The second section (§8.6) begins to develop a theory that models the evaluation of mathematical conditionals. The groundwork for this theoretical attempt is laid in this chapter through the introduction of Toulmin's (1958) argumentation scheme. It is argued that the manner in which the scheme has been applied to mathematics by earlier researchers is inadequate for dealing with all types of mathematical reasoning.

- In the final section (§8.7) the data reported in the earlier two sections is drawn together with Evans and Over's (2004) theory of suppositional conditionals to develop a model of how successful mathematicians evaluate mathematical conditionals.

8.1 The task.

The task used for this experiment ("the Abundant Number Task") is given here. Participants were first given the following information, on a A5 card:

> ***All the numbers below should be assumed to be positive integers.***
>
> **Definition.** An *abundant number* is an integer n whose divisors add up to more than $2n$.
>
> **Definition.** A *perfect number* is an integer n whose divisors add up to exactly $2n$.
>
> **Definition.** A *deficient number* is an integer n whose divisors add up to less than $2n$.
>
> **Example.** 12 is an abundant number, because $1+2+3+4+6+12 = 28$ and $28 > 2 \times 12$. However, 14 is a deficient number, because $1 + 2 + 7 + 14 = 24$, and $24 < 2 \times 14$.
>
> Your task is to consider the following conjectures and determine, with proofs, whether they are true or false.

When participants indicated that they were ready to proceed they were given the first of the conjectures shown in Figure 8.1, again on A5 card. Participants had unlimited amounts of paper on which to work. Only after participants had

dealt with the first conjecture to their satisfaction were they given the card containing the second conjecture, however they retained access to the previous cards that they had been given. Only one participant managed to complete every conjecture in the allotted hour, and discretion was used by the interviewer in determining whether or not to miss out certain conjectures in order to maximise the use of time. The conjectures, were, however, always presented in this order they are shown in Figure 8.1.[1] Enthusiastic readers may wish to study the conjectures themselves before proceeding to the next section.

Conjecture (1). *A number is abundant if and only if it is a multiple of 6.*

Conjecture (2). *If n is perfect, then kn is abundant for any $k \in \mathbb{N}$.*

Conjecture (3). *If p_1 and p_2 are primes, then $p_1 p_2$ is abundant.*

Conjecture (4). *If n is deficient, then every divisor of n is deficient.*

Conjecture (5). *If n and m are abundant, then $n + m$ is abundant.*

Conjecture (6). *If n and m are abundant, then nm is abundant.*

Conjecture (7). *If n is abundant, then n is not of the form p^m for some natural m and prime p.*

Figure 8.1: The Abundant Number Task.

If the participant began to check some examples, or asked whether any examples were available, they were given a further card:

Examples.
The first few abundant numbers are: $12, 18, 20, 24, 30, 36, 40, \ldots$
The first few perfect numbers are: $6, 28, 496, 8128, \ldots$

8.2 Designing the task.

When designing the Abundant Number Task there were several priorities. In view of the aims and objectives of the study it was important to ensure that the conjectures were as realistic as possible. That is to say that they needed to appear to the participants as being natural successors to one another. As a consequence of this objective Conjectures 2–7 were generated during the course of an investigation into the properties of abundant numbers.

[1] i.e. Conjecture n was presented after Conjecture m if $m < n$.

Conjecture 1 was based on previous empirical research into mathematical reasoning. In an extremely interesting, yet unpublished, piece of research from the early eighties Markowitz and Tweney (1981a, 1981b) reported a study that looked at the confirmatory and disconfirmatory reasoning strategies of mathematicians. In earlier work Mynatt, Doherty, and Tweney (1978) had investigated how research scientists test conjectures in a simulated research environment. Markowitz and Tweney's work was an attempt to extend this research to mathematicians; thus allowing the comparison of confirmation and disconfirmation strategies between the disciplines.

To do this they presented twelve research mathematicians (primarily research students) with Conjecture 1, and asked them to first evaluate the conjecture, and then proceed in any manner they chose, bearing in mind their primary goal: to "determine the divisibility properties necessary and sufficient for a number to be abundant". After two hours of solitary work, during which time the experimenter took notes on the participants' behaviour, the participants were asked to organise and summarise their work.

Markowitz and Tweney's (1981a) method was a hybrid of their own construction. They used participants' protocols, the experimenter's notes and a post-experimental interview to construct an account of the participants' reasoning process. From a more recent perspective, their analysis procedure was highly unorthodox. Ten "interesting" conjectures were determined from amongst the participants' replies, and a pseudo-quantitative analysis conducted on the participants' reasoning patterns. For example, participants were awarded 'points' for proposing 'interesting' conjectures, and further points for proving them.

In view of the differing aims of the current study and their's, and the methodological discussion in Chapter 5, it was felt that a clinical interview would be more appropriate than Markowitz and Tweney's (1981a) approach, both methodologically and practically. The richness and non-triviality of the mathematical context that Markowitz and Tweney's (1981a) opening conjecture provided, however, was deeply impressive. It was resolved to use this abundant number environment to explore the current research question: How do mathematicians evaluate conditional statements, and what is the role of System 1 heuristics?

In view of the concerns with making the ordering of conjectures as realistic as possible, Markowitz and Tweney's experimental instructions were retrospectively followed by myself. What follows is a rough introspective account of this attempt, together with some discussion of how the conjectures ended up in their final form on the interview schedule.

Conjecture (1). *A number is abundant if and only if it is a multiple of 6.*

False.

There are easy counterexamples to this conjecture in both directions. However, further investigations reveal that all *proper* multiples of 6 are abundant. The proof of this fact is non-trivial, and relies upon picking an appropriate subset of divisors of $6k$, so that when you add up all the divisors in this subset you get a total of more than $12k$.

Proof of Conjecture 1. $\Rightarrow$: 20 is a counterexample, it is abundant, but not a multiple of 6 ($1+2+4+5+10+20 = 42 > 2\times 20$). $\Leftarrow$: 6 is a counterexample, it is a multiple of 6 and is perfect, not abundant. However, all multiples of 6 apart from 6 are abundant. Take $n = 6k$ for some $k \neq 1$. Then $1, k, 2k, 3k, 6k$ are divisors of n. This gives: $1+k+2k+3k+6k = 12k+1 > 2\times 6k = 2n$. □

Conjecture (2). *If n is perfect, then kn is abundant for any $k \in \mathbb{N}$.* ***False.***

Conjecture 2 is a generalisation of Conjecture 1. Whilst investigating the first conjecture it was found that all proper multiples of (the perfect number) 6 are abundant, so it seemed natural to wonder whether this is true for all perfect numbers. It is (and can be proved in a similar fashion to the argument in Conjecture 1), although the conjecture as written is false since for $k = 1$ it is clear that kn is perfect, not abundant.

Proof of Conjecture 2. The statement is true provided $k \neq 1$. Suppose n is perfect with divisors $d_1, d_2, \ldots, d_r$ (i.e. $2n = d_1 + \ldots + d_r$). Then kn has amongst its divisors $1, kd_1, kd_2, \ldots, kd_r$, and these sum to $2kn + 1 > 2kn$. So kn is abundant. □

Conjecture (3). *If p_1 and p_2 are primes, then p_1p_2 is abundant.* ***False.***

The key issue in proving Conjectures 1 and 2 was finding the correct divisors to add up. Considering the properties of divisors of natural numbers naturally leads you to consider primality, and what relation primality has to abundantness (c.f. Weber's (2001) notion of 'strategic knowledge'). It was easy to see that all prime numbers are deficient, so the properties of numbers that are multiples of two primes were investigated. It is fairly straightforward to show that these numbers cannot be abundant.

Proof of Conjecture 3. $p_1 = 2, p_2 = 3$ is a counterexample. In fact it is true to say that if p_1, p_2 are prime then p_1p_2 is *not* abundant. For this we need to show, assuming $p_1 \neq p_2$, that $1 + p_1 + p_2 + p_1p_2 \leq 2p_1p_2$. This reduces to $1 + p_1 + p_2 \leq p_1p_2$. This is equivalent to $(p_1 - 1)(p_2 - 1) \geq 2$ which is clearly true for all p_1, p_2 other than $p_1 = p_2 = 2$. □

At this stage the investigation moved on to consider the properties of sums of abundant numbers (Conjecture 5). But for the research instrument it was also important to evaluate the role of System 1 heuristics in mathematical reasoning. One of the heuristics that turned out to play a vital role in the Selection Task was the so-called if-heuristic. This heuristic plays the role of directing attention towards the antecedent of a conditional statement, i.e. the P in $P \Rightarrow Q$. The next conjecture was designed to investigate whether this is true in genuinely mathematical contexts.

Conjecture (4). *If n is deficient, then every divisor of n is deficient.* ***True.***

This conjecture is (almost) the contrapositive of Conjecture 2. In fact the proof of Conjecture 2 proves a stronger statement: that kn is abundant (or perfect) for any perfect or abundant n. Conjecture 4 is the contrapositive of this stronger statement, and therefore should be a straightforward corollary of Conjecture 2. However, the if-heuristic predicts that things will not be so simple: attention will be directed towards the statement's antecedent – the case when n is deficient. In fact, to prove this statement using the contrapositive it is necessary to consider the case when n' (a divisor of n) is abundant or perfect. The prediction, then, is that this conjecture sets up a Criterion T situation (see §4.4.7).

Proof of Conjecture 4. Consider the contrapositive: if n is not deficient, then kn is not deficient. Suppose n has divisors $d_1, \ldots, d_r$ and that $d_1 + \ldots + d_r \geq 2n$. Then the set of divisors of kn contains $1, kd_1, kd_2, \ldots, kd_r$. And we know that $1 + kd_1 + \ldots + kd_r \geq 2kn + 1$. Therefore kn is not deficient. This is the same statement as Conjecture 2 (with a small modification). □

Conjecture (5). *If n and m are abundant, then $n + m$ is abundant.* ***False.***

This conjecture asserts that abundantness is preserved under addition, but finding a counterexample is trivial.

Proof of Conjecture 5. 20 and 12 are abundant, but 32 is deficient ($1 + 2 + 4 + 8 + 16 + 32 = 63 < 2 \times 32$). □

In an attempt to extend Conjecture 5, the next conjecture to be investigated asserted that abundantness is preserved under multiplication.

Conjecture (6). *If n and m are abundant, then nm is abundant.* ***True.***

This statement turned out to be surprisingly difficult to prove, despite, on reflection, being a straightforward consequence of Conjecture 2. However, the if-heuristic predicts that this statement will indeed be difficult to prove, as it sets up a Criterion T situation. To recap, the if-heuristic directs conscious attention towards the antecedent of the statement – in this case the situation

where n and m are both abundant – thus disguising the fact that Conjecture 2 is relevant. If attention were not biased towards the antecedent of the statement, it would perhaps be straightforward to see that it is a trivial special case of Conjecture 2. Interestingly, Hadamard (1945) described a similar, yet more extreme, situation:

> "I have several times happened to overlook results which ought to have struck me blind, as being immediate consequences of other ones which I had obtained. Most of these failures proceed [...] from attention too narrowly directed" (p.50)

Hadamard went on to give some clear cut examples from his own mathematical career, and the careers of others (pp. 49-52). In the abundant number context it seems that the if-heuristic directs attention too narrowly – at the antecedent as written – and thus inhibits progression towards the trivial conclusion that this conjecture is a straightforward consequence of Conjecture 2.[2]

Proof of Conjecture 6. This is a specialisation of the claim that all multiples of abundant numbers are abundant, so it is true. □

Conjecture (7). *If n is abundant, then n is not of the form p^m for some natural m and prime p.* ***True.***

The final conjecture was, in some sense, an extension of Conjecture 3. But the final formulation that appeared on the interview schedule was designed to investigate the role of the matching-heuristic. The statement is straightforward to prove indirectly by considering its contrapositive; the matching-heuristic should aid this process as, according to the dual process account, it directs attention

[2]My own initial approaches at proving Conjecture 4 revolved around assuming n and m are abundant, cross multiplying the set of divisors of n with the set of divisors of m, and adding up the resultant divisors. Unfortunately far too many repeat divisors are created in this cross multiplication, and so when the addition process takes place at the end, the sum is artificially large. I also attempted the decomposition of n and m into products of primes and attempting to count up the primes in nm. So, for example, I considered

$$n = p_1^{\alpha_1} \dots p_r^{\alpha_r} \text{ and } m = p_1^{\beta_1} \dots p_s^{\beta_s}$$

where the α_i and β_i may be zero, and it is assumed (without loss of generality) that $r > s$. Multiplying these two products together gives

$$nm = p_1^{\alpha_1+\beta_1} \dots p_s^{\alpha_s+\beta_s} \dots p_r^{\alpha_r}.$$

Using this technique I hoped that the divisors of nm would reveal themselves. Unfortunately they didn't: this strategy proved both complicated and fruitless. Having unsuccessfully attempted to prove the statement, I sought assistance from a colleague who was also familiar with the task context and the first few conjectures. Together we ran through similar arguments to the ones discussed above, but were also unsuccessful at producing a correct proof. Disheartened, I gave up. But, strangely, a couple of days later, whilst driving home from work, the solution floated into my consciousness! Conjecture 6 is a trivial consequence of Conjecture 2! If kn is abundant for any $k > 1$ and abundant n, then certainly it is is the case that nm is abundant for any two abundants m and n.

towards the linguistic surface features of the rule, in this case the case where $n = p^m$.

Proof of Conjecture 7. Consider the contrapositive: if n is of the form p^m then n is not abundant. The divisors of p^m are $1, p, p^2, \ldots, p^m$, so we need to show that $1 + p + \ldots + p^m \leq 2p^m$. Do this by induction on m. Clearly true for $m = 1$. Assume true for $m = k$, giving $1+p+\ldots+p^k \leq 2p^k$. This implies that $1+p+\ldots+p^k+p^{k+1} \leq 2p^k+p^{k+1}$, which gives $1+p+\ldots+p^k+p^{k+1} \leq pp^k+p^{k+1}$ since $2 \leq p$. Therefore $1+p+\ldots+p^{k+1} \leq 2p^{k+1}$ which completes the induction step. □

Before the main study the interview schedule was piloted with a colleague, and minor adjustments were made to the design. The most notable change was the inclusion of the sheet of examples in order to minimise the amount of time that participants spent looking for abundant and/or deficient numbers.

8.3 Participants, method and analysis.

8.3.1 Participants.

Eight postgraduate volunteers were interviewed as part of this study, each participant was paid £10 for taking part. The participants were recruited through an appeal for volunteers sent to all postgraduate students researching mathematics or undergraduate mathematics education at the University of Warwick. All those who responded were interviewed. The backgrounds of the students are summarised in Table 8.1.

Code	Name	Research Area
A	Andrew	Functional analysis and partial differential equations
B	Ben	Quantum chemistry
C	Chris	Conformal field theory
D	David	Algebraic topology
E	Edward	Computational algebra in Lie theory
F	Fred	Philosophy of mathematics
G	Gary	Undergraduate maths education
H	Harry	Undergraduate maths education

Table 8.1: The participants in Experiment 4.

All the students had successfully completed undergraduate mathematics degrees and were now researching some area of mathematics or mathematics edu-

cation. All participants other than Fred were, at the time of interview, studying at the University of Warwick. Fred was at the University of London.[3] Participants A to F intended to follow an academic career in mathematics, and were at various stages in their Ph.D. research. Gary and Harry intended to follow careers as mathematics education researchers. All the participants had been highly successful undergraduate mathematics students, and participants A to F were amongst the most talented mathematics students in the country. It is on these cases that the following analyses will concentrate.[4]

8.3.2 Method.

Data collection.

Participants were interviewed alone in a seminar room or private office. The interviews were recorded with an electronic audio-dictaphone, and later transcribed for analysis. The possibility of video-taping the interviews was considered, but it was felt that such a method would be more intrusive than an audio recording and that the benefits of the extra data provided by a video recording would be minimal. In the early interviews the interviewer made notes during the course of the interview, but participants tended to find this distracting so the practice was stopped. Instead the interviewer wrote down reflective comments immediately on completion of the interview. Participants' rough notes made during the work were retained for analysis.

The interview began with the participants being presented the first card of instructions and being asked to read it aloud. When they were happy to move on the interviewer gave them the card containing Conjecture 1. Participants were asked to "think aloud" during the interview, and to verbalise anything that they wrote down on the rough paper provided.

The interviews followed a 'semi-structured' format (L. Cohen et al., 2001; Robson, 1993). Although the conjectures were never presented in a differing order, sometimes (in order to maximise the use of time) conjectures were missed out. For example, after participants had correctly solved Conjecture 3, most were asked about the modified conjecture "if p_1 and p_2 are primes, then p_1p_2 is **not** abundant." However, for example, Ben had spent disproportionately long investigating Conjectures 1 and 2, so the modified version of Conjecture 3 was omitted from the interview schedule in his case.

When necessary within each conjecture, the interviewer asked for clarifica-

[3]Fred had previously studied at Warwick.

[4]A further two participants were interviewed during the design of the task, all were researching undergraduate mathematics education. Although these data are not reported here, the case-law outlined in the following sections is consistent with incidents seen in these pilot interviews.

tion and explanation, and probed the participants in a fashion associated with the clinical interview method (Ginsburg, 1981; Swanson et al., 1981). When the interviewer felt that too long had been spent on a certain conjecture (typically over 15 minutes on a single conjecture), and that as a consequence no useful data was being collected, two strategies were adopted. Firstly the interviewer attempted to 'push' the participant towards a correct solution by the use of hints and leading questions. If that failed the participant was asked to stop and the interviewer demonstrated the solution, before moving on to the next conjecture. Happily, the first technique needed to be used sparingly, and the second only once. Notwithstanding this, none of the participants seemed to feel that the task was trivial, and only one, David, completed the whole of the task in the allotted hour. Several participants, however, asked to carry on until they had completed the conjecture they were working on to their satisfaction. As an example of a typical interview structure, the entire transcript of the interview with David is given in Appendix A.

Data analysis.

The analysis was conducted as a series of case studies using the quasi-judicial method, with the aim of developing a set of coherent case-law (for a full discussion of the quasi-judicial method of qualitative data analysis see §5.3, or Bromley, 1986). In practice this meant:

1. An interview was selected.
2. The interview was transcribed and imported into qualitative research software for analysis.
3. The interview was coded. Codes were based on the dual process framework (this was, to use Bromley's (1986) terminology, a *prima facie* explanation). The full set of codes used in the analysis are given in Appendix B.
4. Separate arguments within the interview were analysed using Toulmin's (1958) scheme, and warrants coded using Harel and Sowder's (1998) proof scheme framework (another *prima facie* explanation).
5. This process was undertaken for each interview in turn, before common themes were drawn together to form a set of coherent case-law. The report which follows describes this case-law.

In the following sections interview transcripts are reported together with the participant's code a number that represents which conjecture they were working on at the time. So, for example, 'A1' indicates that the transcript is from the portion of Andrew's interview where he was working on Conjecture 1.

8.4 Aside: a note about notation.

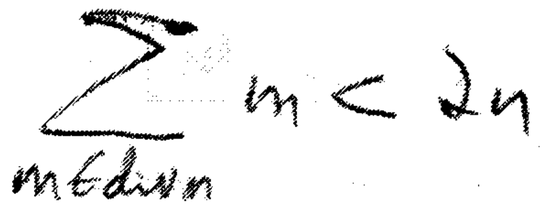
In the extracts that follow the standard notation regarding the divisors of a natural number is used. Given $n \in \mathbb{N}$, the sum of the divisors of n will be denoted by $\sigma(n)$. It is interesting to note that no participant in the study was familiar with this notation, and as a consequence, they were forced to invent their own. There were a wide variety of different notations invented by the various participants.

Some participants used adaptations of set-theoretic notation (e.g. David), some used sentences of text (e.g. Ben), and others used hybrid notations of their own construction (e.g. Fred). Examples of various notations are shown in Figure 8.2.

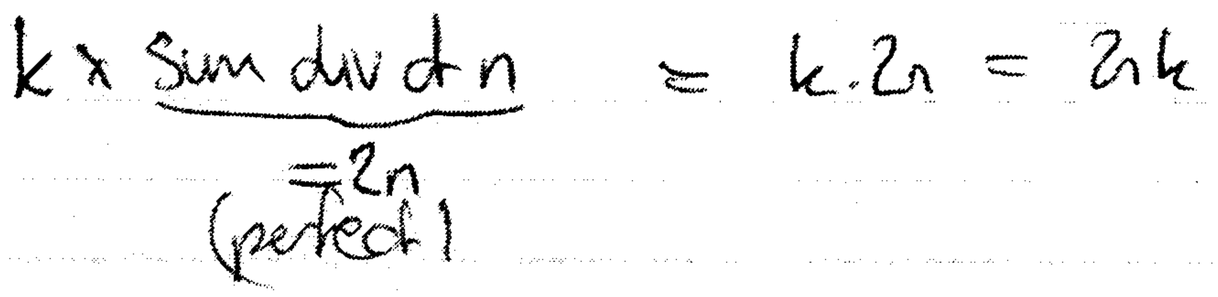
Some participants' notations impacted on their ability to complete the task successfully. For example, Fred's notation naturally led him to the lemma $\sigma(nk) \geq \sigma(n) \cdot k$ (see Figure 8.2). However, participants who had adopted notation which did not reveal this relationship as naturally tended to struggle. For example, Harry used a strange notation where each divisor of n was represented by a dot. This notation, shown in Figure 8.2, seemed to hinder Harry's progress towards a solution he was happy with. The impact that the quality of notation has on success in mathematical research appears to be an area ripe for future research. Perhaps a methodology similar to that adopted here for different research questions could be used to study this question empirically. However, the goal of this thesis is not to study the role of notation, rather to characterise how successful mathematicians reason with conditional statements. This process begins by considering the role of preconscious heuristics in Experiment 4.

8.5 Evidence of preconscious heuristics at work.

The heuristic-analytic dual process theory suggests two stages to reasoning. The first preconscious stage has the role of directing attention to appropriate parts of the environment. As seen in Chapters 6 and 7, two main heuristics have been proposed that are important when participants consider conditional statements as part of the Selection Task: the if- and matching-heuristics. In this section the role of these heuristics in shaping participants responses to the Abundant number task is considered.

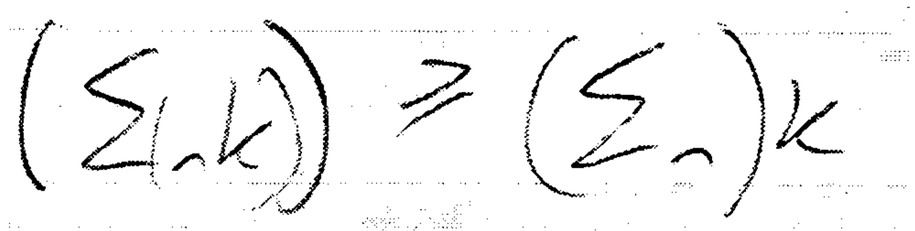


David's notation for $\sigma(n) < 2n$

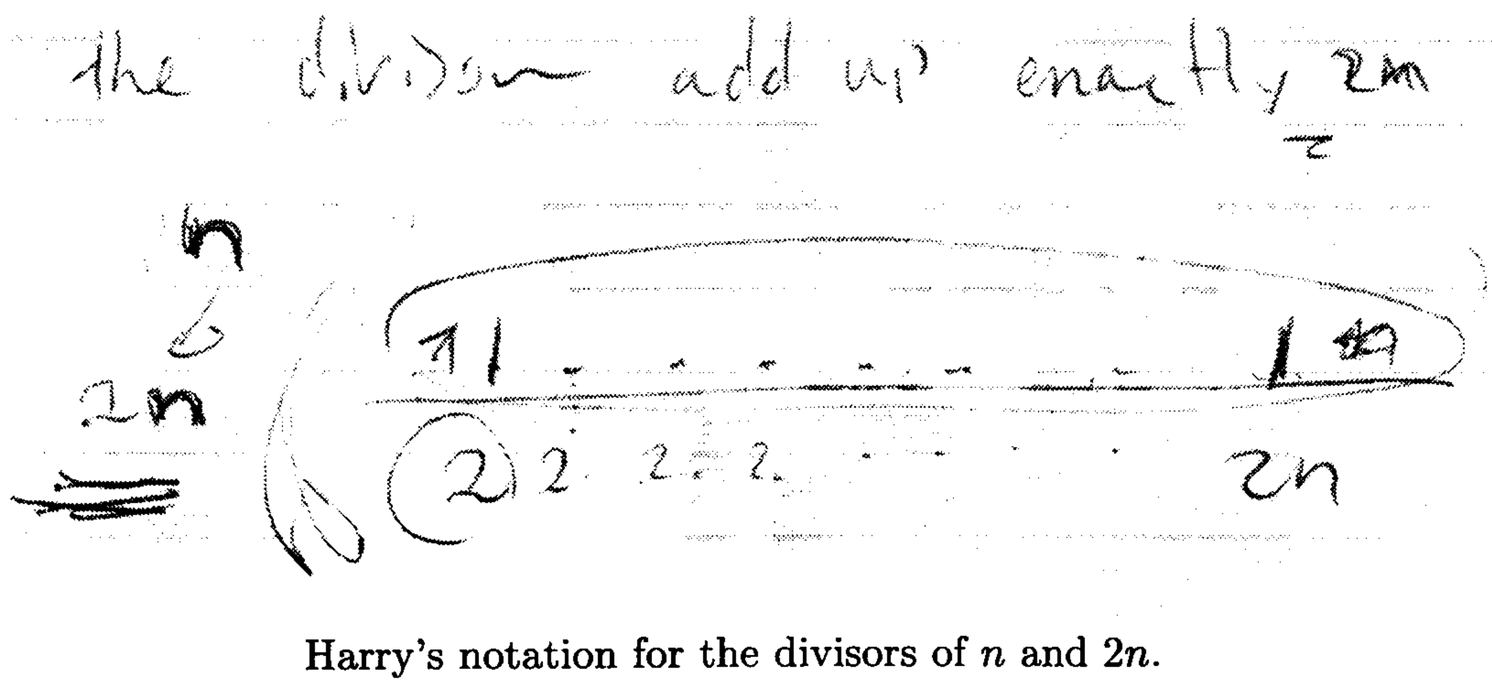


Ben's notation for $k \times \sigma(n) = 2nk$.

Fred's notation for $\sigma(n) = 2n$.



Fred's notation for the lemma $\sigma(nk) \geq \sigma(n) \cdot k$.



Harry's notation for the divisors of n and $2n$.

Figure 8.2: Various notations adopted by the participants in Experiment 4.

8.5.1 The if-heuristic.

Recall that Evans and Over (2004) postulated that the if-heuristic directs attention towards the antecedent part of a conditional statement (the P in the rule 'if P then Q').

Evidence from the abundant number study suggested that participants did indeed tend to initially concentrate on determining the meaning of the antecedent before moving on to try and prove the statement. For example, consider Andrew's reaction to Conjecture 2:

> ANDREW: [*Reads question*] OK, so if n is perfect, then kn is abundant, for any k. OK, so what does it, yeah it looks, so what does it mean? Yeah so if n is perfect, and I take any p_i which divides this n, then afterwards the sum of these p_i's is $2n$. This is the definition. (A2)

Andrew's focus is immediately on the semantic meaning of the antecedent, it is only then that he moves on to consider the consequent, and a possible proof. This behaviour was widespread throughout the study. Ben's reaction to the same conjecture was similar:

> BEN: [*Reads question*] OK, so n being perfect means that you've got the sum of the divisors is exactly n. (B2)

However the more interesting examples of the if-heuristic at work came from Conjectures 4 and 6. Recall that these two conjectures were related to Conjecture 2:

> **Conjecture** (2). *If n is perfect, then kn is abundant for any $k \in \mathbb{N}$.*
>
> **Conjecture** (4). *If n is deficient, then every divisor of n is deficient.*
>
> **Conjecture** (6). *If n and m are abundant, then nm is abundant.*

Conjecture 4 is (almost[5]) the contrapositive of Conjecture 2, and Conjecture 6 is a trivial consequence of Conjecture 2. However, the if-heuristic predicts that participants' attention would be directed away from these relevant factors.

Evidence of the if-heuristic from Conjecture 4.

Having successfully proved Conjectures 1, 2 and 3, David began tackling Conjecture 4 in the following manner:

[5] As discussed earlier, although the two statements are not exact contrapositives, the standard proof of Conjecture 2 also proves the exact contrapositive of Conjecture 4.

DAVID: If n is deficient then every divisor of n is deficient, so the sum of the divisors is, right, err, sum of m in divisors of n, [*David had invented the notation div(n) to describe the set of divisors of n, so here he actually wrote* $\sum_{m \in div(n)} m$] right, let's just write this down as less than or equal to $2n$. And if you've got, err, so if you consider, hmm. Umm, if you consider a divisor of n, itself then you've got, erm... [*long pause*]

INTERVIEWER: What are you thinking?

DAVID: What am I thinking?! I'm not, I'm not sure. I'm trying to take away a term from that, but you can't just take away one term because you might end up with too many. OK, so let's, a good way to do this is sort of fundamental theorem of arithmetic... (D4)

David then proceeded to attempt a proof by decomposing n into its prime factors. After spending a while attempting to find a sufficiently large lower bound on the sum of the divisors of n, David abandoned his initial attempt at a proof and started again, this time trying to formulate an induction argument based on the fact that the product of primes is deficient:

DAVID: Yeah, so if n is deficient then every, ok, so let's think primes again, primes are very deficient, what was that previous one? if p_1 and p_2 are primes then $p_1 p_2$ is erm, not abundant. So, it's either err, so we write our n as a product of primes, p_1 to p_s, and each of these are, each of these are [*very long pause*] err, oh hang on, that [*Conjecture 3*] only says $p_1 p_2$ doesn't it? That doesn't help, it won't give us an induction.

INTERVIEWER: Oh, I see, does that extend to 3 then? If you put a p_3 in there?

DAVID: Well, it wouldn't necessarily extend, I mean look at $2 \times 2 \times 3$, it's 12 which is abundant. (D4)

David's attention here is clearly drawn towards the antecedent of the statement. He first writes down what "n is deficient" means, and then attempts two different methods of direct proof. Despite having proved a logically (near) identical statement ten minutes or so previously, all David's attempts are based on constructing a new direct proof. This is exactly the kind of behaviour that the theory behind the if-heuristic would predict.

After his second failure at a direct proof David re-evaluates, and considers an indirect approach:

DAVID: I'm just trying to think about going the other way, rather than starting with n start with some divisor of n, and then see if we can say, OK, so I guess what I'm trying to do is the contrapositive.

So I'm saying, umm, suppose every divisor of n is not deficient, so suppose this is not deficient and then add a prime and see if you get something that's still not deficient? Is that right? Is that what I'm trying to say? If every divisor of n is not deficient then n is not deficient. Umm.

INTERVIEWER: So what's the contrapositive?

DAVID: Yeah. That's what I've just said. If every divisor of n is not deficient then n is not deficient. So, this is umm, p_1 to, why is this taking me so long? (D4)

Even having formulated the contrapositive of Conjecture 4 it's structural similarity to Conjecture 2 remains hidden to David. The surface linguistic features of the two statements appear to direct attention to two different locations: Conjecture 2 is about abundantness and Conjecture 4 is about deficientness (even when reformulated into a conjecture about non-deficientness – see the discussion of the role of the matching-heuristic below).

After David had successfully proved the statement indirectly (he had used a slightly different method to his proof of Conjecture 2), the interviewer asked him about the connections between Conjectures 2 and 4:

INTERVIEWER: So what's the relationship between that one, number 4, and the one you proved a minute ago [*points at number 2*]? If at all?

DAVID: Umm, if n is perfect then kn is abundant. Errm, oh yeah, so it's the same thing isn't it? [*laughs*] Gosh! It's the contrapositive, well, hang on, it's not quite the contrapositive is it? Because if n is perfect, it doesn't mean, erm, so here we're not assuming perfect, we're assuming not deficient, so it's not quite the same, and we're not quite proving the same thing, because we're proving that kn is not deficient.

INTERVIEWER: Right, so it's nearly the contrapositive?

DAVID: So, it's sort of, it's similar to the contrapositive, but it's not quite.

INTERVIEWER: Yeah, because you've come up with two entirely different proofs which is quite interesting.

DAVID: Yeah gosh, I have, haven't I? [*laughs*] (D4)

When asked directly about the connections between the two conjectures David immediately sees that they are closely related, and is amused that he hadn't noticed. It is when the interviewer forced David to attend to the two conjectures that he sees the similarities. Previously his attention had been directed elsewhere because of the influence of the if-heuristic and the surface linguistic relevance features of the statements.

It is interesting to note David's amusement when he notices the connection. His reaction is somehow similar to how some people react when encountering optical illusions: many people can be amused by *what they see as* the irrationality of their behaviour. (It is important to emphasise the "what *they* see as" so as to avoid any unwelcome distractions from the debates surrounding rationality discussed in §4.5). David's amusement at this situation was not at all atypical, and other examples can be found throughout the extracts reported in this chapter.

David's behaviour when investigating Conjecture 4 was typical. To give another example, here is how Andrew reacted:

ANDREW: If n is deficient then every divisor of n is deficient. Err... [*long pause*]

INTERVIEWER: What are you thinking?

ANDREW: Actually I'm thinking if it's OK or not. But actually I can't, I don't know at the moment, so let's make some investigations. Err, if n is deficient [*long pause*] OK, so OK, let these [*writes* $p_1 \ldots p_k$] be the divisors of n. (A4)

Notice how Andrew's initial reaction is to attempt to decide whether he thinks the statement is "OK or not", this behaviour is discussed in greater depth in section §8.6. After attempting, and failing, to find a direct proof by considering the set of divisors of a typical divisor of n, Andrew asked to see the examples. Having convinced himself of the statement's truth by evaluating a few examples he returned to attempting to prove the statement directly. Eventually, after several further unsuccessful attempts, Andrew looked back through the conjectures he had already worked on and noticed that Conjecture 2 was relevant. As with David, he was highly amused by this finding:

ANDREW: Wow! A number is abundant if and only if it is a multiple of 6. But I think that this wasn't true. If n is perfect then kn is abundant. Oh! Oh! [*collapses laughing*]

INTERVIEWER: What are you whooping at?! What does this mean?!

ANDREW: So actually it is done. Right. (A4)

Two[6] participants fairly quickly realised that they needed to try an indirect proof, but still did not notice the connection with Conjecture 2. Fred's initial reaction was similar to David's: to try to consider n as a product of primes.

[6]Note that although the number of participants in the study who answered in each fashion are included here for interest purposes, no claim of statistical significance is being made. As discussed in Chapter 5, the quasi-judicial method of qualitative data analysis seeks validity and reliability through the cogency of the theoretical reasoning conducted by the analyst, not through the representativeness of the cases.

However, he quickly realised that this wouldn't work, and instead turned his attention to the contrapositive.

FRED: If n is deficient then every divisor of n is deficient. Hmmm. Now I'm going to be a little bit wily, because I think, because going by the last question, it would probably help me to look at the product of two primes. [*long pause*] Umm, no that's not going to help me, because there might be other deficient numbers that aren't the product of two primes... Let me think... I would go towards this... Umm... I think the way to do it is to find... hmm... is the term contrapositive? Is it the contrapositive I'm looking for?... going the other way... there exists a non deficient divisor of n implies n not deficient. Now is that the contrapositive?

INTERVIEWER: What are you trying to do here?

FRED: Err, going through that, I mean, my initial thought is that, I did have it in my mind, but I'm trying to think again now... if n is... [*long pause*]

INTERVIEWER: What are you thinking?

FRED: I did have quite a clear picture of where I was going to go, I'm just trying to get that back... I really do think you've got to go from the opposite direction...

INTERVIEWER: Why?

FRED: Because, because, given a number n, it's tricky to find the divisors, but given a divisor its a lot easier to find n. Like, it's, yeah, it seems a lot less tricky to work up rather than to work down. (F4)

There are several things worth noting from this transcript. Firstly, it is striking that Fred is not at all confident with the term "contrapositive"; although he is perfectly able to formulate the contrapositive correctly, he seems more comfortable referring to it as "going the other way". This type of behaviour was not at all unusual. Ben, for example, consistently and incorrectly used the term "converse" to mean "contrapositive". Andrew preferred to talk only about "contradiction arguments", and Edward used "the opposite of what this is saying" to refer to the contrapositive. It seems reasonable to conclude that the contrapositive, at least in the formal sense that it is taught in logic courses, was not a prominent part of these (highly talented and successful) research students' concept images of implication.

Secondly, although Fred believes that the contrapositive may be of help to him, it is not because he realises that he has already proved it. Instead he notes that it is easier to find multiples of numbers than it is to find divisors of numbers. He uses his strategic knowledge of number theory to help formulate

his proof attempt (Weber, 2001). But he does not immediately recognise that the proof he has already completed will work.

Thirdly, this excerpt indicates that the if-heuristic may not be all-pervasive. Fred's attention, whilst initially directed at the antecedent, soon is redirected into trying an indirect proof. Perhaps Fred was not overly influenced by the if-heuristic, or perhaps his System 2 processes soon overrode any influence that it did have. In either case it seems, contrary to the experiences of David and Andrew discussed above, Conjecture 4 was not a Criterion T problem for Fred. Of the all the interviewees, only two – Fred and Edward – considered the contrapositive of the statement early on in their attempted proofs, and neither apparently decided to do this on account of having already proved a near-identical statement.

After Fred successfully completed his proof of Conjecture 4 the interviewer asked him about connections between Conjectures 2 and 4.

INTERVIEWER: What's the difference between number 2 and number 4? Did you just re-prove number 2?

FRED: Hmmm. [*long pause*] No I haven't, because Conjecture 2... talks specifically about n being perfect, it's going to be a bit confusing to describe this because you're using n to describe the sort of big multiple number in Conjecture 4, but in the first one you're saying the big multiple number... erm, is based on a perfect number, but in Conjecture 4 you're saying the big multiple number is built on, err... well the contrapositive statement which I proved is saying that the big number is based on a perfect number or an abundant number.

INTERVIEWER: So if we changed that to be, in number 2, to be either a perfect or an abundant number it would be exactly the same?

FRED: Umm, let's have a look at my notes... would the prove I have worked? [*Reads through his proof of Conjecture 2*] Yeah, the proof I have would work. [...] I'm not sure why it is the same. Umm, I have to say I can't see much similarity between the statements, it's not jumping out at me that the statements are very similar. (F4)

Even after Fred agrees that the two proofs he has produced are essentially identical, he still feels that the statements do not "seem" that similar. Of course, the reason why may be that the surface linguistic features of the problem bias Fred's attention towards deficiency in Conjecture 4 and abundantness in Conjecture 2.

In summary, evidence was found that Conjecture 4 was a Criterion T situation for most of the participants. The if-heuristic appeared to direct attention towards the antecedent of the rule, which, because of linguistic surface relevance

features of the statement, biased participants away from noticing that they had already proved a near-identical statement. In fact even the two participants who *did* quickly adopt an indirect strategy did not notice the connections with Conjecture 2. This, whilst not a consequence of the if-heuristic, could still be seen as being a consequence of preconscious System 1 relevance heuristics: because of the linguistic structure of the two statements, Conjectures 2 and 4 were seen as being "about" different things. One was concerned with abundantness and one with deficiency.

Evidence of the if-heuristic from Conjecture 6.

Recall that Conjecture 6 asked participants to prove that the product of two abundant numbers was also abundant. This is a trivial consequence of the fact that kn is abundant for abundant n and natural k. However, it was predicted that the if-heuristic would complicate matters by directing attention towards the antecedent of the statement, that is to say the situation where *both* m and n are abundant. There was clear evidence of this from the interviews. Here, for example, is Chris' reaction to being given the statement:

CHRIS: Right, so if n and m are abundant then nm is abundant. That looks more plausible [*than Conjecture 5*], cos they're going to share factors. Anything that divides n...in fact, I mean it might be a quite straightforward proof because the set of factors of n, these same a_i's I've been using [*Chris was using the notation a_i to denote divisors of n*], so we know the sum of the a_i's are strictly greater than $2n$, if I let the b_i's be the divisors of m. So certainly the a, the a_i's and the b_i's are also factors of nm, so the sum of all the factors of nm is going to be greater than or equal to, the sum of the a_i's the sum of the b_i's. (C6)

Chris went on to try to prove the statement by considering the 'cross product' of $\{a_i\}$ with $\{b_i\}$, but realised that this method would fail on account of having repeated divisors in the final sum. The interview ran out of time before Chris completed his solution.

Edward also failed to find a solution during the course of the interview, but asked the interviewer what the correct solution was. When the interviewer pointed out the connection between Conjectures 2 and 4 Edward was highly amused:

EDWARD: Oh yeah [*laughs*]. [...] Oh yeah, of course. That is funny [*laughs*]. That's so blatantly obviously the same as that! [*laughs*]. (E6)

David adopted a similar strategy to proving Conjecture 6 as Chris and Edward:

DAVID: Oh, so n and m are...oh, hang on, so you've got sums here as well, so I've got my two assumptions down there, [*David used the same assumptions that he'd written down for Conjecture 5*] and then umm, I want to say the divisors of nm, so what I do know is the sum k, maybe I should use a different k here, multiplied by k', so k is in the divisors of m, umm, k is in divisors of n, so this is actually greater than $4mn$. (D6)

David continued down the same path as Chris before realising that he too will end up with too many repeated divisors in his sum:

DAVID: Actually, that might not be right [*scribbles everything out*], because you might get terms more than once here, if you multiply these two sums together, errm, umm, you see, [*long pause*]. (D6)

David attempted to rescue the situation by arguing that the divisors will only ever be repeated twice, but eventually rejected this approach as not feasible. Having seen both his approaches so far fail, David reviewed the previous conjectures and notices the connection with Conjecture 2, again he was amused by his failure not to notice this earlier:

DAVID: Umm, I think, let's have a look at what we had before. So we had, these two corollaries 4 and 2 wasn't it, let's see if we can use any of those. If n and m are abundant, so if n's abundant then kn's got to be abundant, I mean if n's perfect then kn is abundant, so mn's got to be abundant, I mean, yeah [*laughs*] I've realised it's kind of a trivial consequence of this [*Conjecture 2*], I mean you can do the same proof as in here can't you? I mean, if you, except you start off with greater than or equal to, $\geq 2n$ and then you prove that, so you prove something different, something weaker than this, well, not logically weaker but, I mean it's got to be, if n is perfect then kn is abundant, well if n's abundant already then...

INTERVIEWER: So if you're making it more abundant?

DAVID: Yeah, yeah, yeah. And so, I mean, you'd need a new proof, you'd need to start off with an inequality there and you'd still have inequalities. (D6)

Fred's interview proceeded in a similar fashion. Having attempted to cross multiply the two sets of divisors he got stuck. Whilst stuck he suddenly noticed the connection with Conjectures 2 and 4:

FRED: Umm, [*long pause*] just because... [*long pause*] oh actually, I think I've got it. Oh, simple, it's much simple, simpler than that! Can I start again? [*Fred seemed to be half annoyed with himself and half amused*]

INTERVIEWER: Yeah, why do you say that?

FRED: The reason is, because from I think, conjecture 4, I proved there, that if a number is abundant you can multiply it by any number... and you get an abundant number. So we're just being a bit more specific about which numbers we can multiply by... [*long pause*]

INTERVIEWER: Right.

FRED: That's it.

INTERVIEWER: Is that it? Is that the end of it?

FRED: Yeah, that's it. That's the end of it. You're multiplying an abundant number by *a* number. And then as we showed in part 4...

(F6)

As with Conjecture 4, however, not every participant found Conjecture 6 to be a Criterion T situation. Two participants – Andrew and Gary – realised immediately that the conjecture was a trivial consequence of their earlier work:

ANDREW: [*Reads Conjecture 6*] Abundant, oh, this is already proven. This is Conjecture 1.

INTERVIEWER: So that's the end of that then? You look surprised. You are entirely happy that that's true?

ANDREW: Yeah [*laughs*]

INTERVIEWER: Why are you laughing?!

ANDREW: No, no, because it seemed that the conjectures are getting tougher and tougher. I was getting incredibly scared! [*laughs*]

(A6)

However, Andrew's quick dismissal of Conjecture 6 could perhaps be because of an earlier experience of a similar conjecture. During his work on Conjecture 4, Andrew had attempted to prove that the product of two perfect numbers couldn't be deficient, despite already having established that any proper multiple of a perfect number was abundant:

ANDREW: So the question is, if we've got one perfect number, a second perfect number and we multiply them, can we get a deficient number? So this is the only thing. OK. [...] OK, so let's dig into these summations, so if I sum this, err... (A4)

Andrew went on trying to prove that the product of two perfects cannot be deficient for some time before noticing that he had already proven a stronger

statement. So Andrew had already encountered a Criterion T situation of similar form to Conjecture 6 during his solution to Conjecture 4, which he had spent some considerable time solving.

Gary also quickly noticed that he had already proved Conjecture 6:

GARY: If n and m are abundant then n times m is abundant. That sounds right.

INTERVIEWER: Why?

GARY: Because multiplication of abundants tend to get even more abundant. I think we have already proved that. I claimed that any multiple of abundant is abundant, so it would be true. Now let's just go back to that little piece to see if it's true, it should be true. (G6)

Interestingly, Gary was the only person in the entire sample who appeared to be unaffected by the if-heuristic on Conjecture 6, or, in the case of Andrew, a similar conjecture earlier.

It is worth pointing out that, because of time constraints[7], Gary was not asked to tackle Conjecture 5 (n, m abundant $\Rightarrow nm$ abundant), the conjecture that immediately proceeded Conjecture 6 for the other participants. Gary, in contrast, had been tackling a structurally very similar statement to Conjecture 6 (Conjecture 4) immediately before. Although, it should be noted that this cannot explain Gary's behaviour in its entirety, as Andrew had also been looking at a similar statement prior to the excerpt from Conjecture 4 quoted above. If meeting Conjecture 5 was a prerequisite for the Criterion T status of Conjecture 6 then one would not have expected Andrew's behaviour in Conjecture 4.

The if-heuristic: A summary of the evidence.

The evidence from Conjectures 4 and 6 supports the notion that the if-heuristic has an important role to play in mathematical reasoning. It was predicted that the two conjectures would satisfy Criterion T, and for the majority of participants this did indeed appear to be the case. That is to say, the structure of the statements biased participants' attention towards unhelpful parts of the problems. This attentional bias was so strong that even though participants had already proved a logically near-identical statement (Conjecture 2), both Conjectures 4 and 6 proved difficult to solve for most participants.

It is also of interest to consider the situations in which participants overcame this bias (if they did). Participants tended to 'get past' the bias after having tried several attempts and got stuck. As part of a 're-evaluation' of the situation,

[7]Gary had taken approximately 45 minutes on Conjectures 1–4.

which often involved explicitly looking back at previous conjectures, the solution would reveal itself. Here, for example, is David solving Conjecture 6:

DAVID: Umm, I think, let's have a look at what we had before. So we had, these two corollaries 4 and 2 wasn't it, let's see if we can use any of those. If n and m are abundant, so if n's abundant then kn's got to be abundant, I mean if n's perfect then kn is abundant, so mn's got to be abundant, I mean, I've realised it's kind of a trivial consequence of this, I mean you can do the same proof as in here can't you? (D6)

This pattern, of getting stuck, re-evaluating the global picture, and then noticing the links to previous conjectures was widespread across the different interviews. Paradoxically, then, mathematicians of higher abilities – those who are less likely to get stuck – may be more likely to have the structure of their final proofs influenced by their preconscious heuristics.

It was also striking at how often participants were amused at their inability to immediately see the relations between the conjectures. Here Andrew recognised the relevance of his proof of Conjecture 1 to a sub-conjecture he made during his solution to Conjecture 4:

ANDREW: OK, so how to prove it. [*long pause, taps on table*] OK, so we've got some number x, and it's deficient. [*long pause*] Well, I mean, it's really very simple [*laughs*] it's really very simple [*laughs*]. Because, yeah, because of this conjecture. (A4)

In summary, evidence was found that is consistent with the hypothesis that the if-heuristic plays a role in mathematical reasoning. In the next section the role of the matching-heuristic is considered.

8.5.2 The matching-heuristic.

Some indication of the influence of the matching-heuristic has already been presented in the discussions above. For example, recall that when Fred was asked about the connections between Conjectures 2 and 4 he remarked that, even though he realised the two proofs he had produced were essentially identical, the statements didn't seem to be that similar (see the dialogue on page 149).

Recall that the matching-heuristic directs attention towards the semantic content of the antecedent and consequent, regardless of the presence of negatives in the rule. Fred's remark that the connections between the statements didn't "jump out" at him could be interpreted as a consequence of the matching-heuristic. For Conjecture 2 this heuristic directs attention towards abundancy

and perfectness; for Conjecture 4 attention is directed towards deficiency. In attentional terms, the two statements are "about" different things, even though logically they are almost identical.

However, more evidence was found for the role of the matching-heuristic with regards to what proof strategies participants adopted for the differing conjectures. The last six conjectures were of the form "if P, then Q", but the immediate strategies adopted by participants varied considerably depending on the presence of a negation in the consequent (i.e. whether Q was written "not R" for some R). Consider Andrew's response to Conjecture 7 ("if n is abundant, then n is not of the form p^m for some natural m and prime p"):

ANDREW: OK, so if n is abundant then n is not of the form p, hmm, hmm. Yeah, well this shouldn't be probably too hard. Because if we take p^m then what are the divisors? Yeah, the divisors are basically $1+p+p^2+$ blah blah blah $+p^m$. And actually this is just a geometric series right? And the sum of this is, I don't know, something like this... (A7)

Andrew immediately starts to consider the negation of the consequent of Conjecture 7, He thinks about the case when n *is* of the form p^m, and starts to deduce consequences from it. Eventually he successfully proves that n cannot be abundant. In this case Andrew's instantaneous reaction was to start a contrapositive proof (although, as we have seen, he did not use this language). Compare his reaction to Conjecture 7 to Conjecture 4 discussed earlier. Although, like Conjecture 7, he ended up using an indirect argument to establish the result, it took him substantial amounts of time to adopt this approach. In Conjecture 7 this was an instantaneous response. David behaved similarly to Andrew:

DAVID: OK, so if n is abundant then n is not of the form, err, so in other words, it's the contrapositive again, if n is of the form p^m then it's, then n is not abundant, so we know this, err, right, so what's the sum of k in p^m? k, err, whatever... (D7)

Again, as with Andrew, David went on to prove that n couldn't be abundant. The contrast between this behaviour and David's approach to Conjecture 4 is striking.

Ben offered an explanation as to why he immediately thought of an indirect proof with Conjecture 7.

BEN: [*reads question*] OK, so, my original idea would be to try and prove the converse, [*Ben persistently used the term 'converse' to mean 'contrapositive'*] because it's "not" which is just nasty. So assume n is p^m. Then you

> want to know about the sum of the divisors of n. Umm, the sum of the divisors of n is just [*long pause*] is just the sum of $1+p+p^2$ up to p^m. Which there's a formula for if I could remember it... (B7)

Ben is absolutely correct to identity the key factor here is the "not" in the statement. Whereas in Conjecture 4 the antecedent was perceived to be about deficiency, in Conjecture 7 the antecedent is, despite the presence of the negative, perceived to be about the form p^m. Consequently the matching heuristic directs attention towards numbers of this form, which in turn leads to an indirect proof based on a contrapositive argument. It should be noted that in this instance the matching heuristic is helpful – by considering this part of the statement a correct proof is more likely to be found. This conjecture does not satisfy Criterion T, indeed the exact opposite: perhaps such a situation could be said to satisfy Criterion U?

In short, the data from Experiment 4 appears to support the idea that the matching-heuristic plays an important role in mathematical reasoning.

8.5.3 Summary of §8.5.

The aim of the quasi-judicial method of multiple case study analysis is to impose pre-existing theory onto qualitative data. The theory used in this section was the heuristic-analytical dual process account of reasoning described, examined and tested in §4.4.7 and Chapters 6 and 7. The empirical evidence collected in Experiment 4 provides strong support for the idea that preconscious heuristics play an important role in directing attention during mathematical reasoning. Two particular heuristics were examined:

- The if-heuristic directs attention towards the situation where the antecedent of an "if P then Q" statement is true. It was found that many empirical observations from across the study fitted with predictions derived from the postulated role of the if-heuristic.

- The matching-heuristic directs attention towards the surface linguistic content of the antecedent and consequent parts of an "if P then Q" statement, regardless of the presence of negatives in that statement. Support was found for the idea that this heuristic plays an important role in mathematical reasoning.

In short then, the evidence suggests that preconscious heuristics have an important job in normal mathematical reasoning: they direct attention towards apparently relevant parts of mathematical statements. Normally these heuristics are useful, they prevent wasteful expenditure of System 2 resources by

biasing attention towards helpful parts of the problem (such situations satisfy Criterion U). However, on some occasions these heuristics are not so helpful; by biasing attention away from pertinent parts of the problem they may detract from an individual's overall ability to solve the task to their own satisfaction (such situations satisfy Criterion T).

In the next section further analyses from Experiment 4 are reported. But whereas this section concentrated on the role of preconscious System 1 heuristics in reasoning, the next considers the processes that underlie conscious System 2 reasoning and argumentation about conditionals: Once attention has been directed towards certain parts of the these statements, how do participants go about evaluating and whether the statements are true or false, and how do they go about proving it?

8.6 Evaluating conditionals.

Recall that the second part of the research question which Experiment 4 was designed to investigate was: what kind of System 2 processes do mathematicians use when working with, and judging the validity, conditional statements? This section attempts to explore this question by representing the arguments that were offered by participants during the interviews. As suggested by Bromley (1986), heavy use will be made of Toulmin's (1958) argumentation scheme. Toulmin's notion of informal logic was briefly discussed in Chapter 3, and is fully reviewed here.

8.6.1 Toulmin's informal logic.

Toulmin (1958) advocated an approach to analysing arguments that radically departed from both traditional and modern approaches to formal logic. Toulmin was less concerned with the logical validity of an argument, and more worried about the semantic content and structure in which it fits. This manner of analysing argumentation has become known as 'informal logic' in order to emphasise its differences from formal logic.

Toulmin's (1958) scheme has six basic types of statement, each of which plays a different role (sometimes a role that is not explicitly verbalised) in an argument:

Conclusion (C) The statement which the arguer wishes to convince their audience of.

Data (D) The foundations on which the argument is based, the relevant evidence for the claim.

Warrant (W) Justifies the connection between data (D) and conclusion (C) by, for example, appealing to a rule, a definition or by making an analogy.

Backing (B) Supports the warrant (W) by appealing to further evidence.

Modal Qualifier (Q) Qualifies the conclusion (C) by expressing degrees of confidence.

Rebuttal (R) Rebuts the conclusion (C) by stating the conditions under which it would not hold.

Toulmin's (1958) use of the word 'warrant' is not identical to how the term has been used by some researchers in the mathematics education literature. Rodd (2000), following Plantinga (1993), defined a warrant as being "that which secures knowledge" (p.222). Rodd saw a warrant as guaranteeing the removal of uncertainty, whereas Toulmin was more flexible, he accepted that a warrant can be qualified with a modal qualifier, thereby potentially only reducing uncertainty.

These six components of an argument are linked together in the structure shown in Figure 8.3. Toulmin's structure can be naturally extended to accommodate longer and more complicated arguments, as shown in Figure 8.4.

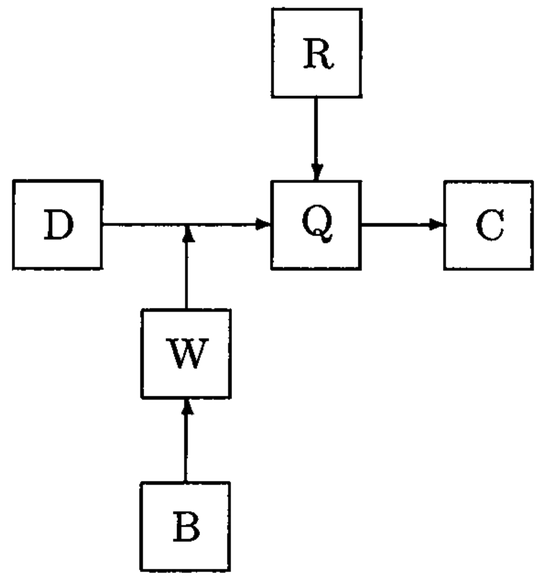


Figure 8.3: Toulmin's model of a general argument.

Somewhat surprisingly, Toulmin's (1958) scheme has been used infrequently for analysing advanced mathematical thought. A notable exception is the philosopher Aberdein who has used to the scheme to analyse the proofs of the Intermediate Value Theorem (Aberdein, 2005) and the Four Colour Theorem (Aberdein, 2006). Aberdein's use of Toulmin's work is interesting, as he only uses it to describe *formal* mathematical arguments. In his original work Toulmin believed that formal mathematics was one of the few domains of explanation where formal logic – the system he was reacting against – adequately described argumentation structures. However, in a later work Toulmin et al. (1984) gave

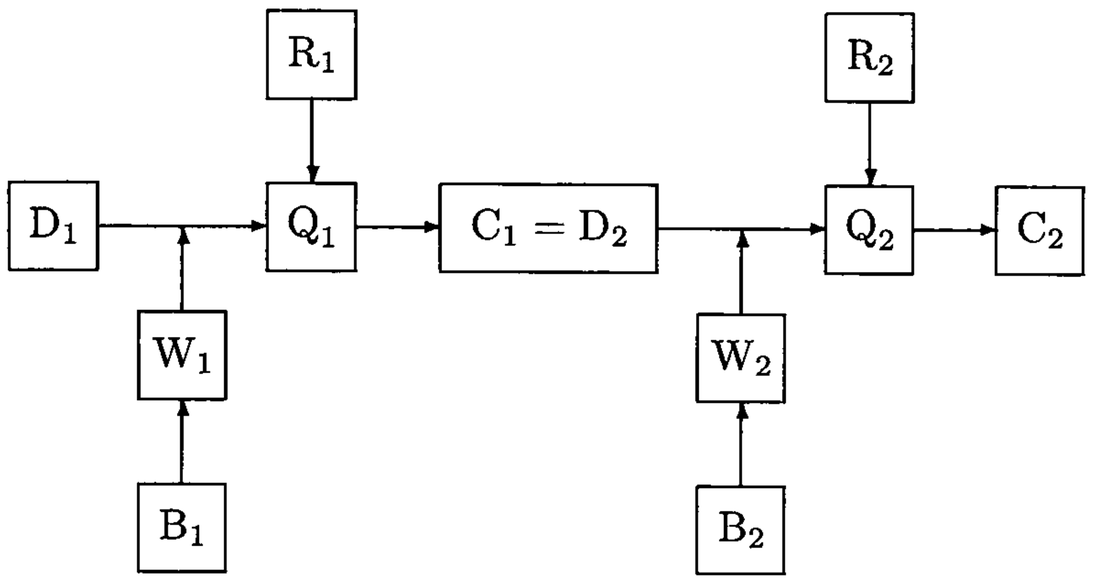


Figure 8.4: A long argument represented in Toulmin's scheme.

an example of a formal mathematical argument expressed in the argumentation structure. Alcolea Banegas (1998) gave further examples of applying Toulmin's structure to mathematics: he looked at a meta-mathematical argument which justified the adoption of the Axiom of Choice. It is on these examples that Aberdein's work builds.

From a psychological perspective, it seems unclear what the purpose of describing formal mathematics using informal logic is. In this specific context many parts of Toulmin's (1958) scheme appear to be trivial or redundant. For example, when Aberdein (2005) analysed the proof of the theorem that asserts there are only five platonic solids the backing he used was "given the axioms, postulates, and definitions of three-dimensional Euclidean geometry"; the modal qualifiers he used was "with strict geometric necessity", "classically" and "constructively"; and the rebuttal he used was "no rebuttals or exceptions". It is clear that these will be effectively identical for all formal mathematical proofs.

Several researchers have attempted to solve this problem by simply omitting the Backing, Modal Qualifier and Rebuttal from Toulmin's (1958) model when applying it to mathematics. Krummheuer (1995), for example, adopted such an approach when using informal logic to describe classroom based mathematical arguments. Many other researchers have followed this approach, and it appears to have become the default position in mathematics education research that has used Toulmin's work. For example, such a position has been taken by researchers studying basic number skills (Evens & Houssart, 2004), logical deduction (Hauk, 2005; Hoyles & Küchemann, 2002; Weber & Alcock, 2005), geometry (Knipping, 2003; Pedemonte, 2003), and proof (Yackel, 2001). Indeed, this position appears to have become so entrenched that, in her recent review of research on proof in mathematics education, Mariotti (2006) described Toulmin's scheme as being a "ternary model".

Note that all these mathematics education researchers were not, as Toulmin et al. (1984) and Aberdein (2005) were, attempting to describe *formal* mathematical reasoning. In contrast they were attempting to describe the *informal* reasoning patterns of students in mathematical situations. One of the major arguments made in this section of the thesis is that both these mathematics educators and Aberdein (2005) have not extracted full benefit from Toulmin's (1958) scheme: Using informal logic to describe *formal* mathematics is psychologically unrevealing; but equally, using only a subsection of Toulmin's scheme to describe *informal* logic is unsound – there are many aspects of informal mathematical reasoning that cannot be adequately modelled without the use of all six components of Toulmin's scheme.

When modelling arguments using Toulmin's (1958) scheme it is often the case that certain parts of the argument (most commonly backings and rebuttals) are not explicitly verbalised by the arguer. In line with earlier researchers who have used the scheme, we dealt with this issue by inferring the backings and rebuttals of participants' arguments where they were not explicitly verbalised. Consequently the diagrams reported in the remainder of this chapter represent plausible models which account for participants' behaviour and utterances; they are not direct one-to-one mappings from utterance to argument. However, it should be noted that, in the examples given here, very few sections of the diagrams are not directly related to the participant's actual spoken words. The methodological issues involved in using Toulmin's scheme for modelling empirical data are discussed in depth by, for example, Bromley (1986) and Simosi (2003).

8.6.2 Modal qualifiers in mathematical reasoning.

Previous work that has applied Toulmin's (1958) argumentation scheme to mathematical reasoning has, as discussed above, downplayed the role of the modal qualifier and rebuttal parts of arguments. This section seeks to argue that appreciating the role of the modal qualifier is crucial to fully understanding how mathematicians deal with conditional statements.

After Chris had correctly identified that Conjecture 3 was false the interviewer asked him whether he thought the conjecture would be true if it was modified to read "if p_1, p_2 are prime, then $p_1 p_2$ is not abundant". He tried two examples (2 and 3, 5 and 97) to investigate the situation and then said:

CHRIS: Since the smallest numbers I could find to put in this equation showed it was perfect and in the larger limit it showed, $p_1 p_2$ was deficient. So it's possible it holds for all p_1, p_2.

INTERVIEWER: Do you think it does?

CHRIS: I think it probably does. But I'm not sure why [*laughs*]. Yeah, the fact that this is, in some ways, sort of monotonic. In other words, I know that this statement is true for large p_1, p_2; I know it's true for small p_1, p_2; so I feel therefore that it should be true for p_1, p_2 in the middle. Umm, but I might have to do some work to show that. (C3)

This piece of argumentation is shown graphically in Figure 8.5. Chris says that

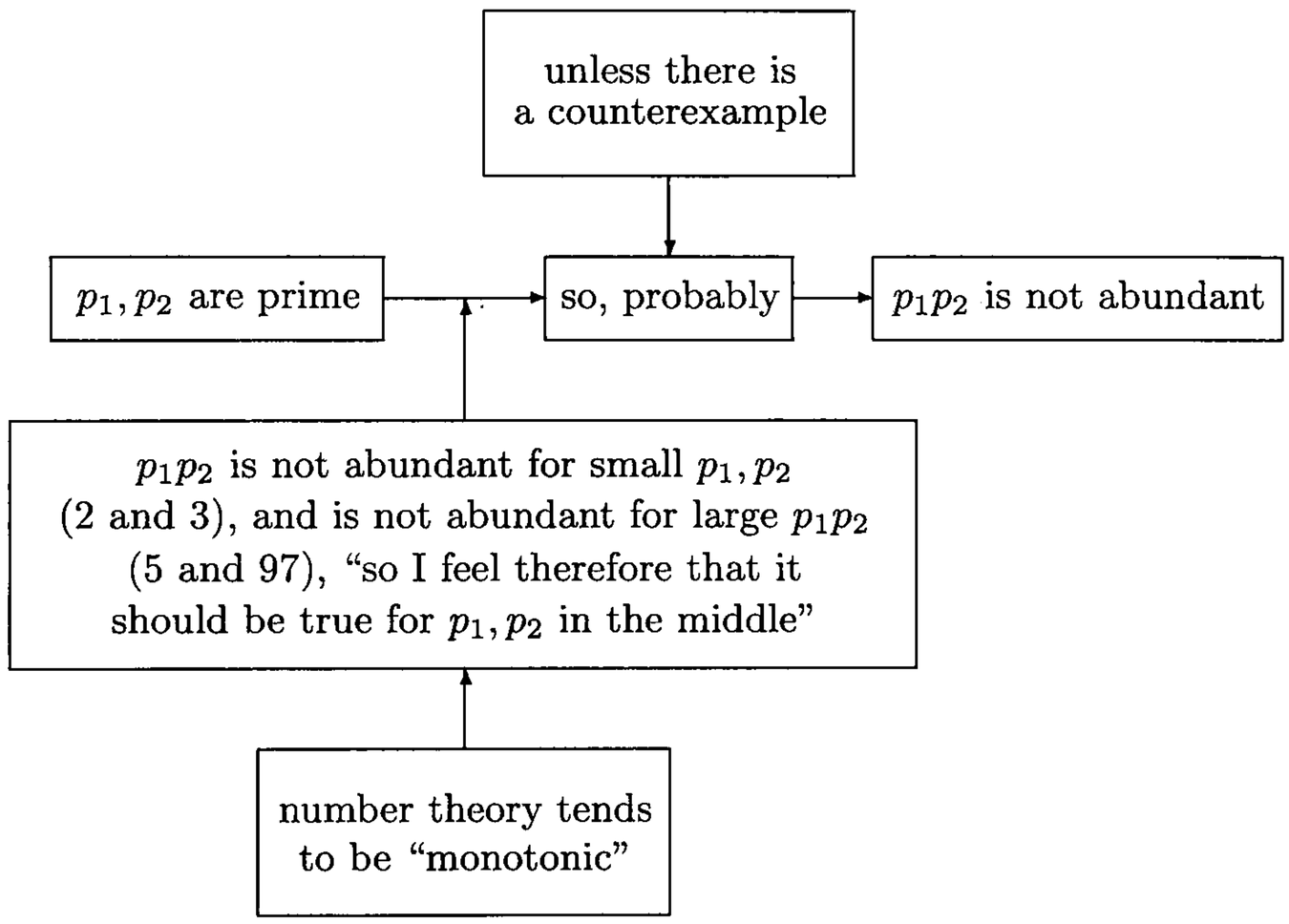


Figure 8.5: Part of Chris's response to Conjecture 3.

he thinks it is "probable" that the statement is true, on account of an argument to do with monotonicity and two examples. He accepts that he has not shown the result formally, but informally, he has persuaded himself that the statement is probably true, and does not feel obliged to carry on and produce a formal proof. In terms of Toulmin's (1958) scheme, his argument revolves around a modal qualifier that does not carry certainty.

In this example Chris appears to be fairly confident that the conclusion can be made, but there are examples of less certain modal qualifiers. Here, for example, David is asked the same question, about the modified version of Conjecture 3:

INTERVIEWER: If I changed it [*Conjecture 3*] then, to be not abundant, what would you say?

DAVID: That would seem more reasonable. Because primes look very deficient.

INTERVIEWER: Do they?

DAVID: Well, they only have 1 and themselves as divisors, so they're about as deficient as you can get right? (D3)

David went on to produce a direct proof of the statement.

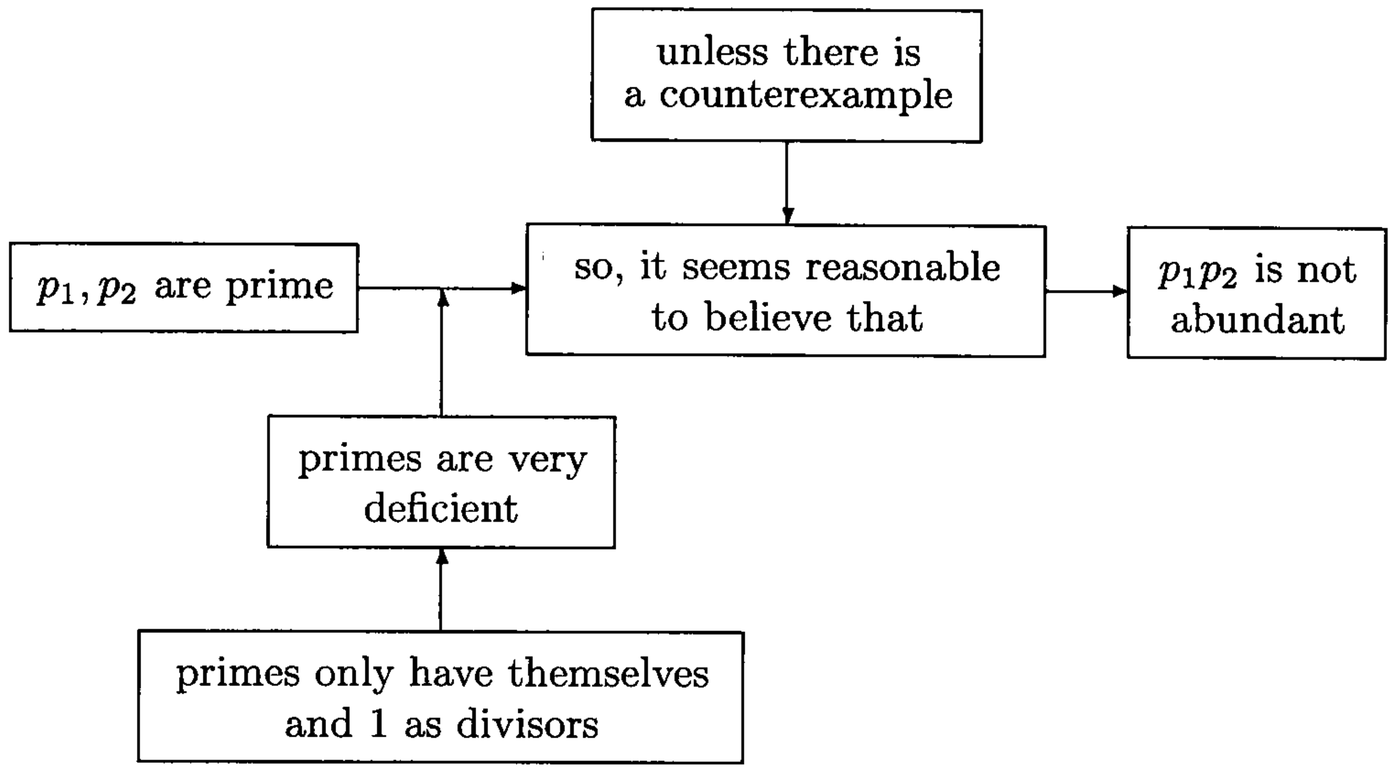


Figure 8.6: Part of David's response to Conjecture 3.

This argument is modelled in Figure 8.6. Compare the different modal qualifiers used by Chris and David. David seems less convinced by his argument than Chris did, and went on to convert this informal piece of reasoning into a formal proof. Chris, on the other hand, was sufficiently convinced of his argument that he didn't feel the need to continue, despite accepting that "some work" would be needed to "show" the result.

The key point here is that Chris and David used different types of warrants in their respective arguments. These different warrants were accompanied by different modal qualifiers, which fixed their degree of belief in the conclusion of the argument. In the next few sections examples of differing types of warrants will be discussed, roughly categorised according to *warrant-types*. To be clear, a *warrant* is a part of a particular argument, however a *warrant-type* is a device with which to categorise individual warrants based on their similarities. Thus, each argument's warrant belongs to a warrant-type. The warrant-types observed fit broadly with Harel and Sowder's (1998) 'proof-schemes' framework (a *prima facie* explanation; see Chapter 2), therefore similar adjectives have been used to describe the warrant-types discussed here.

There is, however, an important difference between the notion of a warrant-type and that of a proof scheme. Harel and Sowder (1998) defines a person's proof scheme as that which helps to "remove her or his own doubts about the truth of an assertion". A proof scheme, then, is about *removing* uncertainty.

Warrants from certain warrant-types, in contrast, may only *reduce* uncertainty. The theoretical relationship between the constructs 'warrant-type' and 'proof scheme' is discussed in full in §8.6.7.

8.6.3 The inductive warrant-type.

Harel and Sowder (1998, p.252) define inductive proof schemes as "when students ascertain for themselves and persuade others about the truth of a conjecture by *quantitatively evaluating* their conjecture in *one or more* specific cases". Many examples of inductive warrants were used by participants in the abundant number study. One example has already been discussed. Figure 8.5 shows an argument offered by Chris during his work on Conjecture 3. Clearly the warrant Chris uses is inductive, he has quantitatively evaluated the conjecture for both small and large numbers, and thus feels that it should be true for all numbers.

In his response to Conjecture 4, having failed in his initial proof attempt, Andrew offered the following argument:

ANDREW: Let's make some experiments [*laughs*]. OK, so the deficient numbers are, for example, 9. 9 is deficient. That's too big because, OK, 10 let's say. We've got 2, 5. Primes are apparently deficient.
INTERVIEWER: Primes are deficient?
ANDREW: Primes are always deficient, yeah, because the sum is equal to the number plus 1. Well, always [*laughs*], no, or is it? no, even 2 is deficient, so it doesn't fail. Yeah, so apparently it works here. Yeah ok, so apparently, it seems to me that it's true.
INTERVIEWER: Why do you say that then? Because it works for 10?
ANDREW: Because [*pause*] because, hmm. [*long pause*] (A4)

At this point, after a long pause, Andrew began another proof attempt that eventually resulted in a correct proof. Andrew's argument in the extract above is modelled in Figure 8.7. Based on one empirical evaluation, Andrew is sufficiently convinced of the conclusion's truth to start a proof attempt. In the language of Balacheff (1988), Andrew conducted a 'crucial experiment' to convince himself of the statement's probable truth.

In a similar example, David constructed a two stage argument to evaluate Conjecture 2. In the first stage he successfully showed that kn was either abundant or perfect if n is perfect (using the same argument deployed during his work on Conjecture 1), and during the second stage he tried to remove the possibility that kn is abundant:

DAVID: Why would it [*the equation* $\sigma(kn) \geq 2kn$] be a greater than? Umm, I don't know, why couldn't it be perfect? I mean you've got some

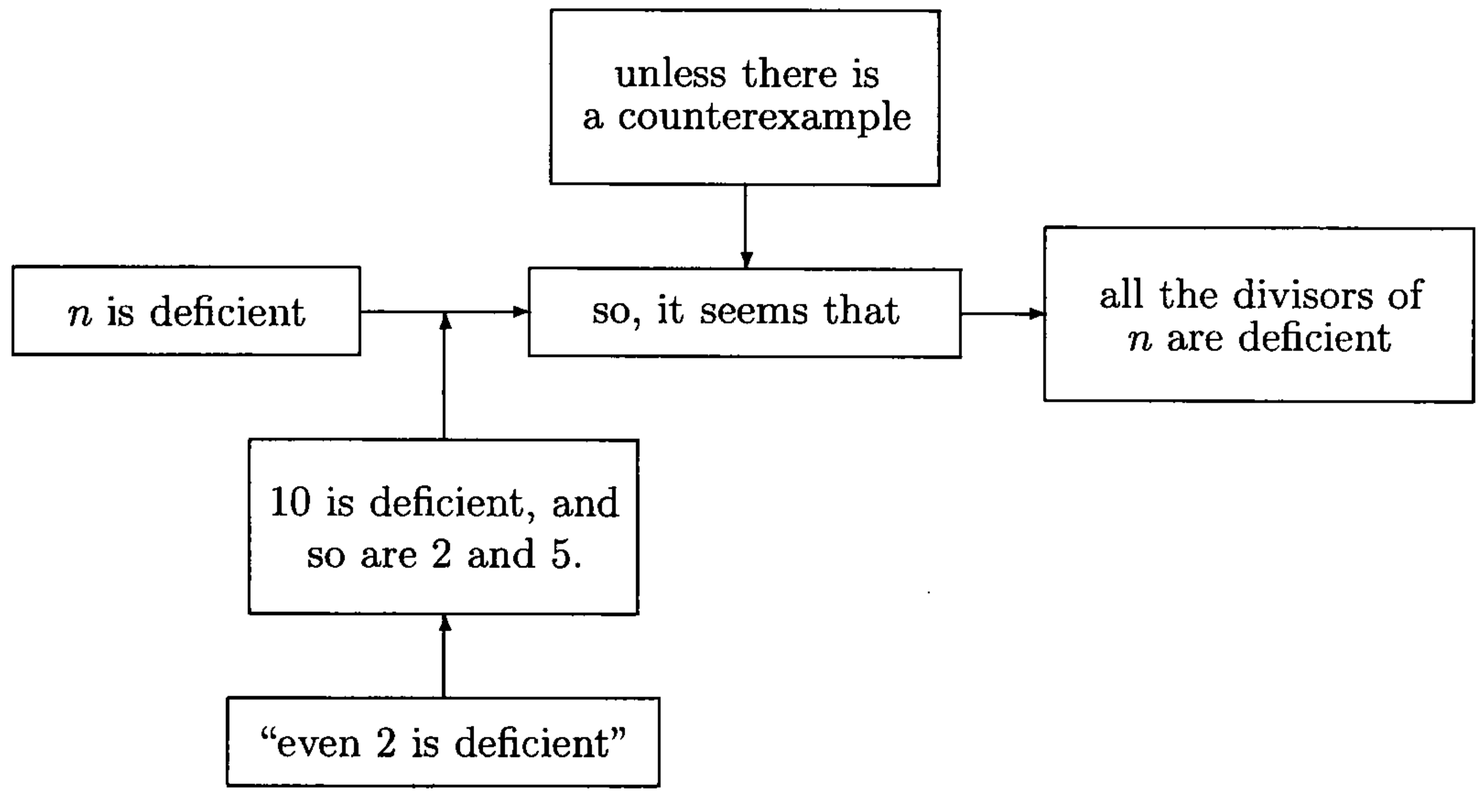


Figure 8.7: Part of Andrew's response to Conjecture 4.

possible counterexamples here, I mean, we might look for one of them, so does 6 divide into that? [*looks at the list of examples of perfect numbers*] I don't know, no it doesn't does it? So does 6 divide into the next thing? So, I can't see any counterexamples there, and for example... So I guess, umm, what was we, what would I, umm, we need to find some divisors that aren't of the form $2m$ for m a divisor of n, don't we? (D2)

David's search for counterexamples is a failure, so he concludes that it is plausible that the statement is true, and attempts a proof. It is notable that before he looked for possible counterexamples, David seemed unsure of whether it was true or not. It is also worth noting that the list of examples David used contained only $28,496$ and 8128. After noting that 6 did not divide any of these three perfect numbers, David was sufficiently satisfied that the conjecture was true that he began a proof attempt. His (two stage) argument is modelled in Figure 8.8. Note that David happily used two different sorts of warrant-qualifier pairs in the argument. The first stage uses a modal qualifier that carries certainty, whereas the second only carries plausibility.

Another example of an inductive warrant was Chris's behaviour when working on Conjecture 6.

INTERVIEWER: So you think it's true?

CHRIS: Umm, I'm not sure, I mean I haven't actually tried an example, because I'm too afraid to work out the prime factorisation of something that big, I suppose, well, ok, 12×18, Right [*laughs*] so [*muttering*] umm, so the factorisation is going to be $3 \times 4 \times 3 \times 2 \times 3$ so 196

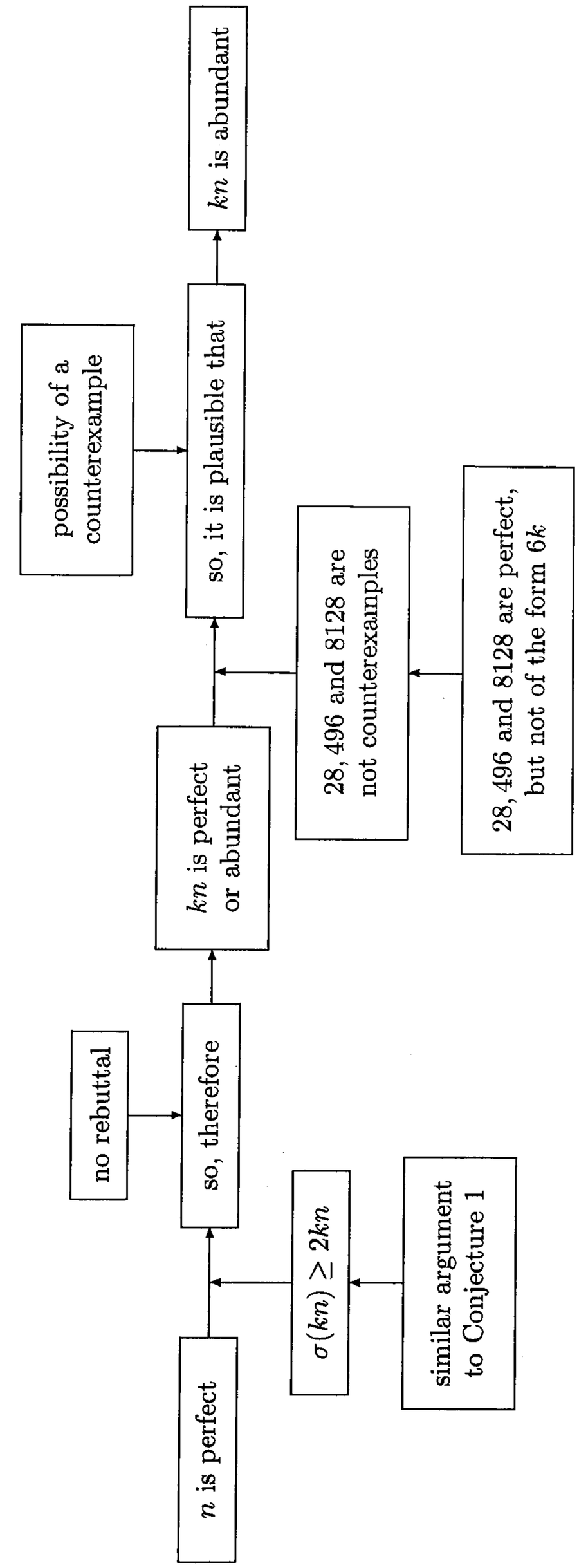


Figure 8.8: Part of David's response to Conjecture 2.

factors to $2^3 \times 3^3$ in fact. [*begins to add up divisors*]

INTERVIEWER: Well, actually, that's going to be abundant.

CHRIS: Right, ok, good. Umm, ok, I mean it certainly looks plausible. (C6)

When the interviewer intervened with the information that 12×18 was in fact abundant, Chris became convinced that the statement was "plausible", and began an attempt to prove it. The interview ran out of time before he was successful. Again, here, Chris conducts an empirical investigation into the situation to evaluate how confident he is in the conclusion of the argument.

Edward also used an inductive warrant to fix his level of belief in a conclusion of an argument. When studying Conjecture 4, he began by thinking it was probably untrue, but became convinced through the use of examples and a lack of counterexamples:

INTERVIEWER: So you think that is unlikely to be true?

EDWARD: No, I'm not sure now. If n is deficient then every divisor of n is deficient. 14 is a deficient number. And certainly all the divisors of 14 are deficient. [*long pause*] Oh, OK, so you've given me a list of the first few abundant, this is the complete list? There aren't any missing?

INTERVIEWER: No, no.

EDWARD: OK, so umm, these are abundant, these are perfect, so obviously all the deficient ones are the ones that aren't in these lists, so, I don't want to go for, I want to go for some interesting numbers. I also want to go for some fairly large numbers because I'm pretty, like 14 is 2×7 and I think that, well it's blatantly obvious, well just one prime number times another prime number, you're not going to get anywhere at all, they're clearly deficient. So, umm, but of course, this is why I'm thinking that this probably is true now, because all of the interesting cases are already grabbed. Err, [*long pause*] so there's no obvious counterexamples that I can see from that list, like numbers that are missing from that list. [*long pause*]. (E4)

This argument is modelled in Figure 8.9.

It is clear that, on many occasions, degrees of belief in conclusions were fixed by the participants with the use of examples. Two distinct strategies emerged:

- The use of examples (e.g. Figures 8.5, 8.7, and Chris's argument in Conjecture 6).
- The use of counterexamples (e.g. Figures 8.8, 8.9).

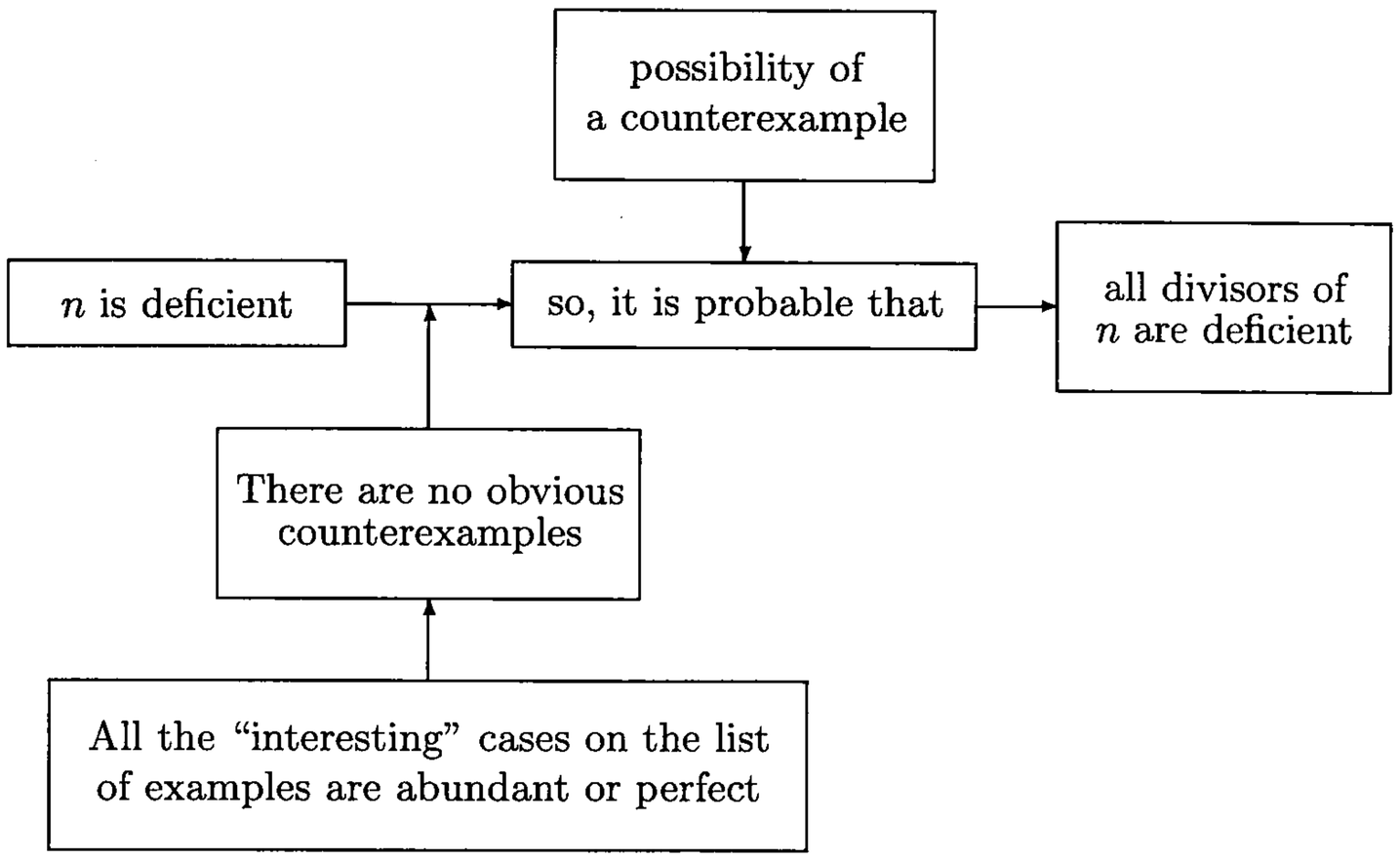


Figure 8.9: Part of Edward's response to Conjecture 4.

The first strategy involves the use of an example as a 'crucial experiment' (Balacheff, 1988) to test whether the Conjecture held in that case or not. If it did, a 'plausible' modal qualifier was used to join the conclusion with the data. The second strategy was somewhat different, in this case participants looked through a series of examples to see if they could find a counterexample. If they couldn't find any counterexamples, again a 'plausible' modal qualifier was used. The second strategy, then, is more like Balacheff's 'naive empiricism'.[8]

In both cases, once participants had fixed their degree of belief in the conclusion, they attempted to prove it formally. No participant used an inductive warrant to deduce *with certainty* that the conclusion followed from the data, and this would not have been expected, since they were all highly talented mathematicians. However, inductive warrants were used widely to establish participant's belief in the conclusion, through moderation by appropriate modal qualifiers.

8.6.4 The structural-intuitive warrant-type.

This section introduces the term 'structural-intuitive' to refer to a participant using observations about, or experiments with, some kind of mental structure, be it visual or otherwise, that persuades a them of a conclusion. Often, but not

[8]It should also be said that there were also several cases of participants using 'generic examples' (Balacheff, 1988). These tended to be useful *after* the degree of belief in the conclusion had been fixed. Generic examples seemed to be most helpful in generating knowledge about the situation in order to produce a formal proof, they tended not to be used to evaluate the conclusion.

necessarily always, this sort of reasoning appears to be of an intuitive type (in the sense of Fischbein, 1987, see §7.2.1).

The notion of the structural-intuitive warrant-type is related to two differing proof schemes in Harel and Sowder's (1998) taxonomy:

- Perceptual: "Perceptual observations are made by means of rudimentary mental images – images that consist of perceptions and a coordination of perceptions" (p.255).
- Transformational: "Transformational observations involves operations on objects and anticipations of the operations' results" (p.258).

It is not at all clear how these two schemes are related, or how it is possible to distinguish between them by observing mathematicians' behaviour. To complicate matters further, Harel and Sowder's (1998) notion of the transformational proof scheme has itself been reconceptualised by Harel (2001, in press). With reference to the current Experiment, based as it is on empirical data, it seems to make more sense to abandon Harel and Sowder's two categories, and categorise warrants using terminology more easily identifiable with actual behaviour.

Chris's reasoning about Conjecture 4 exemplifies this kind of warrant:

CHRIS: So if n is deficient then we get for free that umm, none of it's divisors are perfect, so every divisor must be deficient or abundant. Umm, it would seem odd if they were allowed to be deficient and abundant but not perfect. Because perfect is kind of the middle case, so it looks true. (C4)

This argument is modelled in Figure 8.10. Chris's warrant here is based on some intuitive understanding about how the properties of deficiency and abundantness should behave, and a realisation that if the conjecture was false, it would mean that these properties would have been broken.

Sometimes participants were unable to back their structural-intuitive warrants that they used. Whilst working on Conjecture 4, for example, Ben admitted that he thought it was true:

BEN: So, I think, my initial thought is that it's true, that every divisor of it is deficient.
INTERVIEWER: Why?
BEN: Just 'cos it seems sensible [*laughs*] Umm...
INTERVIEWER: So that's just a gut feeling you've got?
BEN: No, no real mathematics involved at all! [*laughs*] (B4)

Whereas Chris could tangibly justify his structural-intuitive warrant, Ben seemed either unable or unwilling to. Nevertheless, this warrant led him to start a proof search which eventually proved successful.

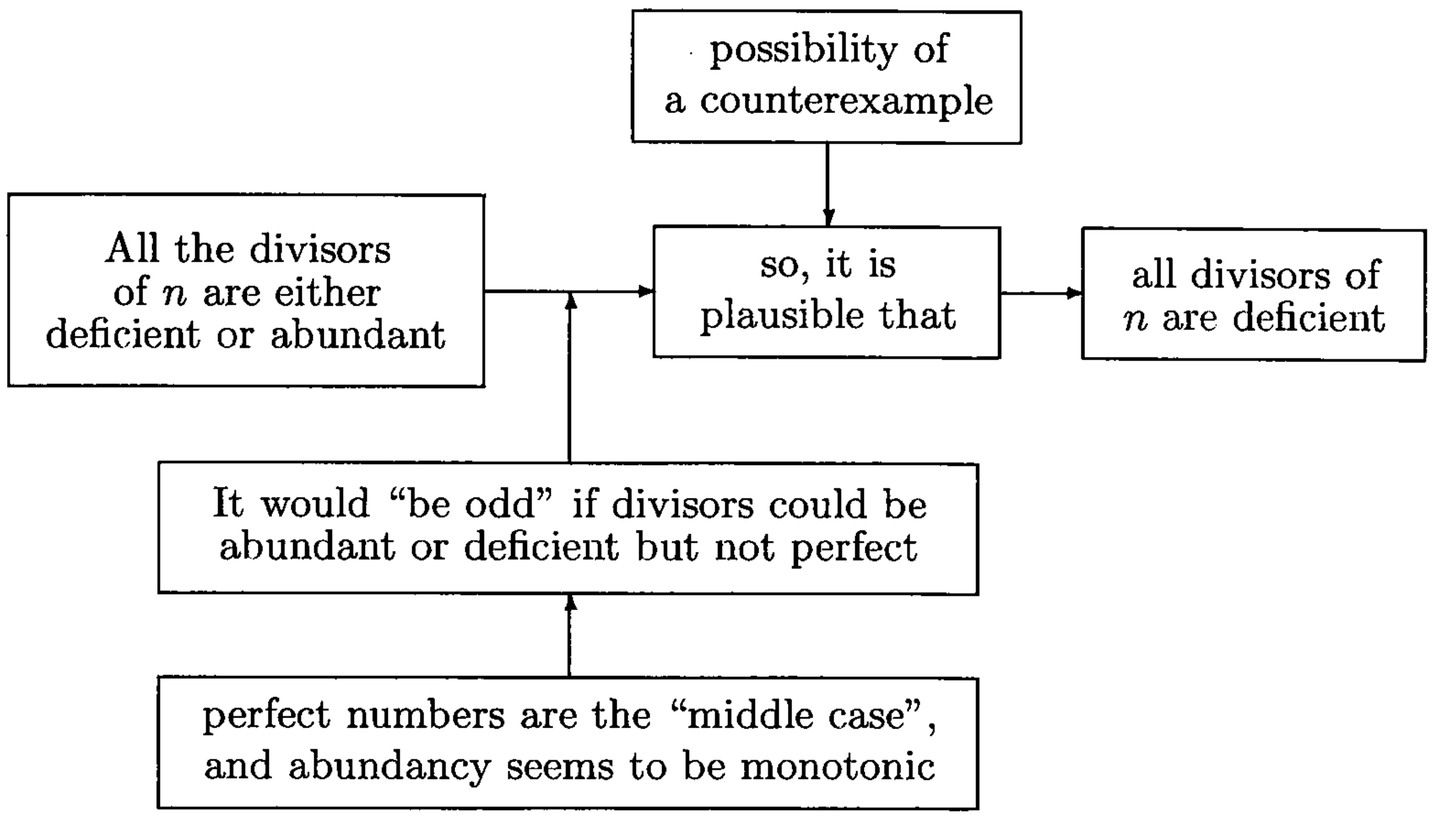


Figure 8.10: Part of Chris's response to Conjecture 4.

Structural-intuitive warrants in Conjecture 5.

Conjecture 5 asked participants to evaluate the statement "if m and n are abundant, then $n + m$ is abundant". This provoked many structural-intuitive warrants which justified the conclusion that the conjecture was false. Take Fred, for example:

FRED: Umm. conjecture 5, if m and n are abundant then $n + m$ is abundant [*long pause*]

INTERVIEWER: Is it true or false?

FRED: I think, going on instinct, it's probably false.

INTERVIEWER: Why?

FRED: Because, err, I mean, think of it... Another typical example, like, saying whether something is abundant is to do with it's divisors, so it's to do with things that divide it, it's to do with multiples. And then, when you add two numbers together, it doesn't necessarily mean that any properties of the divisors stay the same. I mean, like, I don't know, when you add 3 and 5. 3 and 5 have certain divisors, but 8 has completely different divisors.

INTERVIEWER: Yeah.

FRED: Umm, but you never know. So, but abundant is a very sort of wide statement, so, I mean, intuitively you'd expect to apply to roughly half of all numbers, so maybe it's not so absurd to think they would, err, that would hold. So I'll try. (F5)

Fred went on to try to prove the statement before abandoning his attempt and looking for counterexamples.

Fred's behaviour here is interesting. Immediately after having read the statement he seems sure that it is false, and justifies his intuition with a structural-intuitive warrant based on the absence of a link between addition and divisors. However he seems, during the course of his articulation, to try to convince himself not to trust his original intuition. This was a mistake, as he found that the statement is indeed false, and looking for a counterexample is straightforward.

Similar structural-intuitive warrants were used by other participants:

EDWARD: If n and m are abundant then $n + m$ is abundant. [*long pause*] I am in general thinking it should be fairly, again, I'm going to eat my words, my immediate reaction is that it's going to be fairly easy to find a counterexample to this.

INTERVIEWER: Why do you say that?

EDWARD: Because, we're thinking of divisors and multiplication, so there's lots of abundant numbers, err, there's lots of abundant numbers, I'm just thinking more carefully now, but I'm thinking when you add them together you could easily end up with a deficient or a perfect number. [*long pause, looks at examples*] Oh, yeah, 18 and 20 add together to give 38, which according to your list is a deficient number. (E5)

Edward based his decision to look for a counterexample (which he found successfully) on his intuition that the statement was unlikely to be true. Chris used a similar structural-intuitive warrant:

CHRIS: Right, so if m and n are abundant, then $m+n$ is abundant. That doesn't look true.

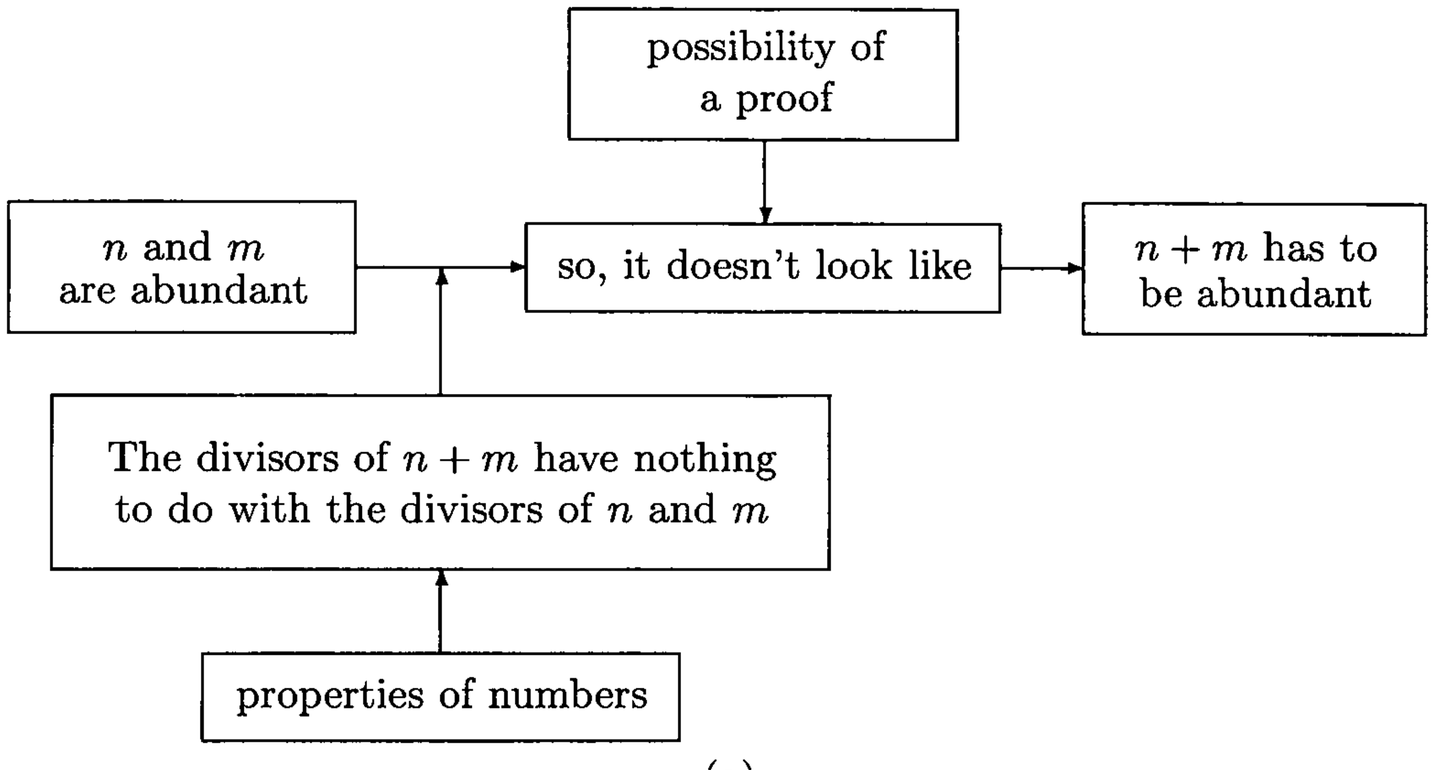
INTERVIEWER: Why not?

CHRIS: Because the factors of $n+m$ don't really have anything to do with the factors of n or m. So it should be fairly easy to construct a counterexample. I say that [*laughs*]. So if I pick two nice abundants, umm... (C5)

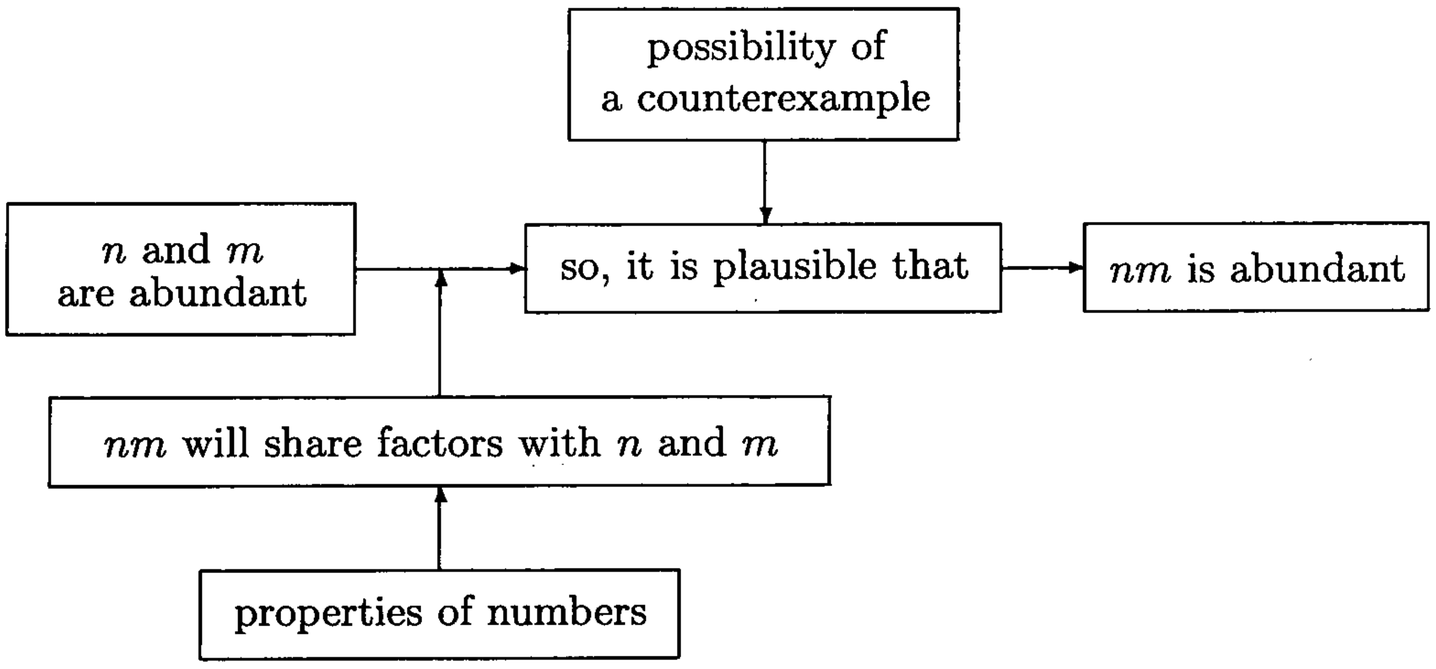
Chris went on to find a counterexample. His argument is modelled in Figure 8.11a.

Andrew's structural-intuitive warrant regarding Conjecture 5 was slightly different. He based his belief that the statement must be false with a structural-intuitive warrant based on the frequency of abundant numbers:

ANDREW: m, n are abundant, $m + n$ is abundant. Ahh, [*laughs*]



(a)



(b)

Figure 8.11: Part of Chris's response to Conjectures 5 (fig a) and 6 (fig b).

INTERVIEWER: So what's your reaction to that?

ANDREW: Ah, this is, this is, this is rubbish [*laughs*].

INTERVIEWER: Why do you say that?

ANDREW: Because you can, get, OK, if I get some for example, some huge prime, maybe 11 will be enough! [*laughs*] Err, yeah, and actually if I've got a prime then I can easily decompose it into, because it seems there are so many abundant numbers right? (A5)

It is interesting to compare participants immediate responses to Conjecture 5 with their corresponding responses to Conjecture 6, which followed immediately afterwards. Consider Edward's initial reaction:

INTERVIEWER: What do you make of that then? [*hands over conjecture 6*]?

EDWARD: [*Reads card*] Now that's more likely to be true. (E6)

Here is Chris's response:

CHRIS: [*Reads card*] Right, so if n and m are abundant then nm is abundant. That looks more plausible, cos they're going to share factors. (C6)

Chris's argument is modelled, in Figure 8.11b, alongside his argument in Conjecture 5 for comparison.

Both these excerpts illustrate the use of similar structural-intuitive warrants to those used in Conjecture 5, but here they carry plausibility rather than unlikeliness.

The pattern of these responses to Conjecture 5 was uniform. Participants' initial intuitions gave them a structural-intuitive warrant which they used to decide that the conjecture was unlikely to be true. This then directed their attempts at solving the problem. Since they had deduced that the conclusion was unlikely to be true, looking for a counterexample was the most appropriate strategy. In Conjecture 6 the situation was reversed. Participants used structural-intuitive warrants to determine that the conjecture was likely to be true, and then based their decision to look for a proof on this judgement.

However, it is also notable that several participants did not fully trust their intuition. Fred, for example, seemed to initially feel strongly that Conjecture 5 was untrue, but then talked himself out of it, and attempted to prove the statement. Similarly, Andrew, after he had initially dismissed the same conjecture as "rubbish" went on to question whether he was correct to do so or not.

As noted in Chapter 7, the reliability of intuition in mathematics has been a recurring subject of discussion by mathematicians and philosophers of mathematics. Hahn (1933/1960), when reflecting on recent developments such as

Peano and Hilbert's work on space-filling curves, even went as far as to argue that intuition is entirely unreliable and should be totally "expelled" from mathematical reasoning:

> "[Mathematicians] learned that it is unsafe to accept any mathematical proposition, much less to base any mathematical discipline on intuitive convictions. Thus a demand arose for the expulsion of intuition from mathematical reasoning, and for the complete formalisation of mathematics." (p.1959)

To back up his argument Hahn gave several examples of counter-intuitive 'monsters': a map of three regions which meet each other at every point along one border, and a curve which intersects itself at every point (e.g. Menger, 1943; E. H. Moore, 1900; Whyburn, 1942). These objects, Hahn argued, should are impossible to reconcile with intuition, and thus intuition needs to be removed from all mathematical reasoning. Other authors have disagreed with this kind of analysis, pointing out that although intuition may be sometimes misleading it is essential for giving direction to mathematical research (Feferman, 2000; Poincaré, 1905). The data from Experiment 4 supports this latter stance. Participants used their intuitive structures to establish a belief in whether the conclusion follows from the data. Structural-intuitive warrants were used to reduce uncertainty. Although, as the next section shows, these warrants were not always mathematically correct.

Are all abundant numbers even? Incorrect structural-intuitive warrants.

In several of the interviews the issue of whether abundant numbers need to be even was discussed. When Andrew was searching for counterexamples to Conjecture 5, for example, he noticed that none of the given examples of abundant numbers were odd:

ANDREW: Hmm. Strange.
INTERVIEWER: Why strange?
ANDREW: Strange that all these numbers are even. All these abundant numbers. (A5)

Whilst working on the same conjecture Fred noticed the same thing:

INTERVIEWER: Well, what numbers are likely to be abundant?
FRED: Err, well my thinking is, odd numbers are not, generally because...
INTERVIEWER: When you say generally what do you mean?

FRED: Just the general idea, cos like if a number is even then one of its divisors is half the number, which is a pretty big chunk, but if a number is odd it's missing a big chunk.

INTERVIEWER: So, you reckon no odd numbers are abundant?

FRED: I think that's quite unlikely. (F5)

During the course of his interview Ben was asked directly about the issue. Ben's argument was similar to Fred's:

INTERVIEWER: Just as an aside, would you say all abundant numbers have to be even?

BEN: Umm, [*long pause*]

INTERVIEWER: I mean that's an incredibly difficult question, but what's your sort of, if you had to stab in the dark about it?

BEN: Umm, [*long pause*] I think it might have to be true.

INTERVIEWER: Why?

BEN: [*long pause*] I think if they're odd, you lose too much of the, sort of, sequence that you can't divide into, if you get what I mean, cos you can't, you can't divide past, so say it was odd and the first one was 3, you'd only have ones up to $\frac{n}{3}$ and then n, whereas if you go up to $\frac{n}{2}$ you get a lot more, well in theory, you could get a lot more possible divisors, so it's based on a sort of size argument rather than anything particularly... But, intuitively, even numbers would certainly be more likely to be abundant than odd numbers. (B5)

The structure of Ben and Fred's arguments is modelled in Figure 8.12. Both used structural-intuitive warrants about the structure of abundant numbers, understanding that they had built up through working on previous conjectures.

Interestingly Andrew used a different structural-intuitive warrant regarding this issue, and ended up with a different conclusion:

ANDREW: Strange that all these numbers are even. All these abundant numbers.

INTERVIEWER: Hmm, interesting, do you think that that will always be the case?

ANDREW: I don't think so. Well, [*pause*] ah, it'll be the case, if we take, for example, 3×7, 3×5 as well, $\times 11$, then I'm pretty confident that this is an abundant number.

INTERVIEWER: Are you? [*Andrew laughs*] Why are you laughing?

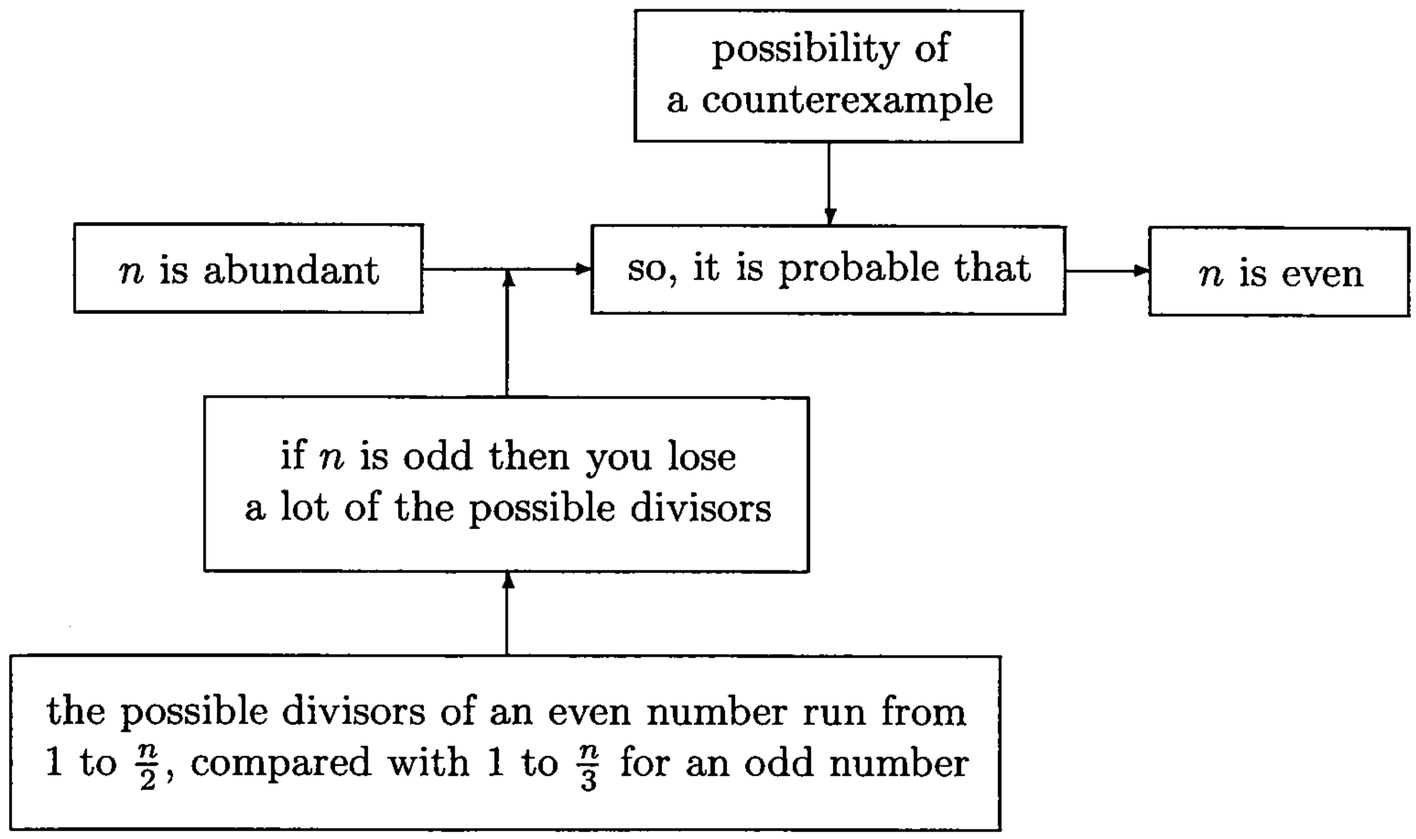


Figure 8.12: Ben and Fred's argument regarding the parity of abundant numbers.

ANDREW: [*laughs*] Because we just showed many times that I massively overuse the word "pretty sure" [*laughs*]. OK, so for instance if we take [pause] well [...] Well, apparently better if these primes probably need to be different, because otherwise one is losing many divisors. Actually, would we be, really need it? If one basically checks the frequencies of the primes? So take some primes, a few odd primes that are very close to each other. I think so. (A5)

Andrew went on to unsuccessfully attempt to construct an odd abundant number, by multiplying together several odd primes. His argument is modelled in Figure 8.13 (note that the backing to Andrew's structural-intuitive warrant is unknown, as he did not verbalise it). Although Andrew reached a different conclusion to Ben and Fred, his argument was of a similar structure. He used a structural-intuitive warrant to draw a conclusion with a 'probable' modal qualifier. In both cases, the warrant was used to make a probabilistic judgement about the conclusion.

Despite Andrew's self-deprecating view of his use of the words "pretty sure", his intuitions were indeed correct. There are an infinite number of odd abundants, with 945 the first.[9] Indeed, surprisingly, it is possible to construct an abundant number whose smallest divisor is arbitrarily high. So Ben and Fred's (highly reasonable) structural-intuitive warrant that odd abundants are unlikely, as the divisors can only 'live' in the lowest third of the number turns out to be

[9]If there is one odd abundant then it follows that there is an infinite number: clearly every odd multiple of 945 will also be an odd abundant.

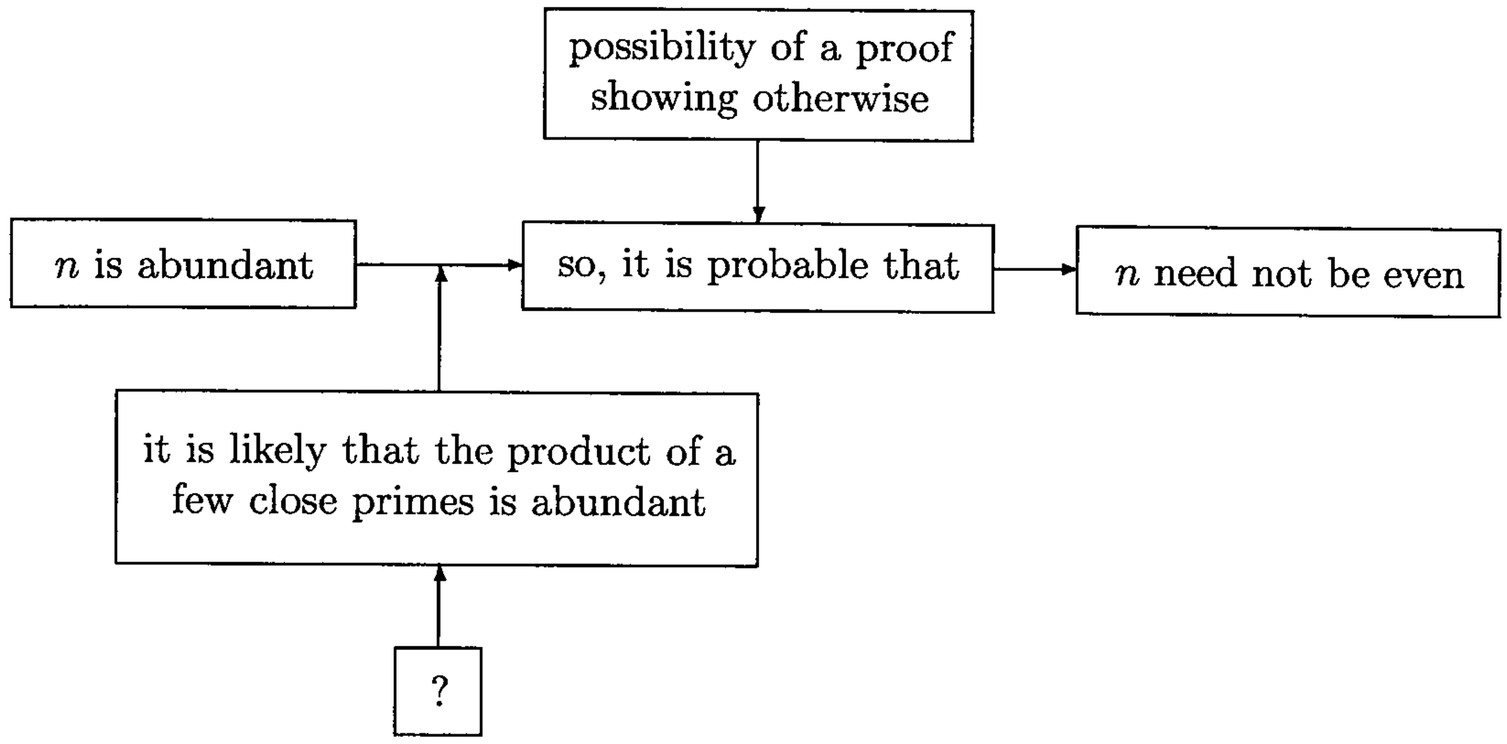


Figure 8.13: Part of Andrew's response to Conjecture 5.

quite wrong. An abundant can be constructed where the proper divisors are less than $\frac{n}{k}$ for any $k \in \mathbb{N}$ (for a proof of this result, see Appendix C). Note that even though Ben and Fred's argument was incorrect, for them, their warrant had the effect of reducing their level of uncertainty about the truth/falsity of the conclusion.

Summary of §8.6.4.

The use of 'structural-intuitive' warrants was found to be widespread in the behaviour of participants in Experiment 4. These warrants refer to the use of a justification of a conclusion on the basis of an, often intuitive, argument regarding the structure of the mathematical objects in question. These arguments can take the form of observations of, or intuitions about, the structure of the mathematical situation, or thought experiments with properties of the structure. In terms of Harel and Sowder's (1998) proof scheme framework, it is unclear whether this category of warrant is more closely related to the transformational scheme or the empirical-perceptual scheme. Either way, when used appropriately (as was the case with all participants in Experiment 4), structural-intuitive warrants carry with them modal qualifiers which are not certain, only probabilistic. These types of warrants tended to be used to determine the likelihood of a statement's truth, a judgement which was then used to determine the most appropriate strategy.

However, there are examples of warrants used by participants which carry different sorts of modal qualifiers to the warrants discussed so far, often deductive warrants *do* carry certainty, and are closer to what would normally be regarded as a formal mathematical deduction.

8.6.5 The deductive warrant-type.

Harel (2001) referred to the most sophisticated proof scheme as the 'deductive-modern-axiomatic' scheme: people who have this scheme use deductions from axioms to establish truth. Harel (2001, in press) sought to distinguish between various different forms of the deductive proof scheme; for example, the Greek-axiomatic is seen as being different to the structural-axiomatic, which is in turn different to the axiomatising-axiomatic. The differences appear to revolve around views of the role of the axioms. Whereas someone with the structural proof scheme sees axioms as being permanent descriptors of a structure, someone with the axiomatising-axiomatic proof scheme is aware that the axioms could be varied, and the consequences of the variation studied. As with the transformational and perceptual proof schemes, it seems extremely difficult to see how these different schemes can be distinguished through behaviour, and so any philosophical benefit accrued through their inclusion in the taxonomy is balanced by an increased confusion for the empirical researcher. No differences between the types of deductive warrants used by participants of the sort described by Harel were observed in the current study.

A similar notion to Harel's (2001) deductive proof scheme is the basis of the deductive warrant-type: formal mathematical justifications are used to warrant the conclusion of the argument in question. These justifications can be of various sorts: deductions from axioms, algebraic manipulations, or the use of counterexamples would all be classified as deductive warrants.

For professional mathematicians, a deductive warrant is seen as carrying formal mathematical necessity: an argument that uses a deductive warrant admits no effective rebuttal. It could be argued that, in complex proofs, mathematicians *do* sometimes have non-trivial qualifiers and rebuttals – such as 'unless there is a flaw in my argument' – but the aim of these forms of argument is to minimise this. Thus, while for professional mathematicians the inductive and structural-intuitive warrant types *aim* to reduce uncertainty, the deductive warrant *aims* to *remove* uncertainty. Although this is the case for professional mathematicians, it may not be for all students: the potential for constructing an inappropriate matching between deductive warrants and modal qualifiers is discussed later in the thesis.

Examples of this kind of warrant abound in the data from Experiment 4, perhaps because the participants were all highly qualified mathematics post-graduates. For example, when Andrew was working on Conjecture 2 he produced the following argument (he had used a different approach in Conjecture 1):

ANDREW: OK, so if n is perfect, then kn is abundant, for any k. OK, so

what does it, yeah it looks, so what does it mean? Yeah so if n is perfect, and I take any p_i which divides this n, then afterwards the sum of these p_i's is $2n$. This is the definition. Yeah, ok, so actually we take kn, then obviously all kp_i divide kn, actually, we sum these and we get $2kn$. Plus, we've got also, for example, we've also got k dividing this, dividing kn. So we need to add this. As far, as basically, there is no disquiet, k would be the same as this. Yeah. And, how would this one go? [*long pause*]

INTERVIEWER: So we've got the same problem as up here [*Conjecture 1*] but in general? With a...?

ANDREW: Yeah. Umm, can we find one? Right, so I don't know. Some example.

INTERVIEWER: I've got some examples for you.

ANDREW: You've got examples of some perfect numbers? OK, so 12, we've got $1+2+3+4+6$, then, ok, $+12$. [*mutters*] But this is not? OK, perfect, I wanted perfect numbers. OK, so let's say 6. Yeah, and we've got $1+2+3+6$ and actually we take 2×6 which is 12. Then yes, I've got divisors $2, 4, 6, 12$. Plus I claim we've got also divisors. Yeah! actually it's simple because, err, because err, the argument is that we've also got 1 which is divisor, and this divisor is no longer contained here if we multiply.

INTERVIEWER: Right so we've always got a spare 1?

ANDREW: OK, so this argument also applies here [*Conjecture 1*]. (A2)

There are two quite distinct stages to this argument, which is modelled in Figure 8.14. Both stages use a deductive warrant. The first establishes that n must be either abundant or perfect. At this stage Andrew realises that he needs to reject the perfect case, and uses the generic example of 2×6 to do this (Balacheff, 1988). In both stages the conclusion follows necessarily from the data. No rebuttals are possible.

Although Andrew believed that no rebuttals were possible, in actual fact there is a possible rebuttal, the case where $k = 1$. When prompted by the interviewer, he immediately recognised this trivial case, and modified the conclusion appropriately.

Not all deductive warrants are of this form, where conclusions are deduced from data by logical implications. Sometimes, for example, participants used counterexamples to warrant their conclusions.[10] Here is part of Fred's argument

[10]It is unclear where counterexamples fit into Harel and Sowder's (1998) taxonomy, arguments based on counterexamples do not really fit into either the axiomatic proof scheme or the empirical proof scheme.

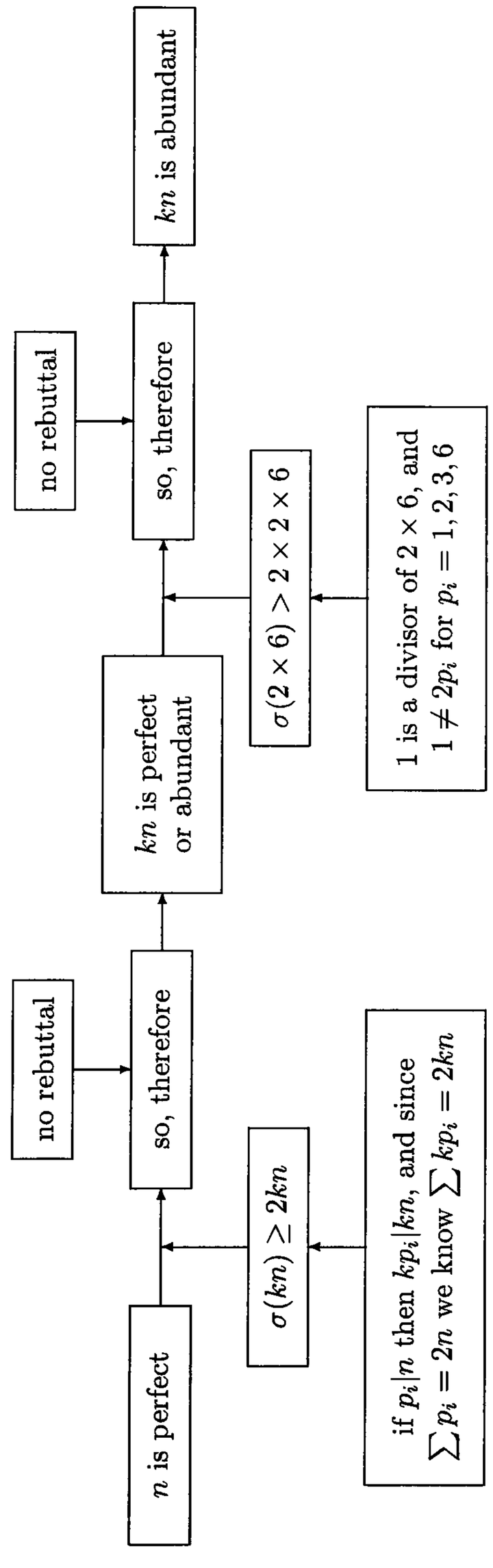


Figure 8.14: Part of Andrew's response to Conjecture 2.

in Conjecture 1. After he had successfully shown that all *proper* multiples of 6 are abundant, he turned his attention to 6 itself:

EDWARD: Yeah. [...] so this leaves the special case of $n = 6$... which is a perfect number, I just know that. Or is it? Hmm. $1+2+3+6 = 12$, yeah, so it's a perfect number. Abundant, does that mean greater or equal to, or just greater? So, that's, that's a counterexample, so when $n = 6$ it's not an abundant number. So... we've got rid of the 'if'. (E1)

Another example of the use counterexamples came from Chris's response to Conjecture 3:

CHRIS: So, think, large primes. OK, I shouldn't have to go too large, umm, but something like, name a large prime? 97 I think is. Oh dear. Umm, I mean, and probably the other one doesn't need to be very big, perhaps 5 would do, because I can multiply by 5. So 97 times 5 is going to be 485. Err, take away 97, 5 and 1 is clearly going to be positive.

INTERVIEWER: So what does that tell us?

CHRIS: So I think this is a counterexample. But possibly I've got this sign the wrong way round, I'll just plug them into here and see if that works. So the question is, is 485 abundant? And I know that the divisors are $1, 5, 97$, and the number itself. Yeah, so that in fact, I was right. OK, so that's a counterexample to this.

INTERVIEWER: So we've shown the thing's false?

CHRIS: Yes. (C3)

This argument is modelled in Figure 8.15. Notice that in this model a modal qualifier of "it is not the case" has been used. Whilst clearly this is different to the "therefore" modal qualifier in Figure 8.14 above, the two modal qualifiers carry the same amount of *uncertainty*, i.e. none. This is the key distinction between the deductive warrants and those discussed in previous sections.

The role of contrapositive arguments, with reference to heuristic biases, was discussed in §8.5. It was argued that certain heuristic biases direct attention towards considering different parts of conditional statements and that this affects the strategy adopted. There were several examples of participants using contrapositive based proofs when studying the different conjectures. The structure of these types of argument is somewhat different to straightforward deductive warrants. With a contrapositive proof, a whole new argument is contained within the warrant/backing structure. Take, for example, Fred's argument in Conjecture 4. After spending some time deducing that a contrapositive argument

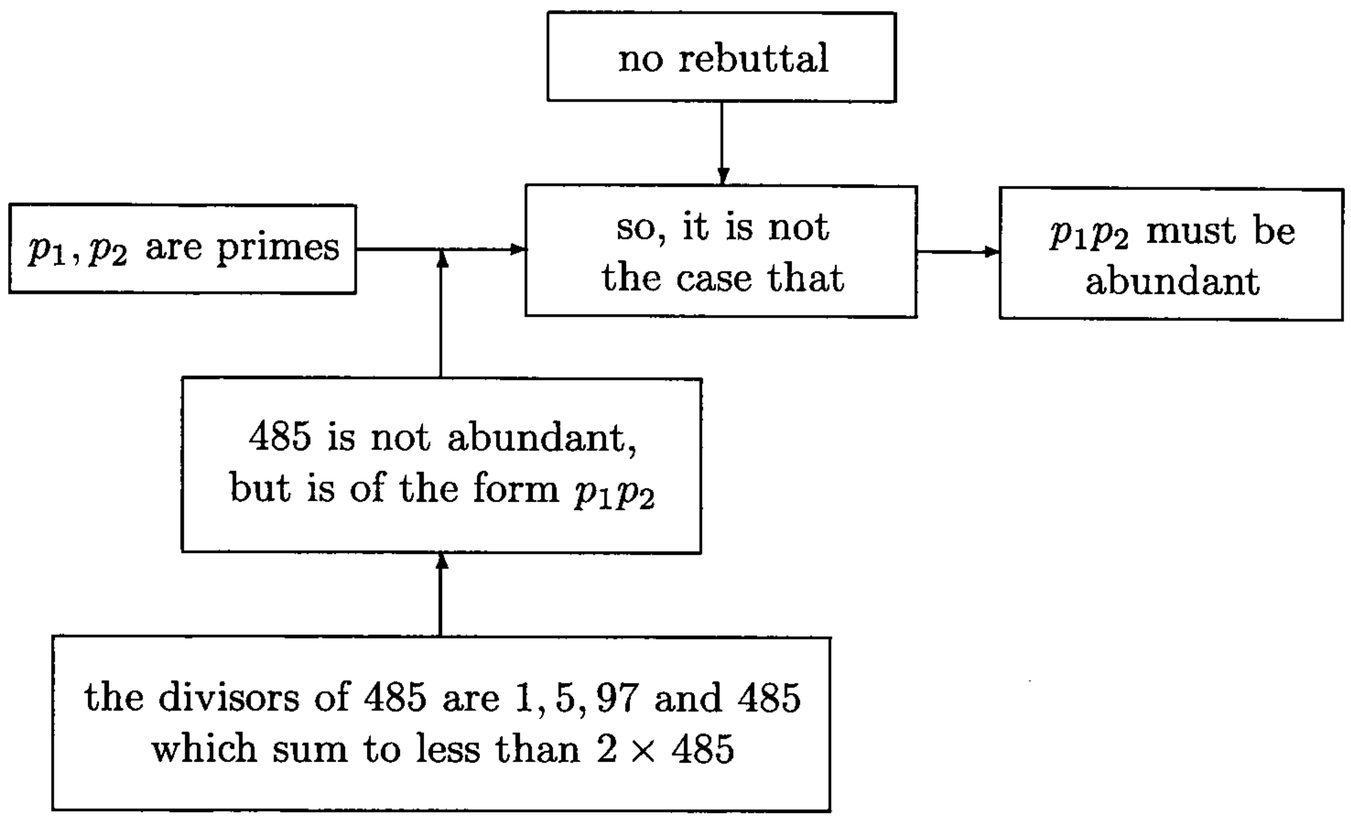


Figure 8.15: Part of Chris's response to Conjecture 3.

might be fruitful, Fred formulated the contrapositive statement, and argued as follows:

FRED: So, let k be non-deficient. This implies, using my notation from there, sigma k in brackets is greater than or equal to $2k$ [*using standard notation this means* $\sigma(k) \geq 2k$, *Fred used the notation* $(\Sigma k) \geq 2k$] ...Err... [*long pause*]

INTERVIEWER: So k here is just any non-deficient number?

FRED: It is yes. Umm, so... [*long pause*] no, it's a non-deficient divisor of n. Oh, actually, am I...? I'll go here. So, n equals mk for some m natural number. Err...OK, so similarly to the way we did Conjecture 1, err...every divisor of k is also a divisor of n. So we've got sigma k in brackets, so that's the sum of all the divisors of k, times m is actually greater than $2km$... which equals $2n$, umm,... so, and then that, the sigma k times m is less than or equal to sigma n in brackets. So, translated that means the sum of divisors of n is greater than or equal to blah blah blah, blah blah blah, and you get out the end is greater than or equal to $2n$.

INTERVIEWER: So we've shown here that, what?

FRED: That there is a non deficient divisor of n. So there exists a non deficient divisor, I'm going to use 'nd', 'ndd', non deficient divisor, of n, implies that n is not deficient. Which I think, yeah, that's the statement that I set out to prove, so the conjecture is true. Conjecture 4 is true. (F4)

This two stage argument is modelled in Figures 8.16 and 8.17. The first diagram

models the top level of the argument, it asserts that because a contrapositive argument exists the conclusion can be drawn from the data. This deductive warrant is backed by two different factors: that the contrapositive is logically identical to the direct statement, and by a whole new argument, shown in Figure 8.17.

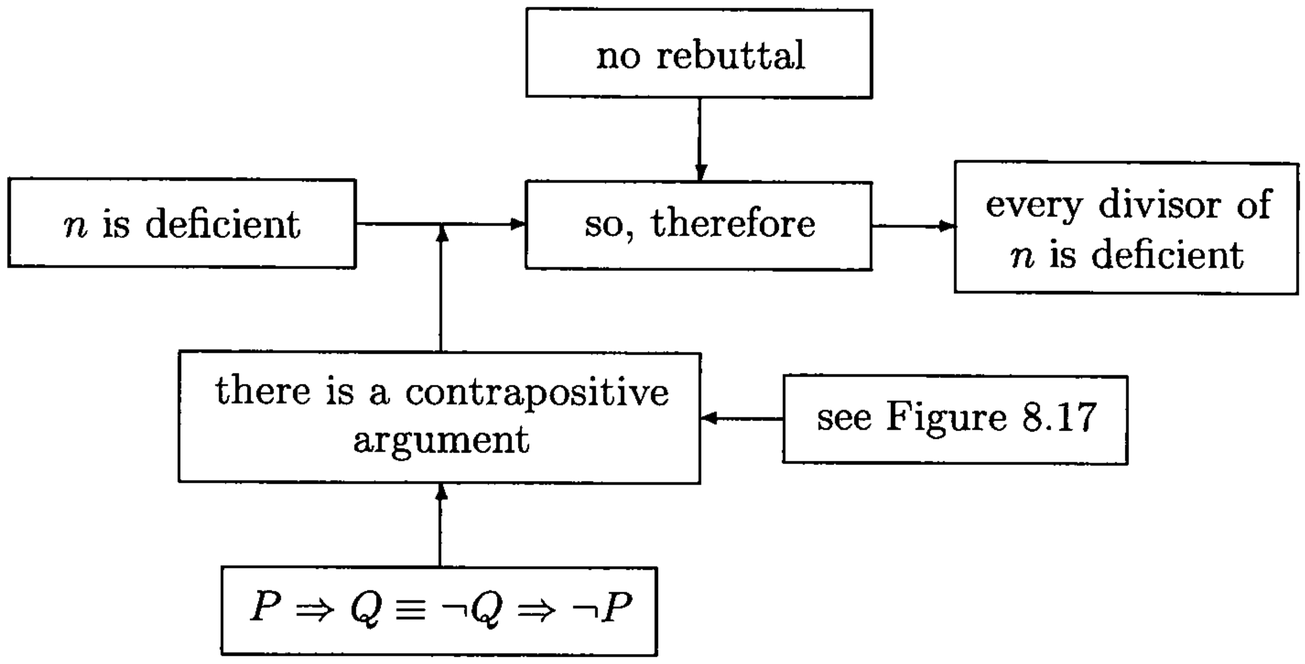


Figure 8.16: Part of Fred's response to Conjecture 4 (part 1).

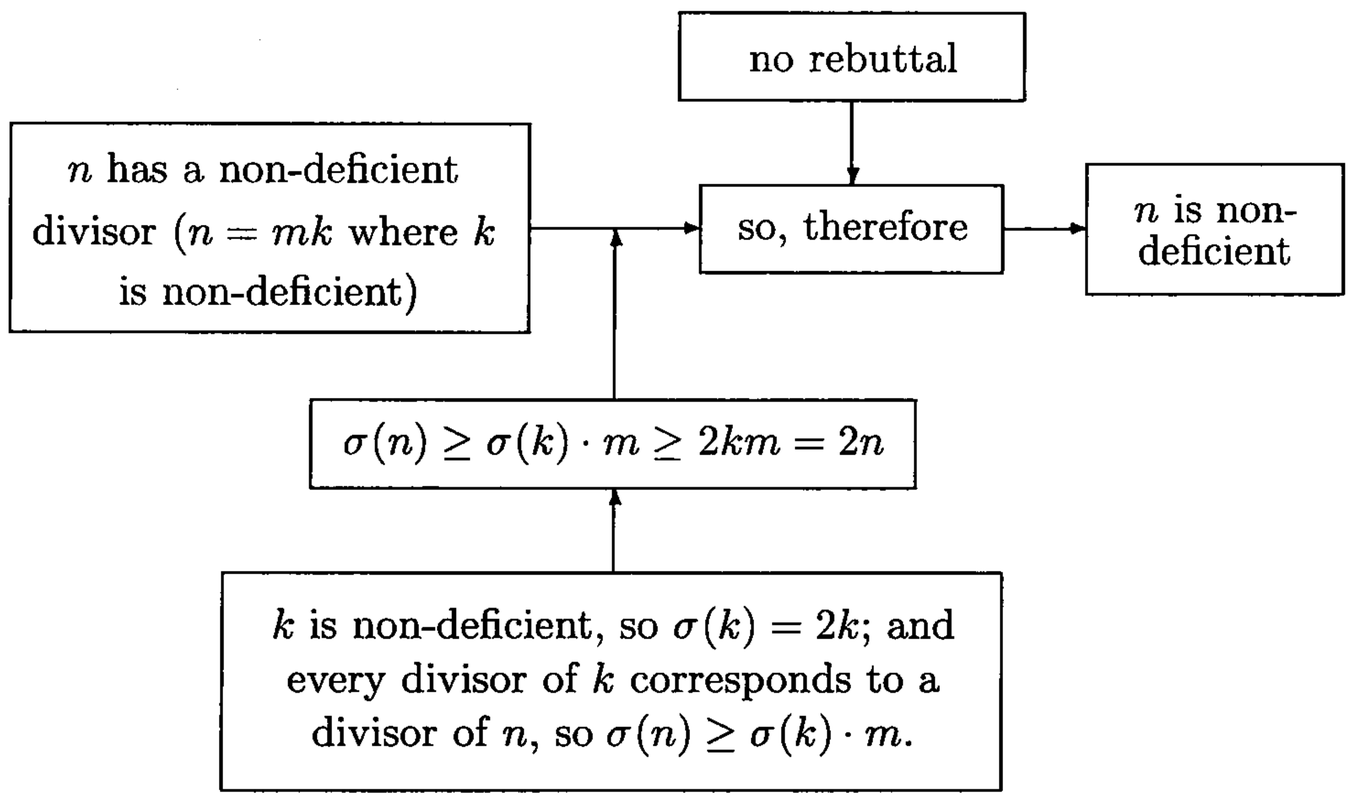


Figure 8.17: Part of Fred's response to Conjecture 4 (part 2).

This use of an 'nested' argument as part of the warrant/backing structure in another argument is an adaptation of Toulmin's (1958) scheme. He does not himself use arguments in this way, however, this structure seems to be closer to the behavioural reality of Fred's reasoning than any other manner in which his argument could be modelled.

This structure of Fred's argument also ties in with the discussion of the if-heuristic in §8.5. Recall that the if-heuristic suggests that attention is directed

towards the P part of 'if...then' statements. The use of Toulmin's model of argumentation gives further suggestion as to why this preconscious heuristic makes ecological sense. If the evaluation of conditional statements revolves around bridging the gap between data (P) and conclusion (Q) through the use of a warrant, it would be odd to initially focus attention on the possibility of a contrapositive argument, buried as it is within a lower level argument.

The idea that the job of a deductive warrant is to allow the reasoner to bridge the gap between data and conclusion – or to, in some sense, travel from data to conclusion safely – came across strongly in the language used by several participants. Implications were, on occasions, talked of as if they were journeys from data to conclusion:

> FRED: I really do think you've got to *go* from the opposite direction [...] I can think of *a way to go* I think [...] that's the statement that I *set out* to prove. (F4, my emphasis)

> EDWARD: I'm *going along* now a *different avenue* to see if I can think about it. (E1, my emphasis)

In both these cases Fred and Edward seem to be talking as if they are going on a journey. They want to traverse the difficult landscape between data and conclusion via the use of deductive warrants (see Lakoff & Johnson, 1980).

Summary of §8.6.5.

This section has only given a few examples of the deductive warrants used by participants in Experiment 4. Their use was widespread throughout the study and were always of the same type. Some kind of formal, or formalisable, mathematical argument that allowed the participants to deduce, with a modal qualifier that carries certainty, a conclusion from data.

Interestingly, these deductive warrants tended to be deployed after the phase where structural-intuitive and inductive warrants were most common. Once participants had convinced themselves of a statement's probable truth or falsity, they could begin to look for a proof that used deductive warrants to bridge between data and conclusion.

Although the vast majority of warrants used by participants fell into the categories discussed above, there were a few cases that fit more comfortably into other categories, again based upon Harel and Sowder's (1998) taxonomy. It is these warrants that are discussed in the next section.

8.6.6 Other warrant-types.

Harel and Sowder's (1998) taxonomy of proof schemes included many more

than the schemes which correspond with the three types of warrants discussed above. One major category of proof schemes they discussed were those based on conviction gained externally. Harel and Sowder had three categories of these:

- Ritual: conviction gained via the appearance of an argument rather than by its correctness.
- Authoritarian: conviction gained via an appeal to an authority figure.
- Symbolic: conviction gained through manipulating symbols "as if they possess a life of their own" (p.250).

There were only a couple of examples of these types of warrants seen in the data gathered in Experiment 4.

When David was working on Conjecture 2 he had successfully shown that if n is perfect then kn must be perfect or abundant, and was convinced that the case where kn is perfect could be eliminated. He argued:

DAVID: Well suppose k divides n and k^2 doesn't divide n, umm, then k^2 divides kn, so that's a number that would make this strictly greater than, and then I guess you're looking at some sort of induction thing, so suppose k^2 divides n but k^3 doesn't divide n, k^3 divides kn erm, and umm,

INTERVIEWER: Can you assume that though? Can you assume you've got a k that divides n?

DAVID: Well, no, just suppose it does so what you're looking at is you're saying let k be a natural number, err, that well k is some natural number, if k doesn't divide n then the theorem's true I think, if it divides n but k^2 doesn't divide n then the theorem's true, if k^2 divides n but k^3 doesn't divide n then the theorem's true. And I think you can just carry on, and k's either going to umm, err, well I mean hopefully, have we done everything now?

INTERVIEWER: you can keep going for ever?

DAVID: [...] Yeah, well I assume that works, as I say, I haven't written it down and I would always say to my students you know, I've got, this seems like an idea, I now think I believe this conjecture, you've now got to sit down and try and write it out. But I wouldn't want to do that and say it at the same time, because it would, you know, you'd have to do that on your own. But it looks plausible I think. (D2)

This argument is modelled in Figure 8.18. David appears to have gained conviction through the appearance of his induction argument. He feels that this kind of argument is of the correct structure – it "looks plausible" – so is hopeful that

it will work. Note, however, that he admits the possibility that the argument may not work. For David, this ritual warrant does not carry an absolute modal qualifier.

Although classified here as being a ritual warrant, David's warrant here has some similarities with the structural-intuitive warrants discussed earlier: it seems almost to be a structural-intuitive warrant about the *appearance* of the argument he has produced.

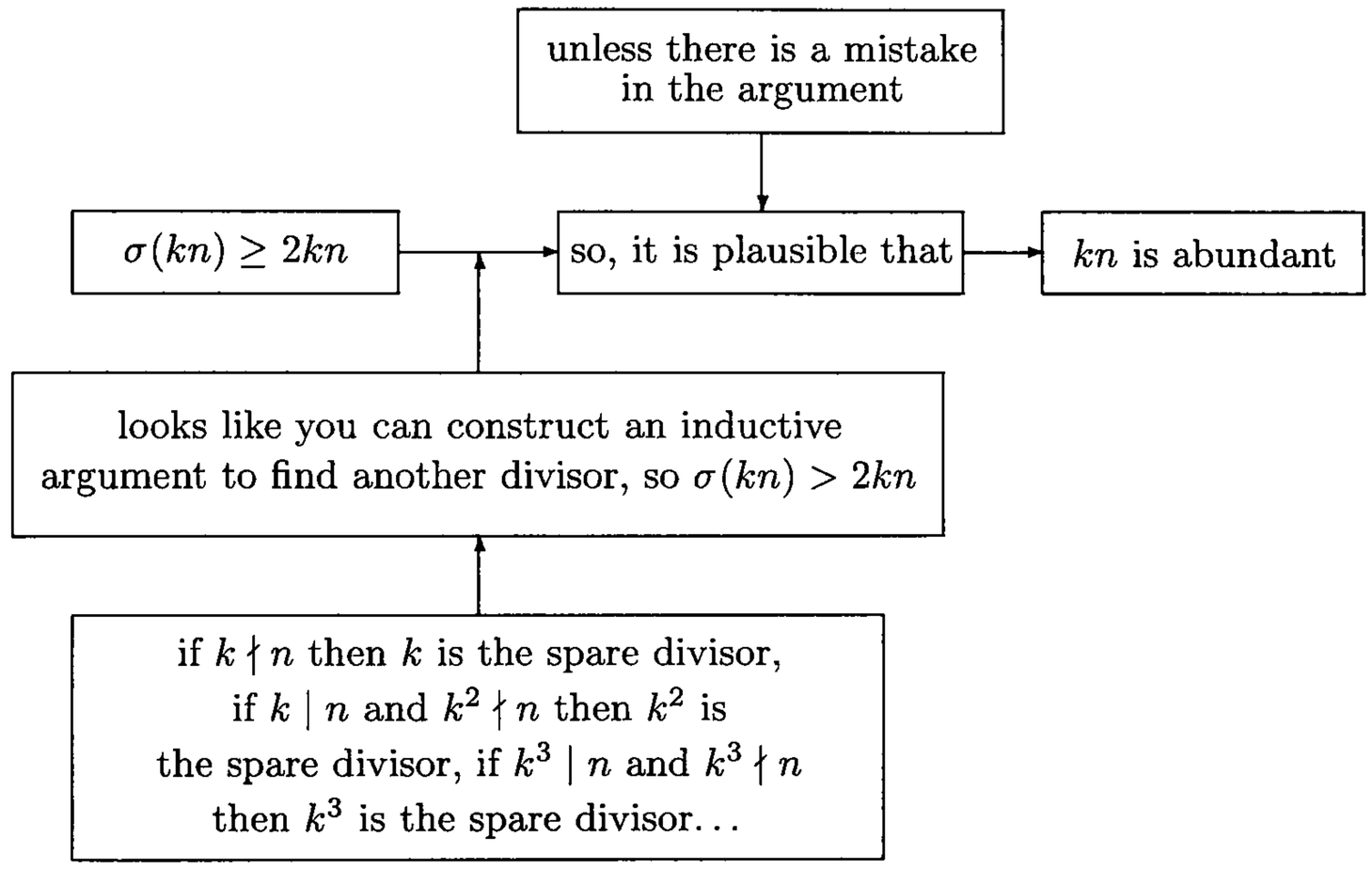


Figure 8.18: Part of David's response to Conjecture 2.

Andrew also exhibited some hints of a warrant-type based upon external conviction, when attempting to construct an odd abundant number he appealed to the fact that you can find large prime numbers that are close together:

ANDREW: So take some primes, a few odd primes that are very close to each other. I think so.

INTERVIEWER: Is that the way to do it?

ANDREW: If the differences in these primes is big, which is still the case here, because this is twice as big as this one, therefore these primes probably need to be really huge. I think there is some theorem, you can get these primes close enough, I don't know these theorems, but I'm pretty confident that you can find, these are never close these primes. (A5)

Here Andrew seems to be appealing to the authority of a theorem that he believes probably exists. He used this theorem to further justify his argument in Conjecture 5.

Note that the relative lack of externally based warrants in the data may be a function of the task design rather than the types of reasoning patterns used by mathematics research students. Andrew appealed to a theorem he believed probably exists, but presumably, if the task had been set within the context of Functional Analysis, his research area, he would have been able to appeal to many theorems that he knew *did* exist.

8.6.7 Discussion and summary of §8.6.

There are two major issues that should be drawn from the data presented in §8.6:

- The modal qualifier plays an important role in informal mathematical reasoning.
- The close relationship between the notion of warrant and proof scheme needs to be clarified, and the implications of this clarification discussed.

These issues have been discussed during the course of the presentation of the data, and will be recapitulated and expanded here.

The role of the modal qualifier.

One of the main research questions that this thesis set out to explore was that of how successful mathematicians deal with conditional statements. Rav (1999) correctly complained that current theories of informal mathematical reasoning are inadequate. This section has attempted to begin to characterise the conscious System 2 reasoning of mathematicians with regards to conditional statements.

Any theory of mathematical reasoning needs to model *both* informal and formal mathematical thought. Toulmin's (1958) argumentational scheme can do this. However, it cannot do it in the manner with which it has previously been applied to mathematics. Neither the approaches of Aberdein (2005, 2006) or of earlier mathematics education researchers (e.g. Hoyles & Küchemann, 2002; Krummheuer, 1995) successfully dealt with informal mathematical reasoning. Whereas the mathematics educators sought to eliminate the role of modal qualifiers, backings and rebuttals in mathematical argumentation, Aberdein sought to trivialise them by talking only of "necessary" qualifiers and "mathematics" as backings. The data presented in this section of this thesis suggests that neither of these approaches are viable.

Modal qualifiers that did not carry certainty were widely used by participants in Experiment 4. Non-formal explorations of the conjectures under study were vital in ascertaining whether the participant believed that the conjecture was

true or false. Without such conviction, participants would have no means of deciding whether to attempt a proof or a disproof.

It was found that certain types of warrants were paired with certain types of modal qualifiers. Non-deductive warrants were paired with non-certain modal qualifiers. To ignore this aspect of mathematical reasoning would be to deny a large part of the behaviour of the highly successful mathematicians recorded in Experiment 4.

The relationship between warrant-types and proof schemes.

The warrant-types discussed in this section are summarised in Table 8.2. The use of inductive, structural-intuitive and deductive warrants was widespread across the sample. The use of ritual and authoritarian warrants was rare, but this may have been a function of the task design. Recall that the quasi-judicial approach to data analysis does not seek to generalise from the observed data on the basis of a representative sample or a large number of participants. It seeks to establish reliability and validity through the rigorous application of the theory under inspection to the data. It is important to emphasise, then, that Table 8.2 should not be seen as an exhaustive taxonomy of warrant-types, but only as a summary of those warrant-types that were observed in the current study.[11]

Warrant-type	Appropriate Qualifier	Section
Inductive	Probable	8.6.3
Structural-intuitive	Probable	8.6.4
Ritual	Probable	8.6.6
Authoritarian	Probable	8.6.6
Deductive	Certain	8.6.5

Table 8.2: A summary of the types of warrants and modal qualifiers discussed in §8.6.

What then, is the relationship between the construct of a proof scheme as introduced by Harel and Sowder (1998), and the notion of a warrant-type as used in this thesis? Harel and Sowder (2005, p.33) write that "a person's proof scheme consists of what constitutes ascertaining and persuading for that person", it is the strategies and arguments which "an individual employs to remove her or his own doubts about the truth of an assertion" (Harel, in press). All the warrants and modal qualifiers in Table 8.2 were used to ascertain and

[11] Indeed it is not even clear that these warrant-types need be distinct. For example, Chris's argument, given in Figure 8.5, seems to combine both inductive properties (he quantitatively evaluated several examples) with structural-intuitive properties (he believes that number theory tends to be monotonic).

persuade the participant about the truth or falsity of the various conjectures. The key difference seems to be that, for Harel, a proof scheme is the strategies that a person uses to *remove* doubts, whereas a warrant may merely *reduce* doubts (although as seen in §8.6.5, a warrant may *also* remove doubts entirely).

All the participants, at various stages in their work used inductive, structural-intuitive and deductive warrants to gain conviction about the truth or falsity of mathematical statements. This conviction was often not certain: their doubts were often not *removed*, but instead were merely *reduced*. This, then, is the key distinction between Harel and Sowder's (1998) notion of a proof scheme and that of a warrant-type. However, once this broader notion – of the reduction of uncertainty rather than the removal of uncertainty – has been adopted there are implications for what it means to develop mathematically.

A new view of mathematical development?

When developing their proof schemes framework, Harel and Sowder (1998) argued that in order to succeed at advanced level mathematics students must abandon inductive, transformational and external proof schemes and instead adopt a deductive scheme:

> "[T]he goal of instruction must be unambiguous; namely, to gradually refine current students' proof schemes toward the proof scheme shared and practised by the mathematicians of today" (Harel, 2001)

Similar arguments have been made by other researchers. Tall (2004), for example, argued that as students deepen their cognitive development, their 'warrants for truth' (in the sense of Rodd, 2000) also deepen, hopefully with the result that formal proof becomes the only acceptable warrant when working in the formal-axiomatic world (Tall, 2004).

Whilst it is certainly true that no student will be successful at advanced mathematics if they accept a conclusion *with certainty* on the basis of non-deductive warrants, the data from this chapter indicates these non-deductive warrant-types play a crucial role in mathematical argumentation, *as long as they are paired with appropriate modal qualifiers.* When a person enters Tall's (2004) formal-axiomatic world, or when they develop Harel and Sowder's (1998) axiomatic-deductive proof scheme, rather than reducing the range of warrant-types they use, they retain the use of the warrants that have been used in previous 'worlds' or 'proof schemes', but *they qualify them appropriately.* These data show that mathematicians do not abandon inductive and intuitive arguments; instead, they learn to pair them with appropriate modal qualifiers and rebuttals. It is this pairing that is so crucial to successfully developing as a mathematician.

Inappropriate warrant-qualifier pairings.

There is evidence that this key skill – the ability to appropriately pair warrants with modal qualifiers – is not always present. In this section two examples taken from the literature which show students constructing inappropriate pairings are briefly discussed.

Weber (2003) reported a student's purported proof of the statement "for every odd integer n, $n^2 - 1$ is divisible by 8":

> "$1^2 - 1 = 0$ which is divisible by 8. $3^2 - 1 = 8$ which is divisible by 8. $5^2 - 1 = 24$ which is divisible by 8. And so on. Therefore if n is odd, $n^2 - 1$ is divisible by 8."

As a consequence of their use of an inductive warrant, Harel and Sowder (1998) would describe this student as having an inductive proof scheme, and in terms of Tall's (2004) framework they are yet to reach the axiomatic-formal world. In terms of Toulmin's (1958) framework for modelling argumentation, this argument has a modal qualifier which is inappropriately matched with its warrant. The key difference between the perspective developed in this thesis and that of Harel and Sowder and Tall, however, is that the use of an inductive warrant is not inappropriate per se. It is only when it is inappropriately paired with an absolute modal qualifier that the argument become problematic.

In this case, Weber's (2003) student inappropriately paired a non-deductive warrant with an absolute qualifier, but there are also reported cases of students doing the reverse: pairing deductive warrants with non-absolute qualifiers.

As part of his work on learning styles, Simpson (1995) discussed responses to the so-called 'Arithmagons' problem (Mason et al., 1982):

> "A secret number is assigned to each vertex of a triangle. On each side of the triangle is written the sum of the secret numbers at its ends. Find a simple rule revealing the secret numbers."

Simpson reported one student's behaviour:

> "Having been asked to prove a result which she had stated after some time working on the [Arithmagons problem] she wrote a quite delightful little proof which, though just essentially algebraic manipulation, made me feel that she had grasped the essence of the problem and gave a quite general solution.
>
> On the next page, she wrote 'I wonder if it works for big numbers?'"
> (Simpson, 1995; see also Duffin & Simpson, 1993).

This student, despite having presented an apparently perfect deductive proof, did not pair it with an absolute modal qualifier. For her, the deductive war-

rant she had written only allowed her to conclude that the statement was true about small numbers. The possibility that large numbers could form a rebuttal remained a concern for her.

In short, this student used a deductive warrant successfully, but was unable to qualify it suitably. The use of deductive warrants alone is not sufficient in advanced mathematics: they must be paired with appropriate modal qualifiers.

8.7 The modified Ramsey Test.

Looking back at the data presented in this chapter reveals one consistent theme: Participants evaluated their belief in the truth/falsity of the statement 'if P then Q' by evaluating what type of modal qualifier will allow them to reach the conclusion Q from the data P. In short, they assumed P was true, and evaluated their level of belief in Q. This is exactly the mechanism that lies behind the Ramsey Test (discussed in §3.1.8): Ramsey (1931/1990, p.247) wrote that, when arguing about the truth/falsity of 'if P then Q' participants hypothetically add P to their stock of knowledge, and argue on this basis about Q.

Ramsey's (1931/1990) model for evaluating conditional statements is almost a redescription of the manner in which Toulmin's (1958) argumentation scheme was used in §8.6. The data suggest the following model: participants hypothetically add P to their stock of knowledge, and look for a warrant with which to conclude (or deny) Q. This warrant is then paired with an appropriate modal qualifier, and Q is concluded with the degree of confidence given by this qualifier. The same level of confidence is given to the whole statement 'if P then Q'. In short, mathematicians implicitly conduct a kind of Ramsey Test to evaluate the level of belief that they have in the conditional statement. However, there appear to be some differences between the type of Ramsey Test that is used by mathematicians and that proposed by Over and Evans (2003) for standard indicative conditionals.

Over and Evans (2003) suggested that the Ramsey Test for everyday indicative conditionals takes place in two stages. The degree of belief[12] in $P \wedge Q$ and $P \wedge \neg Q$ are separately evaluated, and then compared. If $\mathbb{P}(P \wedge Q)$ is higher than $\mathbb{P}(P \wedge \neg Q)$ then $\mathbb{P}(Q \mid P)$ is high and, consequently, $\mathbb{P}(P \Rightarrow Q)$ is judged to be high. Similarly, if $\mathbb{P}(P \wedge Q)$ is lower than $\mathbb{P}(P \wedge \neg Q)$, then $\mathbb{P}(Q \mid P)$ is low and so $\mathbb{P}(P \Rightarrow Q)$ is judged to be low.

The crucial difference between the type of Ramsey Test described by Over and Evans (2003) and that used by the mathematicians in Experiment 4 is the the role of probabilistic warrants. Standard indicative conditionals, Over and

[12]Recall that the degree of belief a person has in event X is denoted $\mathbb{P}(X)$.

Evans suggested, are evaluated by judging $\mathbb{P}(Q \mid P)$, in line with the Ramsey Test. However, Over and Evans interpreted the function $\mathbb{P}(X)$ (degree of belief in event X) as being much closer to the actual formal probability of X than seems to be the case in mathematical contexts. For example, Evans et al. (2003) described an experiment where participants were given a pack of cards which contain a particular (named) distribution of colours and shapes. They were then asked "how likely" a series of conditional statements about the situation is. Evans et al. found that participants' evaluations are closely correlated with the formal probability of Q given P. Over and Evans suggested then, that there is a close correspondence between the functions "degree of belief in" and "the probability of".

It is clear that this reliance on pure probabilistic judgements did not happen in the mathematical context of Experiment 4. For example, every participant noted that Conjecture 1 was still false, despite being true for every multiple of 6 other than 6 (for example, see the transcript of Edward's response, p. 180). Even though the theorem had been established as being true for all natural numbers $k > 1$, the statement as a whole was still regarded as being false. Clearly this mathematical statement wasn't being evaluated on the purely probabilistic grounds that Over and Evans suggest for standard indicative conditionals.

The data reported in this thesis suggests that, in mathematical contexts, a modified version of the Ramsey Test is used. Instead of evaluating $\mathbb{P}(Q \mid P)$, mathematicians seemed to evaluate $\mathbb{P}[\mathbb{P}(Q \mid P) = 1]$. That is to say, rather than evaluating their degree of belief in Q given P, they evaluated their degree of belief in the claim that Q, given P, is certain.[13]

Consider again Conjecture 1 from Experiment 4. Edward had deduced that $6k$ is abundant for all $k > 1$, but that this was not the case for $k = 1$. Here P is the statement 'n is a multiple of 6' and Q is 'n is abundant'. The two Ramsey Tests in this case give different results:

- In the standard test $\mathbb{P}(Q \mid P)$ is evaluated. For this example this quantity is very high indeed: $\mathbb{P}(P \wedge Q)$ is considerably greater than $\mathbb{P}(P \wedge \neg Q)$.

- In the modified test $\mathbb{P}[\mathbb{P}(Q \mid P) = 1]$ is evaluated. For this example this quantity is zero: although $\mathbb{P}(Q \mid P)$ is high, Edward knew it is not 1 because of the case $n = 6$.

[13]Note that the way Evans et al. (2003) phrased their question prompted participants to respond in a probabilistic manner (they were asked "how likely" it was that 'if P then Q'). It is clear that the task used by Evans et al. doesn't really make sense in the context of the modified Ramsey Test: all the statements are trivially impossible to judge, as no information is given about the relations between the colours and shapes, participants only have frequency data. It is therefore unclear how mathematics students would respond to the materials used by Evans et al. (2003).

It seems clear that, in the mathematical context of the Abundant Number task, the modified Ramsey Test was used by the mathematicians. Further evidence to support this assertion was reported by Inglis and Simpson (2006) in the context of the Maze Task (see §3.2.1, Durand-Guerrier, 2003). Inglis and Simpson compared a mathematical and non-mathematical version of the Maze Task and found that the more mathematical the context, the more likely participants were to evaluate conditional statements with a modified Ramsey Test rather than the original version.

The notion of the modified Ramsey Test fits with evidence from other conjectures in Experiment 4. Take Chris's argument in Conjecture 5, for example (see Figure 8.11). In the language of suppositional conditional theory, Chris hypothetically added the belief that "n and m are abundant" to his stock of knowledge. He then argued, on this basis, about the conclusion "$n+m$ is abundant". He evaluated $\mathbb{P}[\mathbb{P}(Q \mid P) = 1]$ as being low (his modal qualifier was "it is unlikely that"), by using the warrant "the divisors of $n+m$ have nothing to do with the divisors of n and m". Note that, even if, for a large percentage of the abundant pairs (n, m), $n+m$ was abundant, this would probably not make Chris change his conclusion. As with the example of Conjecture 1, it is simultaneously possible to believe that $\mathbb{P}(Q \mid P)$ is high, but $\mathbb{P}[\mathbb{P}(Q \mid P) = 1]$ is low.

It is now possible to tie together the theories of Ramsey (1931/1990) and Toulmin (1958) in the context of mathematics: When a mathematician evaluates an "if...then" statement they set up an argument in the sense of Toulmin with P as the data and Q as the conclusion. They then judge the modal qualifier of the argument using a modified version of the Ramsey Test. Instead of evaluating $\mathbb{P}(Q \mid P)$, they evaluate $\mathbb{P}[\mathbb{P}(Q \mid P) = 1]$. The evaluation is based upon the *type of warrant* used to 'transmit' validity from P to Q.

The important question that this suppositional structure raises is: how does a person find their warrant? Where do they begin to look? The evidence presented in this thesis suggests that part of the answer can be found in the notion of *preconscious heuristics*. Drawing together the strands of research presented in the thesis is the goal of the next, and final, chapter.

8.8 Summary of Chapter 8.

This chapter has looked at two distinct stages that are important in understanding how mathematicians use and evaluate conditional statements. Following on from Chapter 6, where it was experimentally demonstrated that Evans's (1996) heuristic-analytic dual process theory can successfully account for the behaviour of mathematicians when dealing with the Wason Selection Task, the

same framework was used to study how conditionals are evaluated in a more realistic mathematical context. The main components of the case-law developed in this chapter were:

- Preconscious System 1 heuristics have a major role in directing the attention of mathematicians when they encounter conditional statements. This can, and does, influence their choice of proof strategies.

- To accurately model mathematical argumentation it is necessary to admit the role of the modal qualifier and rebuttal in Toulmin's (1958) argumentation model. Expert mathematicians appear to accurately match up the warrants they use with appropriate modal qualifiers.

- The notion of warrant-type is broader than that of proof scheme. A warrant-type does not necessarily remove *all* doubts, it may only remove *some* doubts. The use of warrants that could be associated with 'inappropriate' proof schemes is not inappropriate as long as they are qualified suitably. The goal of instruction should be to help students better tie together these warrant-types with appropriate modal qualifiers.

- Mathematicians evaluate mathematical conditionals using a modified form of the Ramsey Test. Rather than evaluating $\mathbb{P}(Q \mid P)$, they evaluate $\mathbb{P}[\mathbb{P}(Q \mid P) = 1]$.

In the concluding chapter that follows the evidence from this and previous chapters is reconsidered, and the roles of preconscious heuristics, warrants and the modal qualifier and the Ramsey Test are drawn together to develop one coherent theory of mathematical conditional evaluation. Particular attention is given to distinguishing between the 'intuition' associated with structural-intuitive warrants and the 'intuition' associated with System 1 preconscious heuristics.

Chapter 9

The Theory

The research question that this thesis set out to answer was a direct response to Rav's (1999) observation that

> "As things stand now, we have remarkable mathematical theories of formal logic, but inadequate logical theories of informal mathematics." (p.14)

Specifically, the primary goal of the thesis was to develop a model of how mathematicians evaluate mathematical conditionals. To do this two quite separate strands of research have been reported in the thesis so far. The purpose of this, the concluding chapter, is to summarise the content of the experimental work reported in earlier chapters, and to synthesise the two strands into one coherent theory.

9.1 Summary of empirical work.

9.1.1 Picking a framework.

The experimental work in the thesis began by attempting to distinguish which of the various theories of reasoning discussed in Chapter 4 is most applicable to studying how mathematicians reason with conditional statements. To this end a direct comparison between mathematics students' and history students' responses to the Wason Selection Task was conducted (Experiment 1). The Wason Selection Task has, historically, been the instrument which has proved to be most insightful for reasoning researchers, and the interesting (and surprising) results from Experiment 1 confirmed this.

The results of Experiment 1 were unexpected: whereas mathematicians did select the normatively correct answer more often than the control group, their

success was not overwhelming. Moreover, the typical mistakes made the mathematics students were *different* to those of the control group. Experiment 2 followed this finding up and demonstrated that success on the Selection Task is not related to a mathematics undergraduate's experience of, and attainment in, degree level mathematics. Furthermore, it was found that the time taken to solve the task by participants who answered correctly was significantly greater than those who made either the typical mathematician's mistake or the typical general population mistake, but that there was no significant difference in length of timings between these latter two groups. The results of Experiments 1 and 2 provided a challenge to the major theories of reasoning. In Section 6.3 the adaptations necessary to the various theories in order to accommodate the experimental findings were discussed. It was concluded that each of the mental models, mental rules and heuristic-analytic dual process theories could be adapted successfully.

In the final experiment in Chapter 6, Experiment 3, the mental models and mental rules theories were eliminated. Based on the work of Evans (1996) and Ball et al. (2003), an inspection time eye-tracker methodology was used to demonstrate that the mathematics students are preconsciously biased towards the same cards as the general population, but that they are substantially better at analytically considering the cards they are biased towards and overriding these biases. It was argued that neither the mental models not the mental rules theories can completely account for these data.

As a consequence of the empirical work in Chapter 6, only one of the major theories of reasoning can be said to efficiently account for the behaviour of successful mathematics students on the Wason Selection Task. This is an important result both in the context of the goals of this thesis, and of the wider psychology of reasoning literature.

9.1.2 Applying the framework.

Having demonstrated that only one theory of reasoning can account for the empirical data collected during Experiments 1, 2 and 3, the next stage of the empirical work conducted in this thesis was to apply this framework to more genuinely mathematical contexts. This was the goal of Experiment 4.

As befitting an experiment designed with the dual process framework in mind, there were two broad stages to the analysis of Experiment 4. The first demonstrated that preconscious heuristics do play an important role in directing attention during mathematical work. It was shown that careful manipulation of apparently irrelevant surface linguistics content can, as a consequence of the if- and matching-heuristics, change an apparently simple problem to a surprisingly

difficult one.

In the second stage of the analysis, conscious processes involved in the evaluation of mathematical conditionals were considered. In line with Bromley's (1986) quasi-judicial approach, each participant's interview was considered as a case study, and analysed using Toulmin's (1958) framework, with Harel and Sowder's (1998) 'proof schemes' theory as a prima facie explanation.

It was suggested that, when evaluating mathematical conditionals, the role of the modal qualifier has been seriously underestimated by earlier researchers. Furthermore, in Section 8.7 it was argued that, when evaluating conditionals, successful mathematics students search for warrants using a modified version of the Ramsey Test (Ramsey, 1931/1990; Evans & Over, 2004).

The goal of this, the final chapter, is to integrate these two strands of empirical work – corresponding to preconscious System 1 thinking, and conscious System 2 thinking – together into one coherent theory. This process begins by exploring the role of preconscious heuristics in warrant finding.

9.2 Preconscious heuristics and warrant finding.

The two strands of argument developed in Chapter 8 can now be drawn together to explore how preconscious heuristics interact with the modified version of the Ramsey Test. In the context of the standard Ramsey Test, Over and Evans (2003) wrote:

> "There are a number of ways in which people can [conduct the standard Ramsey Test]. Sometimes they will know relevant frequency information, and they can use that to make an explicit comparison. [...] More widely, [preconscious] heuristics will sometimes be engaged." (p.346)

The evidence from Experiment 4, reported in §8.5, suggests a similar role for preconscious heuristics as that postulated by Over and Evans: the if- and matching-heuristics direct attention towards considering apparently relevant features of the statement, and this affects where the participant looks to find a warrant, which in turn affects the type of warrant they may use.

In short, the evidence presented in this thesis strongly supports the claim that preconscious System 1 heuristics play a major role in directing attention during the evaluation and use of conditional statements in advanced mathematics. Experiment 4 showed that changing the surface linguistic features of conditional statements can dramatically alter the manner in which they are dealt with by mathematics students. The following terminology was adopted in this thesis to aid classification of conditional statements:

Criterion T. A conditional satisfies Criterion T if it is phrased in such a way as to ensure that people tend to be preconsciously biased towards considering features of the situation which will hinder them finding warrants.

Criterion U. A conditional satisfies Criterion U if it is phrased in such a way as to ensure that people tend to be preconsciously biased towards considering features of the situation which will help them find warrants.

In Experiment 4 it was shown that, apparently irrelevant surface content can significantly influence whether a mathematical problem is either a Criterion U or a Criterion T situation.

Previous research on dual process theories of reasoning, reviewed in Chapter 7, have used comparatively trivial everyday tasks and tests to investigate the bimodal structure of reasoning posited by dual process theorists. As Leron and Hazzan (2006) pointed out, evidence that similar heuristics also play a role in higher level mathematical reasoning will come as a surprise to many. Traditionally higher level mathematics is seen as being abstract and formal, it is a situation in which System 2 responses would be expected to dominate. That System 1 preconscious heuristics play an important role in the adoption of proof strategies – which sometimes leads to non-normative responses – is an important finding. This thesis, then, supports and extends the argument put forward by Leron and Hazzan (2006).

However, whilst preconscious heuristics have been shown to play an important role in allocating attention, they should not be confused with the notion of structural-intuitive warrants introduced in §8.6.4. Clarification of the differences between these somewhat similar constructs – preconscious heuristics on the one hand, and structural-intuitions on the other – is the goal of the next section.

9.3 System 1 heuristics and intuition.

The form of the dual process framework that has been adopted in this thesis is Evans's (1984, 1989, 1996, 2006) heuristic-analytic theory. Evans argued that there are two stages of reasoning which he originally called *heuristic* and *analytic* (Evans, 1984). In the first stage of reasoning preconscious heuristics select certain aspects of the environment as being *relevant*. In the second stage conscious analytical attention is 'spent' on these relevant parts of the environment. This heuristic-analytic account has since become incorporated into the standard dual process framework, and the more generic terms System 1 and System 2 have overtaken the heuristic-analytic terminology.

The role of preconscious heuristics, in the heuristic-analytic theory, is to direct attention towards relevant parts of the environment. Specifically, two heuristics appear to be of importance when considering conditional statements:

- The if-heuristic directs attention towards considering the case when the antecedent of the conditional is true.

- The matching-heuristic directs attention towards the surface linguistic content of the rule rather than the semantic meaning of the terms in the rule.

These heuristics, when seen from an ecological point of view, are helpful, they direct attention towards parts of the the environment that are most relevant. However, as this thesis has demonstrated, in unusual situations they can be unhelpful: Chapters 6 and 8 gave numerous examples of mathematical Criterion T situations: where the reasoner is led to consider unhelpful aspects of the environment which disguise normative solutions.

In the mathematical suppositional conditional model discussed above the role of System 1 preconscious heuristics is quite different to that of structural-intuitive warrants. Although Fischbein (1987) and Schmalz (1988) saw 'intuitions' as being "immediate", "certain" and standing in opposition to rational thought – all characteristics of System 1 heuristics – they play a different role.

Structural-intuitive warrants are the result of intuitions about the structure of a person's internal representations. Intuitions, in this sense, are about beliefs resulting from internal representations, they are not about the way in which attention is directed to find these beliefs. Burton's (2004) participants, all professional mathematicians introspecting about their work, described the role of intuition:

- "I don't think you would ever start anything without intuition."

- "[Intuition is] a sense of the possible or even likely."

- "You are always thinking in a not-straightforward, deductive way. You are always looking for some hint from within that this might be an interesting thing to do" (pp.76-77).

These characterisations of intuition fit with this thesis's notion of structural-intuitive warrants: when looking for a structural-intuitive warrant you are looking for a "hint from within" about whether a conclusion is "possible or even likely", without this kind of warrant it is unlikely you "would ever start" a proof attempt. These characteristics, however, do not fit with the Dual Process idea of a preconscious heuristic. Intuition is *what* you are looking for, but *where* you look is determined by preconscious System 1 heuristics.

9.4 The evaluative model: A summary.

The ground has now been prepared sufficiently for the statement of the evaluative model of mathematical conditional evaluation as developed in §8.7-9.3.

Evaluating mathematical conditionals.

The evaluative model posited by the theory developed in this thesis is shown in Figure 9.1.

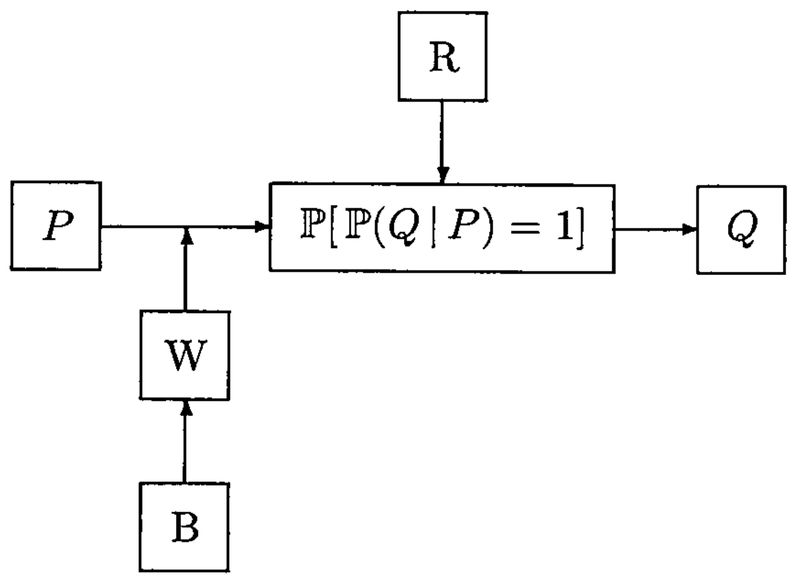


Figure 9.1: The model: how mathematicians evaluate mathematical conditionals.

To evaluate a conditional "if P then Q" the mathematician hypothetically adds P to their stock of knowledge, by putting P as the data in Toulmin's (1958) argumentation scheme. They then attempt to construct an argument where Q is the conclusion. Based on System 1 preconscious heuristics, the mathematician's attention is directed to certain parts of the environment and they attempt to evaluate the modal qualifier using a modified Ramsey Test: $\mathbb{P}[\mathbb{P}(Q \mid P) = 1]$ is the level of belief the mathematician has in the claim that Q is certain, given P. This evaluation is based on the type of warrant that the mathematician finds during their investigation. The conditional statement "if P then Q" is then concluded with the level of certainty given by the modal qualifier.

When a conditional is evaluated using a deductive warrant no rebuttal is admitted and $\mathbb{P}[\mathbb{P}(Q \mid P) = 1] = 1$. However, the situation is more complicated for a contrapositive deductive warrant. In these cases the backing is an entirely new argument, complete with its own data, conclusion, warrant and so on. This structure is shown in Figure 9.2.

The evaluative model versus formal logic.

The model outlined above is not merely a rephrasal of formal logic. Whereas Inhelder and Piaget's (1958) model of reasoning (at least in the stage of formal

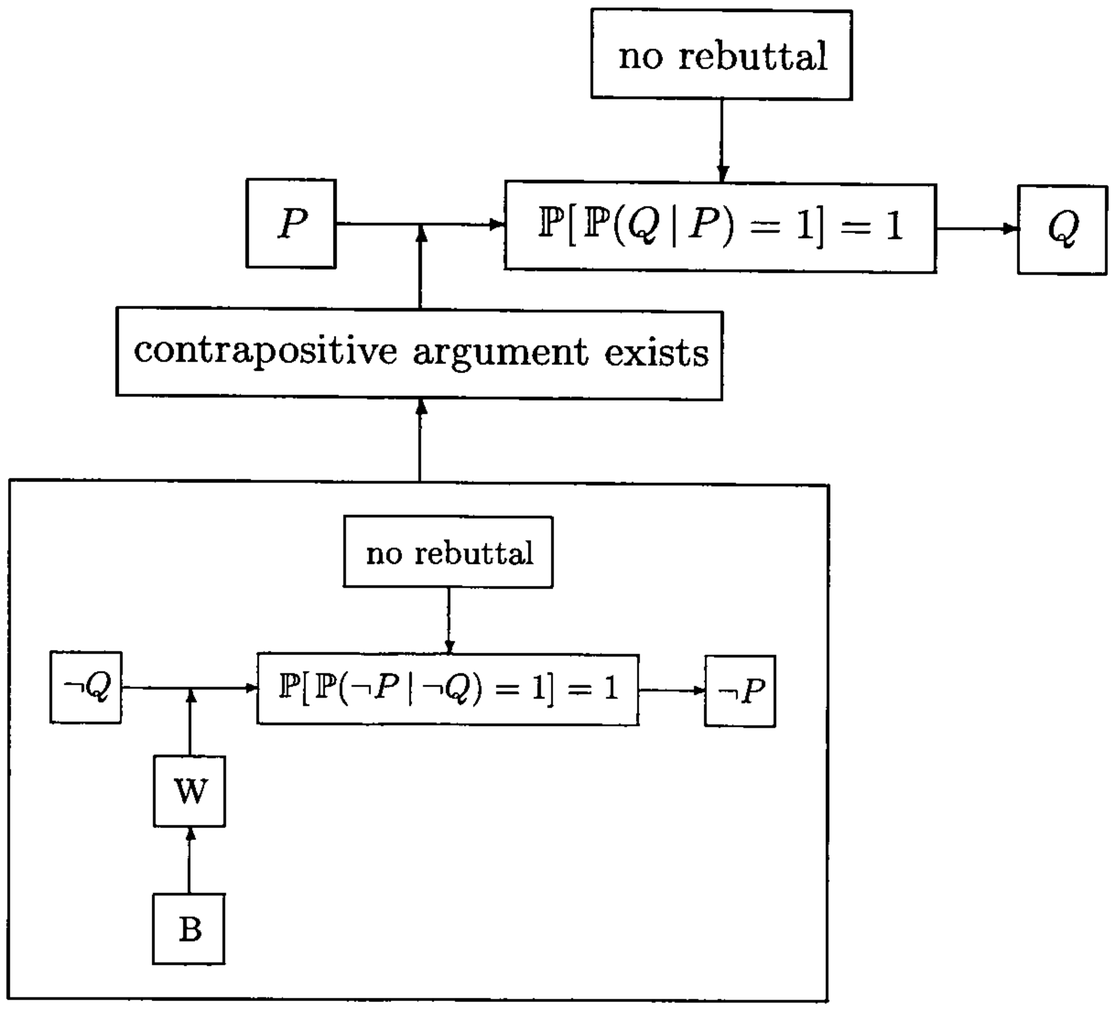


Figure 9.2: The model: how mathematicians evaluate contrapositive arguments.

operations) was overtly logical, the model proposed here has several important differences:

- *The role of the modal qualifier.* In formal logic "if...then" statements are either true or false. In the evaluative model developed in this thesis conditionals admit degrees of confidence. Mathematicians can, and do, believe that certain conditionals are 'almost certainly true', 'probably true', 'plausibly true' and so on.

- *The role of preconscious heuristics in directing attention during the warrant search.* In the evaluative model, the various warrants that are possible when proving a statement can be hidden or highlighted depending on the interaction between the statement's linguistic structure and the mathematician's preconscious biases. A prover's attention may be directed towards looking for a warrant in helpful or unhelpful places; this can, and does, highlight certain warrants and hide others. Certain arguments can therefore be privileged over others.

- *The difference in status between direct and indirect arguments.* In formal logic there is no difference in 'status' between a direct and an indirect proof of a statement. They both establish the result with certainty using one logical step (modus ponens or modus tollens). In the evaluative model, however, this is not the case. Establishing a result indirectly involves

a complex argument structure with a whole new argument forming the backing to the (deductive) warrant that establishes the conclusion with certainty.

The evaluative model, then, is entirely different to formal logic: it posits a process by which mathematicians evaluate conditional statements which admits the possibility of degrees of confidence, and in which the evaluative process is significantly affected by preconscious System 1 heuristics.

The evaluative model versus everyday reasoning.

As well as being a different model to the formal logical theory of Inhelder and Piaget (1958), the evaluative model developed here is also at odds with current models of everyday reasoning with conditionals. The evaluative model is compatible with none of the models discussed in Chapter 3. Instead it is an adaptation of the suppositional conditional model described by Edgington (2003), Evans and Over (2004) and Over and Evans (2003). To re-emphasise the key differences between these two theories:

- The suppositional model of *indicative* conditionals suggests that a person fixes their degree of belief in a conditional statement 'if P, then Q' by considering their degree of belief in Q, given P.

- The evaluative model for *mathematical* conditionals suggests that a person fixes their degree of belief in a conditional statement 'if P, then Q' by considering their degree of belief in Q being certain, given P.

- Symbolically, the suppositional model for indicative conditionals argues that $\mathbb{P}(P \Rightarrow Q) = \mathbb{P}(Q \mid P)$, whereas the evaluative model for mathematical conditionals suggests $\mathbb{P}(P \Rightarrow Q) = \mathbb{P}[\mathbb{P}(Q \mid P) = 1]$.

However, notwithstanding these differences the overall structure is the same. When evaluating an 'if...then' statement an argument is constructed with the antecedent as the data and the consequent as the conclusion. A warrant is sought to bridge between these two parts of the argument and an appropriate modal qualifier for this warrant gives the degree of belief in the conditional.

9.5 Speculations on the sources of the differences between mathematical and general cognition.

As discussed above, the evaluative model developed in this chapter is different to that proposed by Over and Evans (2003) for the general population. This

begs the question: what causes this difference in argumentation practice, and to what extent does it extend to non-mathematical settings? Do mathematicians evaluate day-to-day indicative conditionals, such as (∗) below, using the Ramsey Test or the modified Ramsey Test?

(∗) If you're in Coventry, then you have a good choice of Indian food.

Common sense seems to suggest that mathematicians are only mathematicians 'in the office': that they would not reject (∗) simply because of the possibility that you could be in a distant suburb of Coventry where there are no Indian restaurants. However, there is currently no empirical evidence upon which to base this common sense impression (see also Inglis & Simpson, 2006). More research is needed on individual and contextual differences in indicative conditional evaluation. However, some care is needed when designing such experiments. As discussed above, the wording of Evans et al.'s (2003) materials[1] may bias participants into evaluating conditionals using the standard Ramsey Test rather than the modified version. Experimental materials which do not privilege one version of the Ramsey Test over another will be required to successfully explore this research question.

The source of the differences in behaviour of mathematicians and non-mathematicians on the Wason Selection Task is also of interest. On this task, recall, mathematics students were biased in the same manner as the general population, but were significantly better at analytically considering the cards they were biased towards and detecting mistakes. Further, it seemed that the magnitude of difference was not related to the students' degree of experience of, or attainment in, advanced mathematics.

How mathematics students cope with preconscious biases is an area of reasoning research that appears to be underdeveloped. There is, for example, some evidence which suggests that mathematics students are significantly less affected by belief bias on conditional inference tasks to the general population (Inglis, 2006). But what is the cause of these differences?

There are two reasonable hypotheses:

- Mathematics students gain superior analytical skills during their mathematical studies which allow them to override misleading preconscious biases.
- Some parts of the population are innately better the kind of analytical cognition which is necessary to override misleading heuristic biases, and people with this characteristic are more likely to be filtered into mathematics education.

[1] Participants were asked to judge "how likely" a conditional was.

The first hypothesis would fit uneasily with the finding that undergraduate experience has no relation with success on the Selection Task, but it may be possible that the skill required to conduct this analysis is relatively minor – Chapter 6 suggested that it may merely be the recognition that $P \Rightarrow Q$ is not the same as $P \Leftrightarrow Q$ – and that this skill might be developed in pre-university courses. Research evidence from studies that have looked at the effect of intelligence and education on reasoning is mixed (e.g. Lehman, Lempert, & Nisbett, 1988; Lehman & Nisbett, 1990; Stanovich, 1999).

In order to distinguish between the two hypotheses outlined above further empirical work is needed. Alongside the longitudinal study that would be necessary to answer such questions, the work contained in this thesis raises many other, currently open, questions. For example, how are mathematics students affected by negative conclusion bias? Do the findings regarding belief bias reported by Inglis (2006) extend to the more traditional setting of Aristotelian syllogisms? Do mathematicians interact with heuristics more associated with the decision making literature (as discussed Chapter 7) in the same way as they seem to with those more associated with the reasoning literature? There appear to be many unanswered questions.

9.6 Final remarks.

The purpose of this thesis was to respond to Rav's (1999) comment that there is currently an inadequate understanding of the logic of informal mathematics. The work reported here has contributed to answering this criticism of the literature by developing a coherent empirically based model that accounts for how mathematicians evaluate mathematical conditionals. The goal of future research should be to further explore and refine the evaluative model, and to begin to approach the important questions that this thesis raises with regards to the relationship between individual differences in reasoning behaviour and the study of advanced mathematics.

Appendix A

Transcript of David's interview.

DAVID: [*reads question*] OK, OK, my goodness, a number's abundant if and only if it's a multiple of 6. Well, I guess 6 is perfect right? [*laughs*]

INTERVIEWER: You just knew that?

DAVID: Well yeah I knew that anyway, so umm, if something's a multiple of 6 then you've got, what have you got? so if you've got 6n then umm, err as divisors you've got of course 1, 2, 3 and 6, and n of all these, n, $2n$, $3n$, $6n$, so the sum of all them is going to be $12n$, is that right? yeah. So, but you may have more divisors I guess, cos you, so, so you've got at least... so maybe, yeah so it may be perfect I guess...I mean 6 is a multiple of 6 and that's perfect [*laughs*]

INTERVIEWER: So what does that tell us?

DAVID: Well that tells us it's false.

INTERVIEWER: That tells us the whole thing is false?

DAVID: Yeah, because if and only if it's a mulitple of 6, so that's a counterexample. So, well, it says 6 is a counterexample to the fact that a number is abundant, you know 6 is not abundant.

INTERVIEWER: Right, so that discounts the whole conjecture, but if you had to split it up into an if and an only if what would that tell us?

DAVID: Umm, er, so a it is a multiple of 6 then it's abundant, well, I mean that's not true right?

INTERVIEWER: Right, so what about the other direction?

DAVID: If it's abundant then it's a multiple of 6, well umm, I guess you just need to find a counterexample to that. Let's have a look, have we got any examples down here?

INTERVIEWER: I've got some examples for you to save having to do any work, so there's some examples.

DAVID: So, the first few abundant numbers, so 20 that's not a multiple of 6.

INTERVIEWER: So that's dead easy then? That's the end of the question?

DAVID: That's the end of the question.

INTERVIEWER: Marvellous, ok, let me give you another one then.

DAVID: If n is perfect then kn is abundant for any k. Right, ok, so we know that's it's not deficient, that's by the same argument that we have down there, so, why would it be abundant? So you've got, n is perfect, so we've got a new n there, and umm, so this will have some divisors, I don't know, sum of the divisors, I don't know how to write the divisors [*laughs*], sum of, what shall I call them? m in divisors of n, is err, equal to $2n$. And then we've got sum of m in the divisors of kn, right we know that it's greater than or equal to, to 2 of kn by the same argument.

INTERVIEWER: Sorry, why is that?

DAVID: That's by this argument again, you see, if you've got umm, so the sum of the divisors, so each divisor of err, each divisors of n, well k times each divisor of n is a divisor of kn right? so therefore this is true. Why would it be a greater than? Umm, I don't know, why couldn't it be perfect? I mean you've got some possible counterexamples here, I mean, we might look for one of them, so does 6 divide into that? I don't know, no it doesn't does it? So does 6 divide into the next thing? So, I can't see any counterexamples there, and for example. So I guess, umm, what was we, what would I, umm, we need to find some divisors that aren't of the form $2m$ for m a divisor a n, don't we?

INTERVIEWER: $2m$?

DAVID: Sorry, km, I'm thinking of that, it's km. So if m is a divisor of n, and, I guess k [*laughs*].

INTERVIEWER: k?

DAVID: Well, k's not necessarily going to be a divisor, so k's going to be a divisor of k, so k divides kn, ok? And k may or may not be a divisor of n. But that doesn't prove it [*laughs*]. If n is perfect then kn is abundant. Err, so that proves it for any k that doesn't divide n. So, let's write that. So if k doesn't divide n, then umm, kn, err, whatever, k divides kn, and err, well, it would take me a long time to write it down, I don't know if you want me to?

INTERVIEWER: No, don't worry

DAVID: OK, so k doesn't divide n, so suppose k divides n. Umm, I'm trusting this is all correct, because I haven't written it down properly [*laughs*]. But this is where I'm going, suppose k divides n, then, well suppose k^2 doesn't divide n, then k^2 would divide kn.

INTERVIEWER: Oh, sorry, you'll have to...

DAVID: Well suppose k divides n and k^2 doesn't divide n, umm, then k^2 divides kn, so that's a number that would make this strictly greater than, and then I guess you're looking at some sort of induction thing, so suppose k^2 divides n but k^3 doesn't divide n, k^3 divides kn erm, and umm,

INTERVIEWER: Can you assume that though? Can you assume you've got a k that divides n?

DAVID: Well, no, just suppose it does so what you're looking at is you're saying let k be a natural number, err, that well k is some natural number, if k doesn't divide n then the theorem's true I think, if it divides n but k^2 doesn't divide n then the theorem's true, if k^2 divides n but k^3 doesn't divide n then the theorem's true. And I think you can just carry on, and k's either going to umm, err, well I mean hopefully, have we done everything now?

INTERVIEWER: Have we?

DAVID: Yeah, let's see, probably fundamental theorem of arithmetic or something. So if k divides n then the primes of k are in the primes of n, err, yeah, umm, I think err, if k divides n, I think that's right, cos then we just need to consider that case, but we get left over with that case and then,

INTERVIEWER: you can keep going for ever?

DAVID: Inductively, that will probably work. Is there a really simple argument? [*laughs*] Have I missed that?

INTERVIEWER: Well, no, no, I mean the goal is not to find the best answer, the goal is to see how people go about it.

DAVID: Yeah, well I assume that works, as I say, I haven't written it down and I would always say to my students you know, I've got, this seems like an idea, I now think I believe this conjecture, you've now got to sit down and try and write it out. But I wouldn't want to do that and say it at the same time, because it would, you know, you'd have to do that on your own. But it looks plausible I think.

INTERVIEWER: Sure, right, the only thing that I would think, what would happen if k equalled k^2 equalled k^3?

DAVID: Well then k is 1, so it doesn't, so that's just if n is perfect then, oh well, so that's [*laughs*] so that's I guess actually that's if n is perfect

then kn is abundant, this has got to be false for $k = 1$, so actually I could have just said, if I'd have realised that that's false, but possibly for k above 1 this is true, and if k is above 1 then you haven't got this problem. So if $k = 1$ then it's false, yeah.

INTERVIEWER: But otherwise we think it's true?

DAVID: It *looks* like it's true [*laughs*].

INTERVIEWER: OK, marvellous let me give you another one then.

DAVID: OK, if p_1 and p_2 are primes then $p_1 p_2$ is abundant. Umm, well that can't be true, because 2 and 3 are primes and 6 is perfect [*laughs*].

INTERVIEWER: Right, ok, so that's the end of that?

DAVID: That's the end of that [*laughs*].

INTERVIEWER: If I changed it then, to be not abundant, what would you say?

DAVID: That would seem more reasonable. Because primes look very deficient.

INTERVIEWER: Do they?

DAVID: Well, they only have 1 and themselves as divisors, so they're about as deficient as you can get right? Except you know... so, you know, p_1, p_2, umm, well ok, if it's deficient, if it's not abundant, so it might be perfect for 2 and 3. So I guess you can do a kind of direct proof. Because $p_1 + 1$, sorry, $p_1 p_2$ has as divisors, p_1 and p_2 and $p_1 p_2$, and er, 1. So the only way in which that could be greater than or equal to $2p_1 p_2$ is if $p_1 + p_2 + 1$ is greater than or equal to $p_1 p_2$. And err, ok, so sorry we're trying to show that that's false right, for 2 and 3 that doesn't work, and for any higher? Well, if p_1 and p_2 are any higher than 2 and 3 then you should be able to do some induction thing here. So this is false for 2 and 3, so you could do induction on p, I mean there's two things, there's two sort of induction things isn't there. So, if you add 1 to p_1 then it's going to add more than 1 to the right hand side, and if you add 1 to p_2 then it's going to add more than 1 to the right hand side. So you're never going to get this inequality.

INTERVIEWER: So it's kind of a get's bigger faster argument?

DAVID: Yeah, this gets bigger faster than that, so you'd need to do an induction on p_1 and then say well, ok, right

INTERVIEWER: That's going to be quite tedious though

DAVID: What's that?

INTERVIEWER: That's going to be quite tedious to do that.

DAVID: Well, no, it's a proof by induction, but I'm not gonna. That I'm very convinced that that is completely false.

INTERVIEWER: So, let's just recap then, so this conjecture as it stands, as it's written is false?

DAVID: Yeah.

INTERVIEWER: But if you change it to not abundant it's?

DAVID: It's probably a typo [*laughs*].

INTERVIEWER: Let me give you another one.

DAVID: If n is deficient then every divisor of n is deficient, so the sum of the divisors is, right, err, sum of m in divisors of n, right, let's just write this down as less than or equal to $2n$. And if you've got, err, so if you consider, hmm. Umm, if you consider a divisor of n, itself then you've got, erm [*long pause*]

INTERVIEWER: What are you thinking?

DAVID: What am I thinking?! I'm not, I'm not sure, I'm trying to take away a term from that, but you can't just take away a term because you might end up with too many. OK, so let's, a good way to do this is sort of fundamental theorem of arithmetic, so n can be written as p_1 up to p_s, not necessarily unique primes, and you're saying that two of these, so you've got, well certainly this is true, this isn't all the divisors of course, this is just the prime divisors.

INTERVIEWER: [*Reading what David has written*] So you've got the sum of all the prime divisors is less than 2 times $p_1 p_2 \dots p_s = n$.

DAVID: And then a divisor of n is going to be, well you remove at least one prime, so let's do an induction thing, so you're subtracting p from that side and you're subtracting some p_i from that side. Umm, and erm, then you're dividing by, so on this side you're, [*long pause*] hmmm.

INTERVIEWER: Sorry, I'm a bit lost now.

DAVID: Yeah, so I'm just saying that, maybe I'm not going about this in a good way. I'm just saying that this is certainly less than the sum of all the divisors, the sum of all the primes. So I'm just trying to write this out, so therefore this $[p_1 + \dots + p_s]$ is certainly less than that $[p_1 \dots p_s]$. And I'm just seeing, well, what is a divisor of n? Well, a divisor of n is obtained by taking away some of the primes. So just taking away one prime, and then saying let's see what happens. But I'm kind of a little bit worried, because, you know, I haven't got enough on that side really, so, I haven't got all the divisors on the left hand side. So I'm not sure if I want to proceed, so I'm kind of pausing.

INTERVIEWER: You've smelt a rat with that technique.

DAVID: Yeah, so if n is deficient then every, ok, so let's think primes again, primes are very deficient, what was that previous one? if p_1 and p_2 are primes then p_1p_2 is erm, not abundant. So, it's either err, so we write our n as a product of primes, p_1 to p_s, and each of these are, each of these are [*long pause*] err, oh hang on, that only says p_1, p_2 doesn't it? That doesn't help, it won't give us an induction.

INTERVIEWER: Oh, I see, does that extend to 3 then? If you put a p_3 in there?

DAVID: Well, it wouldn't necessarily extend, I mean look at 2 times 2 times 3, it's 12 which is abundant. I'm just trying to think about going the other way, rather than starting with n start with some divisor on n, and then see if we can say, OK, so I guess what I'm trying to do is the contrapositive. So I'm saying, umm, suppose every divisor of n is not deficient, so suppose this is not deficient and then add a prime and see if you get something that's still not deficient? Is that right? Is that what I'm trying to say? If every divisor of n is not deficient then n is not deficient. Umm.

INTERVIEWER: So what's the contrapositive?

DAVID: That's what I've just said. If every divisor of n is not deficient then n is not deficient. So, this is umm, p_1 to, why is this taking me so long?

INTERVIEWER: No this is good.

DAVID: So if p_1 to p_n is not deficient, so this is umm, [*long pause*] so, oh, I've got that written. So this is where I was, so let's multiply both sides by another prime. If we multiply that by another prime, and then we want to show that this is, so we're saying that this is not deficient, right m divisors of m is not deficient, so it's greater than or equal to $2n$. And if we multiply by a prime.

INTERVIEWER: So n here is now a divisor?

DAVID: Is now a divisor yeah, sorry. Different n. Now if we multiply by a prime what do we get? Err, do we get that this is, oh hang on, what have I done? I only want to multiply, I want to multiply this side by a prime, np, I'm multiplying n by a prime. And then we want, so we've got every divisor, inside this sum, we've got every umm, mp from this m up here, so this is, and we may have more, so this is greater than or equal to, too many crossings out around here, so my signs are going, m in divisors of, oh, hang on, what am I saying? The sum of m in divisors of, so I'm trying to use this thing up here, in np, is greater than or equal to, err, the sum of k, erm, sorry, not

k, pm. m is a divisor of n, which is greater than or equal to $2np$. Right? So we've assumed that and then got...

INTERVIEWER: So we've assumed the divisors are deficient.

DAVID: Yeah, so pick some divisor n, and then umm, err, multiply it by p, so you'd need to do some induction to get back up, so assume all the divisors are deficient, so assume one of them's deficient, then by an induction step np must be deficient, so you're original n whatever it was is also deficient. And since you've just picked any divisor, it works for every single one.

INTERVIEWER: So why have you chosen a p? Why is it a prime that you've picked?

DAVID: Umm, yeah there's no need to actually, yeah I picked a prime because I was thinking along the lines of prime factorisations, but actually looking over this proof you don't need to pick a prime, so infact you don't need to do any induction, you just pick some, you just pick a number.

INTERVIEWER: Yeah, so what's the relationship between that one, number 4, and the one you proved a minute ago [*Conjecture 2*]? If at all?

DAVID: Umm, if n is perfect then kn is abundant. Errm, oh yeah, so it's the same thing isn't it? [*laughs*]. Gosh! It's the contrapositive, well, hang on, it's not quite the contrapositive is it? Because if n is perfect, it doesn't mean erm so here we're not assuming perfect, we're assuming not deficient, so it's not quite the same, and we're not quite proving the same thing, because we're proving that kn is not deficient.

INTERVIEWER: Right, so it's nearly the contrapositive?

DAVID: So, it's sort of, it's similar to the contrapositive, but it's not quite.

INTERVIEWER: Yeah, because you've come up with two entirely different proofs which is quite interesting

DAVID: Yeah I have, haven't I? [*laughs*]

INTERVIEWER: Can you use this, or that, to prove this and vice versa? Or because of the slight difference are they entirely unrelated?

DAVID: I prefer this one [*laughs, points at Conjecture 4*] because it doesn't have this [*points at the k divides k^2 argument*]. I don't know, can we do it? So what do we have? We had if n is perfect, that's what we have, and then we say the sum of the divisors of kn is greater than or equal to that, yeah I mean you would because you've got this greater than or equal to, this thing would be an equals, so you'd have, oh hang on, no it's abundant, so you'd have a greater than

or equal to, you'd still need to show that it's a greater than, and I think this is what this does.

INTERVIEWER: Right, so that strict inequality makes this difference?

DAVID: Yeah.

INTERVIEWER: OK, so can I give you another one then?

DAVID: OK, this is the sort of thing I hate because, whenever you're thinking factors of things, and you're multiplying, and then someone throws in an addition and you think right ok, got to be a bit careful because you've got two binary operations going on. OK, if m and n are abundant then you've got, OK, so umm, let's write this definition down so the sum can't be, the sum of k is a divisor of n, err, is greater than 2, strictly greater than, k is a divisor of m, greater than $2m$. Yeah, ok, the divisors of $m + n$, I mean it would be nice and easy if the divisors of $m + n$ were just the sums of those, but err, if m and n are abundant then $n + m$ are abundant. Errm, let's look at some of these examples again, oh, ok so we've got $12 + 18$ is 30, $18 + 20$ is 38, so we've got two abundant numbers and the sum of them doesn't have it on the list anyway, [*laughs*] I mean, it says the first few abundant numbers, so I'm trusting you [*laughs*] I'm trusting that that list is correct, that it hasn't missed it out, so $n + m$ can't be abundant.

INTERVIEWER: So that's the end of it?

DAVID: I think so, I mean, 38, is it abundant? I mean, I'm assuming it's not.

INTERVIEWER: No, no, it's not, if it's not on the list, you've got to trust my list! I've done all the calculations. So is that enough to disprove the whole thing?

DAVID: If n and m are abundant then, $n + m$, yeah because n and m are abundant, 12 and 18 are abundant but the sum is not, so yeah [*laughs*].

INTERVIEWER: Marvellous, good stuff, let's move on to this one.

DAVID: Oh, so n and m are... oh, hang on, so you've got sums here as well, so I've got my two assumptions down there, and then umm, I want to say the divisors of nm, so what I do know is the sum k, maybe I should use a different k here, multiplied by k', so k is in the divisors of m, umm, k is in divisors of n, so this is actually greater than $4mn$

INTERVIEWER: So you've just multiplied those two inequalities?

DAVID: Yeah, I've multiplied those two inequalities, I'm wondering how, now I'm doing a very sort of blind thing like that, I'm wondering

whether all these terms are going to be divisors of nm, well, if k' is a divisor of m, and k is a divisor of n, then kk' is a divisor of nm right?

INTERVIEWER: Yep

DAVID: So, so yeah, so this is greater than or equal to, sorry, so, it's probably equal is it actually? Is it equal? I don't know, anyway it's certainly, things of kk' yeah, of course, this is fundamental theorem of arithmetic isn't it? You're not going to get, you're going to get any other, cos you've got 1 in there. I think that's actually equal to, but it's certainly greater than or equal to the sum of the divisors maybe k'', k'' in divisors of mn.

INTERVIEWER: Is that right?

DAVID: It's certainly less than or equal to, because every single kk' of this form is some k'', but I think k'' is you see, if k'' divides nm, so what are we saying? If k'' divides nm then, err, k'' divides n, so it can be written as $n = ck''$ and it divides m, so $n = c'k''$, so errm, so we've got a divisor, err, well anyway, you don't need, I don't know, you need. Well I guess, the other thing is, do you get things more than once? Actually, that might not be right [*scribbles something out*], because you might get terms more than once here, if you multiply these two sums together, errm, umm, you see, umm, [*long pause*] if n and m, I mean we could go back to our examples, but I don't particularly want to because we'd be multiplying high numbers and it isn't long enough [*laughs*].

INTERVIEWER: I haven't brought a calculator and I should have done.

DAVID: Yeah, if, if, I mean, [*long pause*], yeah, so you might, you see my problem here? There might be a 3 there and a 2 there, and there might be a 2 there and a 3 there. So, we've got duplications and that would only be counted once in there, so you can't, so that's not actually true. Ah, but if you've got a duplication, it's only ever going to be twice isn't it? I mean, are you going to get it more than once? I mean you might have a 2 there and a 3 there and a 3 there and a 2 there which would give you two 6s, but you're never going to get two different, well, or are you? Yeah, I suppose you could have something like a 1 and an 8, a 4 and a 2 and a 2 and a 4, so I mean, my idea was half that and half that, but errm,

INTERVIEWER: I suppose it's conceivable that you could only have big duplications?

DAVID: Yeah, I suppose if m and n only have something like 12 as a divisor then you could have all sorts of things yeah. Umm,

INTERVIEWER: what's the way round that then?

DAVID: Umm, I think, let's have a look at what we had before. So we had, these two corollarys 4 and 2 wasn't it, let's see if we can use any of those. If n and m are abundant, so if n's abundant then kn's got to be abundant, I mean if n's perfect then kn is abundant, so mn's got to be abudant, I mean, I've realised it's kind of a trivial consequence of this, I mean you can do the same proof as in here can't you? I mean if you, except you start off with greater than or equal to, greater than or equal to $2n$ and then you prove that, so you prove something weaker than this, well, not logically weaker but, I mean it's got to be, if n is perfect then kn is abundant, well if n's abundant already then...

INTERVIEWER: so if you're making it more abundant?

DAVID: Yeah, yeah, yeah. And so, I mean, you'd need a new proof, you'd need to start off with an inequality there and you'd still have inequalities.

INTERVIEWER: So, why does number 6 follow from that?

DAVID: It doesn't quite follow, as I say, but the proof works exactly the same way because well you just forget that m is, pick one of them, say that that's abundant, what I'm going to prove is a stronger thing, if n is abundant then nm is abundant for any $m \in \mathbb{N}$, right? And I think you can just prove it in the same way as here, in the same was as conjecture 2, because you start off with the sum of the divisors is greater than or equal to $2n$, and then you've got the sum of the divisors of kn, sorry different notation and everything, is greater than or equal or to, is still going to be greater than or equal to, right? And then you have the same process.

INTERVIEWER: Right, so you're basically saying that numbers 4,6 and 2 are pretty much the same.

DAVID: Well, not the same, but they're giving you similar things, you're going to prove them using similar techniques. I actually think 4 and 2 are relatively different because, I think, somehow, 2 required this extra step [*points at the k^2 argument*], this kind of simple step was really conjecture 4, similar to the one in conjecture 4.

INTERVIEWER: OK, so where are we? We're happy that number 6 is? What's the situation with number 6?

DAVID: Yeah, yeah, I think it's true.

INTERVIEWER: Entirely true, good stuff.

DAVID: And I think there's a stronger statement, that one of these doesn't need to be abundant.

INTERVIEWER: OK, let's do another one then.

DAVID: Umm, if n is abundant then n is not of the form p^m for some natural n and prime p. Well ok, this is not true for $m = 2$ right? Because we've proved that before.

INTERVIEWER: Errm, have we?

DAVID: Errr, this was, where is it? p_1, p_2 primes then $p_1 p_2$ is not abundant, so if it's abundant then it can't be $p_1 p_2$, but anyway. We only need to concern ourselves with n bigger than or equal to 3. So I guess what we're proving is p^m is not abundant, [*laughs*], so

INTERVIEWER: So what have you done there? What are we doing?

DAVID: OK, so if n is abundant then n is not of the form, err, so in other words, it's the contrapositive again, if n is of the form p^m then it's, then n is not abundant, so we know this, err, right, so what's the sum of k in p^m? k, err, whatever. Sum of k in divisors of p^m. Umm, so what is this? Well we've got a 1, we've got a p, we've got a p^2, etc. plus p^m. And, err, well I guess, let's do, we know it's true for $m = 2$, so let's do an induction thing. What are we doing? We're proving that this is not abundant, so this thing is less than or equal to $2p^m$, is that right? Yeah, not abundant, yeah, that's right. Umm, so then, so then we want to prove that $2p^{m+1}$ is greater than or equal to all this, so, umm, so what have I got to do? So sum of k in divisors of p^{m+1} is equal to $1 + p + \ldots + p^{m+1}$ which is, umm, so if you multiply both sides, if you multiply this by p, sorry my mind's gone, so we want to show that this is less than or equal to $2p^{m+1}$, so umm, that's done by multiplying that by, well $2p^{m+1}$ is certainly less than $p + p^2 +$ ah, ha, I think I might have a problem. And then I want to put the equality, I mean you've got that haven't you? [*laughs*] you've got inequalities going both ways so you've got a bit of a problem, umm, err, so I wonder whether there's a, a false thing here. Oh, well hang on, no this is curious, so you've got 36 here, and 36 is abundant, and is of the form p^2, but hang on, didn't we prove that p^2, that p_1, p_2 are primes then $p_1 p_2$ is not abundant? What about 6 and 6?

INTERVIEWER: So what's gone wrong here?

DAVID: Sorry, so there's an abundant number 36, right? so 36 is equal to, oh, sorry, I'm being right, no, you've been getting me doing maths for too long now [*laughs*] my brain's going funny, 6 is not prime, of course [*laughs*]. Umm, yeah, watch me do maths for long enough and I'll do something really stupid. OK, so, err, fine, so

INTERVIEWER: So you're convinced this is true are you?

DAVID: Well, I, I think so, I mean I don't know [*laughs*] I'm not sure, I mean it looks, I mean these two things, I mean they're only 1 apart, this is not, I don't know whether, I mean, maybe it isn't, maybe at some point this catches up on you, umm, err, I mean it's difficult, you see, there's certainly no counterexamples there, so I wouldn't want to look for a counterexample now, I mean I'd plug it into a computer or something to see what happens, so I mean it just looks a bit fishy. I mean, how else may I, might I prove this? Umm, ok, so I'm using induction, maybe I should try to do it more directly. What did I do with these two primes then? So, there's a similar sort of thing here? So I said, hmm, what other conjectures have we got? Umm, if n and m is abundant, so we don't want that that's useless, we want two non-abundant numbers, we had some, if n's perfect, so that doesn't help us, if n is deficient then every divisor, that doesn't help us. Umm, I reckon it may be false, but I'd need to, well, I'd need to do a computer program to calculate higher abundant numbers and umm, err, see if it happens, and then the next thing I'd do is try to find a counterexample, because I have a feeling that each time you do your induction step this gets bigger and bigger and bigger and maybe it starts to exceed this, something goes wrong. This adding 1 each time means that the difference between these two things means that less and less and less, and then at some point it overtakes. I think in fact at some point it would [*laughs*], I mean, let's, well, so assume $2p^m - 1 + p + \ldots + p^m$ is some number, some number, errm, err, I'm running out of letters, err, α, start running out of letters go to Greek, so err, and err, $2p^{m+1} - (1 + p + \ldots + p^{m+1})$, what's this equal to? So, err, my mind's gone blank, so we've got a, err, $2p^m - p$, so we've got a $+p^{m+1} - (1 + p + \ldots + p^m)$ but this, right, this now is greater than or equal to p, p is prime so it is greater than or equal to $2p^m - (1 + p + \ldots + p^m)$, which is equal to α, ahh! So the difference between this and this thing is actually bigger, so err, actually this proves it.

INTERVIEWER: So that's the end of it?

DAVID: Yeah, that's the end of it, yeah, ok. So, if α is positive which is our induction assumption, our inductive assumption, then, then, the difference between that and this thing, the difference between $2p^{m+1}$ and all it's prime factors is even bigger, so therefore has positive, so therefore $2p^{m+1}$'s got to be greater than that.

INTERVIEWER: And that's your induction step?

DAVID: And that's your induction step.

INTERVIEWER: Marvellous. So are you entirely convinced by that?

DAVID: Yeah, I'm entirely convinced of this, yeah, but that first way doesn't work [*laughs*]. (D1–D7)

Appendix B

Details of Coding on Experiment 4.

This appendix details the different codes that were used during the quasi-judicial case study analysis of interviews in Experiment 4. As is appropriate with the quasi-judicial method, most of these codes were predetermined based on the prima facie theories. However, some non-essential codes were incorporated as a result of the analytical process. These codes are marked with a (∗). Naturally, the codes listed below are not necessarily mutually exclusive.

Examples –

Example – use of examples.

Counterexamples – use of (or search for) counterexamples.

Heuristic biases –

If-heuristic – evidence of the if-heuristic.

notif-heuristic – evidence of the if-heuristic not playing a role.

matching-heuristic – evidence of the matching-heuristic.

Implication as a Journey (∗) – language which talks of an implication as if it were a journey.

Meta-comment (∗) – miscellaneous comment about meta-mathematics.

Warrant –

Authoritarian – a warrant based on authority.

Deductive –

Formalisable – 'Proper' mathematics that can be formalised.

Non-formalisable – Use of generic examples, for example.

Hard to classify – difficult to classify.

Inductive – a warrant involving evaluating one or more specific cases.

Perceptual – conviction is gained through perceptions of mental images.

Transformational – conviction is gained through 'thought experiments' on mental structures.

Ritual – conviction is gained through the appearance of an argument.

Symbolic – conviction is gained through meaningless symbol manipulation.

Logic –

Contrapositive – mention of the contrapositive.

$\Leftrightarrow$ – interpretation of '$\Rightarrow$' as '$\Leftrightarrow$'.

$\Rightarrow$ – interpretation of '$\Rightarrow$' as '$\Rightarrow$'.

Specific/general – comment on whether '$\Rightarrow$' needs to be true in general.

Negations $(*)$ – statements being negated.

Strategic knowledge $(*)$ – evidence of the application of strategic knowledge.

Appendix C

Constructing Odd Abundants

Theorem. *If an integer n is multiplied by a sequence of k consecutive prime numbers that are greater than n, then the resultant number is abundant for a sufficiently large k.*

Proof. Pick an integer n with prime decomposition $n = q_1 q_2 q_3 \dots q_r$. Pick a prime p where $p \neq q_i$ for any i. Assume n has distinct divisors $d_1, d_2, \dots, d_s$, which sum to $\sigma(n)$. Consider the divisors of pn. As p is not a divisor of n this ensures that, for every d_i, pd_i is not a member of the set of divisors of n, i.e. for every i there is no j such that $pd_i = d_j$. This guarantees that all the numbers in the list $d_1, d_2, \dots, d_s, pd_1, pd_2, \dots, pd_s$ are distinct. Also, they are all divisors of pn. These divisors sum to $\sigma(n) + p \cdot \sigma(n) = (p+1) \cdot \sigma(n)$. Therefore, $\sigma(pn) \geq (p+1) \cdot \sigma(n)$.

Consider the 'abundance', $A(n)$, of a number n, given by $A(n) = \frac{\sigma(n)}{n}$. If $A(n) > 2$ then n is abundant. We know, from the above, that:

$$A(pn) = \frac{\sigma(pn)}{pn} \geq \frac{(p+1) \cdot \sigma(n)}{pn} = \frac{\sigma(n)}{n} \cdot \frac{p+1}{p} = A(n) \cdot \frac{p+1}{p}$$

That is to say, multiplying a number n by an appropriate prime p increases n by a factor of p, but it increases $A(n)$ by a factor of at least $\frac{p+1}{p} > p$.

Consider the infinite product $\prod_i \frac{p_i+1}{p_i}$, where the p_i are consecutive odd primes (i.e. $p_1 = 3, p_2 = 5 \dots$). First we will look at the partial product $\prod_{i=1}^{n} \frac{p_i+1}{p_i}$

$$\prod_{i=1}^{n} \frac{p_i+1}{p_i} = \left(1 + \frac{1}{p_1}\right)\left(1 + \frac{1}{p_2}\right) \dots \left(1 + \frac{1}{p_n}\right)$$

When this is factored out we will, for each i, get a term of the form

$$1 \cdot 1 \cdot 1 \cdot \ldots \cdot \frac{1}{p_i} \ldots 1 \cdot 1.$$

Consequently

$$\prod_{i=1}^{n} \frac{p_i + 1}{p_i} \geq \frac{1}{p_1} + \frac{1}{p_2} + \ldots + \frac{1}{p_n} = \sum_{i=1}^{n} \frac{1}{p_i}$$

Since the harmonic series of primes, $\sum_i \frac{1}{p_i}$, is known to diverge (e.g. Nagell, 1951), we know that the *sequence* $\left(\sum_{i=1}^{n} \frac{1}{p_i}\right)_{n \in \mathbb{N}}$ diverges.

So the sequence $\left(\prod_{i=1}^{n} \frac{p_i+1}{p_i}\right)_{n \in \mathbb{N}}$ also diverges, and hence so does the infinite product $\prod_i \frac{p_i+1}{p_i}$. Therefore, multiplying a number n by consecutive primes that are not in the prime decomposition of n (giving rise to the sequence $\{n_i\}_{i \in \mathbb{N}}$) will ensure that $A(n_i) \to \infty$. Hence, for some sufficiently large i, $A(n_i) > 2$ and n_i will be abundant. □

Corollary. *There exist odd abundant numbers.*

Proof. Let $n > 1$ be odd. Then all prime numbers not in the prime decomposition of n and that are greater than n will be odd. The product of odd numbers is odd, so by the theorem, multiplying by consecutive primes we will eventually get an odd abundant number, and all odd multiples of this number will also be abundant. □

Corollary. *The product of the first k odd prime numbers is an odd abundant number for some sufficiently large k.*

Proof. If you pick $n = 3$ in the theorem, then the sequence of increasing odd primes satisfy the property that they do not coincide with any primes in the prime decomposition of n. Hence $3 \times 5 \times 7 \times \ldots \times p_k$ will be an odd abundant number for a sufficiently large k. □

Corollary. *An abundant number can be constructed which has an arbitarily large smallest factor.*

Proof. Choose a large prime number q. Pick the next largest prime p_1, and form a sequence of consecutive primes $p_1, p_2, \ldots$, starting with p_1. Then, by the theorem, $qp_1p_2 \ldots p_k$ will be abundant for a sufficiently large k, and the smallest factor of $qp_1p_2 \ldots p_k$ is q. □

Bibliography

Aberdein, A.(2005). The uses of argument in mathematics. *Argumentation, 19*, 287-301.

Aberdein, A. (2006). The informal logic of mathematical proof. In R. Hersh (Ed.), *18 unconventional essays on the nature of mathematics* (p. 56-70). New York: Springer.

Ahn, W.-K. & Graham, L. M.(1999). The impact of necessity and sufficiency in the Wason four-card selection task. *Psychological Science, 10*(3), 237-242.

Aigner, M. & Ziegler, G. (2000). *Proofs from the book* (Second ed.). Berlin: Springer-Verlag.

Alcock, L. & Simpson, A. (1999). The rigour prefix. In O. Zaslavsky (Ed.), *Proceedings of the 23rd International Conference on the Psychology of Mathematics Education* (Vol. 2, p. 17-24). Haifa, Israel: IGPME.

Alcock, L. & Simpson, A. (2002). Definitions: dealing with categories mathematically. *For the Learning of Mathematics, 22*(2), 28-34.

Alcolea Banegas, J. (1998). L'argumentació en matemàtiques. In E. C. i Moya (Ed.), *XIIè Congrés valenciaà de filosofia* (p. 135-147). Valencià: Diputació de Valencià.

Almor, A. & Sloman, S. A.(1996). Is deontic reasoning special? *Psychological Review, 103*, 374-380.

Anderson, J. R.(1991). Is human cognition adaptive? *Behavioural and Brain Sciences, 14*, 471-485.

APMEP. (1992). *Publication no 88: EVAPM91/2 evaluation des programmes de mathématiques seconde 1991.* France: Association des Professeurs de Mathématiques de l'Enseignement Public.

Arbib, M. A. (1990). A Piagetian persepective on mathematical construction. *Synthèse, 84*, 43-58.

Balacheff, N.(1987). Processus de preuves et situations de validation. *Educational Studies in Mathematics, 18*(2), 147-176.

Balacheff, N.(1988). Aspects of proof in pupils' practice of school mathematics. In D. Pimm (Ed.), *Mathematics, teachers and children* (p. 216-235).

London: Hodder & Stoughton.

Ball, L. J., Lucas, E. J., Miles, J. N. V., & Gale, A. G. (2003). Inspection times and the selection task: What do eye-movements reveal about relevance effects? *Quarterly Journal of Experimental Psychology, 56A*, 1053-1077.

Ball, L. J., Lucas, E. J., & Phillips, P. (2005). *Eye-movements and reasoning: Evidence for relevance effects and rationalisation processes in deontic selection tasks.* Proceedings of the 27th Annual Conference of the Cognitive Science Society, Stresa, Italy. Online Article [accessed 15/07/2005]: `http://www.psych.unito.it/csc/cogsci05/frame/talk/f697-ball.pdf`.

Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. N. (1997). Deitic codes for the emodiment of cognition. *Behavioural and Brain Sciences, 20*, 723-767.

Bell, A. W. (1976). A study of pupils' proof conceptions in mathematical situations. *Educational Studies in Mathematics, 7*, 23-40.

Bickley, W. G. (1966). Some thoughts on mathematical thinking. *The Mathematical Gazette, 50*(371), 1-8.

Bills, L. & Tall, D. O. (1998). Operable definitions in advanced mathematics: the case of the least upper bound. In A. Olivier & K. Newstead (Eds.), *Proceedings of the 22nd International Conference on the Psychology of Mathematics Education* (Vol. 2, p. 104-111). Stellenbosch, South Africa: IGPME.

Bingolbali, E. & Monaghan, J. (2004). Identity, knowledge and departmental practices: mathematics of engineers and mathematicians. In M. J. Høines & A. B. Fuglestad (Eds.), *Proceedings of the 28th International Conference on the Psychology of Mathematics Education* (Vol. 2, p. 127-134). Bergen, Norway: IGPME.

Boole, G. (1854/1958). *An investigation into the laws of thought.* New York: Dover.

Bringsjord, S., Noel, R., & Bringsjord, E. (1998). In defense of logical minds. In *Proceedings of the twentieth annual conference of the cognitive science society* (p. 173-178). Hillsdale, New Jersey: Lawrence Erlbaum.

Bromley, D. B. (1986). *The case-study method in psychology and related disciplines.* Chichester: John Wiley & Sons.

Burton, L. (2004). *Mathematicians as enquirers: Learning about learning mathematics.* Dordrecht: Kluwer.

Carroll, L. (1988). *Alice's adventures in wonderland.* London: Julia MacRae Books.

Chao, S. & Cheng, P. W. (2000). The emergence of inferential rules: the use of pragmatic reasoning schemas by preschoolers. *Cognitive Development, 15*(1), 39-62.

Charness, N., Reingold, E. M., Pomplun, M., & Stampe, D. M.(2001). The perceptual aspect of skilled performance in chess: Evidence from eye movements. *Memory & Cognition*, *29*, 1146-1152.

Chazan, D.(1993). High school geometry students' justification for their views of empirical evidence and mathematical proof. *Educational Studies in Mathematics*, *24*, 359-387.

Cheng, P. W. & Holyoak, K. J. (1989). On the natural selection of reasoning theories. *Cognition*, *33*, 285-313.

Cheng, P. W., Holyoak, K. J., Nisbett, R. E., & Oliver, L. M.(1986). Pragmatic versus syntactic approaches to training deductive reasoning. *Cognitive Psychology*, *18*, 293-328.

Cleeremans, A. & Jiménez, L. (2002). Implicit learning and consciousness: A graded, dynamical perspective. In R. M. French & A. Cleeremans (Eds.), *Implicit learning and consciousness: An empirical, philosophical and computational consensus in the making* (p. 1-40). Hove: Psychology Press.

Clement, J., Lockhead, J., & Monk, G.(1981). Translation difficulties in learning mathematics. *American Mathematical Monthly*, *88*, 286-290.

Coe, R. & Ruthven, K. (1994). Proof practices and constructs of advanced mathematical students. *British Educational Research Journal*, *20*(1), 41-53.

Cohen, L., Manion, L., & Morrison, K. (2001). *Research methods in education* (5th ed.). London: Routledge/Falmer.

Cohen, L. J.(1981). Can human irrationality be experimentally demonstrated? *Behavioural and Brain Sciences*, *4*, 317-370.

Cosmides, L. (1989). The law of social exchange: has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition*, *31*, 187-276.

Cosmides, L. & Tooby, J.(1992). Cognitive adaptations for social exchange. In J. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture.* Oxford: OUP.

Crowley, L., Thomas, M., & Tall, D. O. (1994). Algebra symbols, and translation of meaning. In J. P. da Ponte & J. F. Matos (Eds.), *Proceedings of the 18th International Conference on the Psychology of Mathematics Education* (Vol. 2, p. 240-247). Lisbon, Portugal: IGPME.

Cuff, E. G. & Payne, G. C. F.(1979). *Perspectives in sociology.* London: George Allen & Unwin.

Cummins, D. D. (1996). Dominance hierarchies and the evolution of human reasoning. *Minds and Machines*, *6*, 463-480.

Davis, C. (1850/1970). The logic and utility of mathematics. In J. K. Bidwell & R. G. Clason (Eds.), *Readings in the history of mathematics education*

(p. 39-62). Washington DC: NCTM.

Davis, P. & Hersh, R. (1983). *The mathematical experience.* Harmondsworth: Penguin.

Dawkins, R. (1976). *The selfish gene.* Oxford: Oxford University Press.

de Groot, A. D. (1978). *Thought and choice in chess* (Second ed.). The Hague: Mouton.

de Villiers, M. (1990). The role and function of proof in mathematics. *Pythagoras, 24*, 17-24.

Deloustal-Jorrand, V. (2002). Implication and mathematical reasoning. In A. D. Cockburn & E. Nardi (Eds.), *Proceedings of the 26th International Conference on the Psychology of Mathematics Education* (Vol. 2, p. 281-288). Norwich, UK: IGPME.

Devlin, K. (2001). *The maths gene: Why everbody has it, but most people don't use it* (paperback ed.). London: Orion Books.

Devlin, K. (2004). *When is a proof?* Online article [accessed 28/04/2004]: `http://www.maa.org/devlin/devlin_06_03.html`.

Dreyfus, T. (1991). On the status of visual reasoning in mathematics and mathematics education. In F. Furinghetti (Ed.), *Proceedings of the 15th International Conference on the Psychology of Mathematics Education* (Vol. 1, p. 33-48). Assisi, Italy: IGPME.

Duffin, J. & Simpson, A. (1993). Natural, conflicting and alien. *Journal of Mathematical Behavior, 12*, 313-328.

Dunham, W. (1994). *The mathematical universe.* New York: John Wiley & Sons.

Durand-Guerrier, V. (1996). *Conditionals, necessity, and contingence in mathematics class.* Rutgers University Symposium, Online article [accessed 24/07/2004]: `http://www.cs.cornell.edu/Info/People/gries/symposium/durand.htm`. Rutgers University.

Durand-Guerrier, V. (2003). Which notion of implication is the right one? From logical considerations to a didactic perspective. *Educational Studies in Mathematics, 53*(1), 5-34.

Edgington, D. (2003). What if? questions about conditionals. *Mind and Language, 18*, 380-401.

Edwards, L. D. (1998). Odds and evens: Mathematical reasoning and informal proof among high school students. *Journal of Mathematical Behavior, 17*(4), 489-504.

Ejersbo, L. R., Inglis, M., & Leron, U. (2006). WS08: Intuitive vs analytical thinking: A view from cognitive psychology. In J. Novotná, H. Moraová, M. Krátká, & N. Stehlíková (Eds.), *Proceedings of the 30th International Conference on the Psychology of Mathematics Education* (Vol. 1, p. 208).

Prague, Czech Republic: IGPME.

Ericsson, K. A. & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87*, 215-251.

Evans, J. St. B. T. (1972). Interpretation and 'matching bias' in a reasoning task. *Quarterly Journal of Experimental Psychology, 24*, 193-199.

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

Evans, J. St. B. T. (1984). Heuristic and analytic processes in reasoning. *British Journal of Psychology, 75*, 451-468.

Evans, J. St. B. T. (1989). *Bias in human reasoning: Causes and consequences.* Hove: Erlbaum.

Evans, J. St. B. T. (1995). Relevance and reasoning. In S. E. Newstead & J. St. B. T. Evans (Eds.), *Perspectives on thinking and reasoning* (p. 147-171). Hove, UK: Lawrence Erlbaum.

Evans, J. St. B. T. (1996). Deciding before you think: Relevance and reasoning in the selection task. *British Journal of Psychology, 87*, 223-240.

Evans, J. St. B. T. (1998a). Inspection times, relevance, and reasoning: A reply to Roberts. *Quarterly Journal of Experimental Psychology, 51A*, 811-814.

Evans, J. St. B. T. (1998b). Matching bias in conditional reasoning: Do we understand it after 25 years? *Thinking and Reasoning, 4*, 45-82.

Evans, J. St. B. T. (2003). In two minds: dual-process accounts of reasoning. *Trends in Cognitive Science, 7*, 454-459.

Evans, J. St. B. T. (2004a). Dual processes, evolution and rationality. *Thinking and Reasoning, 10*, 405-410.

Evans, J. St. B. T. (2004b). History of the dual process theory of reasoning. In K. I. Manktelow & M. Cheung Chung (Eds.), *Psychology of reasoning: Theoretical and historical perspectives* (p. 241-266). Hove: Psychology Press.

Evans, J. St. B. T. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin and Review, 13*, 378-395.

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

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

Evans, J. St. B. T. & Curtis-Holmes, J. (2005). Rapid responding increases belief bias: Evidence for the dual-process theory of reasoning. *Thinking and Reasoning, 11*, 382-389.

Evans, J. St. B. T., Ellis, C. E., & Newstead, S. E. (1996). On the mental representation of conditional sentences. *Quarterly Journal of Experimental*

Psychology, *49A*(4), 1086-1114.

Evans, J. St. B. T., Handley, S. J., & Over, D. E. (2003). Conditionals and conditional probability. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *29*, 321-335.

Evans, J. St. B. T., Legrenzi, P., & Girotto, V. (1999). The influence of linguistic form on reasoning: The case of matching bias. *Quarterly Journal of Experimental Psychology*, *52A*(1), 185-216.

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

Evans, J. St. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human reasoning: The psychology of deduction.* Hove, UK: Lawrence Erlbaum.

Evans, J. St. B. T. & Over, D. E. (1996a). *Rationality and reasoning.* Hove, UK: Psychology Press.

Evans, J. St. B. T. & Over, D. E. (1996b). Rationality in the selection task: Epistemic utility versus uncertainty reduction. *Psychological Review*, *103*, 356-363.

Evans, J. St. B. T. & Over, D. E. (2004). *If.* Oxford: OUP.

Evans, J. St. B. T., Over, D. E., & Manktelow, K. I. (1993). Reasoning, decision making, and rationality. *Cognition*, *49*, 165-187.

Evens, H. & Houssart, J. (2004). Categorizing pupils' written answers to a mathematics test question: 'I know but I can't explain'. *Educational Research*, *46*, 269-282.

Fallis, D. (2003). Intentional gaps in mathematical proofs. *Synthèse*, *134*, 45-69.

Feferman, S. (2000). Mathematical intuition vs. mathematical monsters. *Synthèse*, *125*, 317-332.

Fiddick, L., Cosmides, L., & Tooby, J. (2000). No interpretation without representation: the role of domain-specific representations and inferences in the Wason selection task. *Cognition*, *77*(1), 1-79.

Fischbein, E. (1987). *Intuition in science and mathematics.* Dordrecht: Reidel.

Galton, F. (1880). Visualised numerals. *Nature*, *21*, 252-256.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "heuristics and biases". In W. Stroebe & M. Hewstone (Eds.), *European review of social psychology* (Vol. 4). Chichester: John Wiley & Sons.

Gigerenzer, G. (1996). On narrow norms and vague heuristics. *Psychological Review*, *103*, 592-596.

Gigerenzer, G. & Hug, K. (1992). Domain-specific reasoning: social contracts, cheating, and perspective change. *Cognition*, *43*(2), 127-171.

Ginsburg, H. (1981). The clinical interview in psychological research on mathematical thinking: Aims, rationales, techniques. *For the Learning of Mathematics*, *1*(1), 4-11.

Glaser, B. G. & Strauss, A. L.(1967). *The discovery of grounded theory: Strategies for qualitative research.* Chicago: Aldine.

Goel, V. & Dolan, R. J. (2003). Explaining modulation of reasoning by belief. *Cognition, 87*, B11-B22.

Goetting, M.(1995). *The college students' understanding of mathematical proof.* Unpublished doctoral dissertation, University of Maryland.

Griggs, R. A. & Cox, J. R. (1982). The elusive thematix-materials effect in Wason's selection task. *British Journal of Psychology, 73*, 407-420.

Griggs, R. A. & Randell, S. E.(1986). Scientists and the selection task. *Social Studies of Science, 16*, 319-330.

Hadamard, J. (1945). *The psychology of invention in the mathematical field* (1954 ed.). New York: Dover Publications.

Hahn, H. (1933/1960). The crisis in intuition. In J. R. Newman (Ed.), *The world of mathematics* (Vol. 3, p. 1956-1976). London: Allen and Unwin.

Hanna, G. (1991). Mathematical proof. In D. O. Tall (Ed.), *Advanced mathematical thinking* (p. 54-61). Dordrecht: Kluwer.

Harel, G.(2001). The development of mathematical induction as a proof scheme: A model for dnr-based instruction. In S. Campbell & R. Zazkis (Eds.), *Learning and teaching number theory* (p. 185-212). Westport, CT: Ablex Publishing Corp.

Harel, G. (in press). Students' proof schemes revisited: historical and epistemological considerations. In P. Boero (Ed.), *Theorems in school: From history, epistemology and cognition to classroom practice.* Rotterdam: Sense.

Harel, G. & Sowder, L. (1998). Students' proof schemes: Results from exploratory studies. In A. H. Schoenfeld, J. Kaput, & E. Dubinsky (Eds.), *Research in collegiate mathematics III* (p. 234-282). Providence, RI: American Mathematical Society.

Harel, G. & Sowder, L.(2005). Advanced mathematical-thinking at any age: Its nature and development. *Mathematical Thinking and Learning, 7*, 27-50.

Hartson, W. R. & Wason, P. C. (1983). *The psychology of chess.* London: Chrysalis Books.

Hattori, M.(2002). A quantitative model of optimal data selection in Wason's selection task. *Quarterly Journal of Experimental Psychology, 55A*, 1241-1272.

Hauk, S. (2005). *Preservice elementary teachers' understanding of logical inference.* Online article [accessed 13/08/2005] `http://hopper.unco.edu/faculty/personal/hauk/barzilai/paper/PSTreasoning040421.pdf`. University of Northern Colorado.

Hazzan, O. & Leron, U.(1996). Students' use and misuse of mathematical theo-

rems: The case of Lagrange's theorem. *For the Learning of Mathematics, 16*(1), 23-26.

He, P. & Kowler, E. (1992). The role of saccades in the perception of texture patterns. *Vision Research, 32*, 2151-2163.

Healy, L. & Hoyles, C. (2000). A study of proof conceptions in algebra. *Journal for Research in Mathematics Education, 31*(4), 396-428.

Hempel, C. G. (1945). Studies in the logic of confirmation. *Mind, 54*, 1-26.

Hersh, R. (1993). Proving is convincing and explaining. *Educational Studies in Mathematics, 24*(4), 389-399.

Hersh, R. (1998). *What is mathematics, really?* London: Vintage Books.

Hoyles, C. & Küchemann, D. (2002). Students' understanding of logical implication. *Educational Studies in Mathematics, 51*(3), 193-223.

Ikehara, C. S. & Crosby, M. E. (2005). Assessing cognitive load with physiological sensors. In *Hicss '05: Proceedings of the 38th Annual Hawaii International Conference on System Sciences* (p. 295.1). Washington, DC, USA: IEEE Computer Society.

Inglis, M. (2004). *Relevance theory explains the maze task.* Presented at the 2nd YERME Summer School, Pôdebrady, Czech Republic.

Inglis, M. (2006). Belief bias and the study of mathematics. *Working Papers of the Warwick Suminer Group, 2*, 35-48.

Inglis, M. & Mejia-Ramos, J. P. (2006). Applying informal logic to mathematics. In *Proceedings of the 3rd International Conference on the Teaching of Mathematics at the Undergraduate Level.* Istanbul, Turkey: Turkish Mathematical Association.

Inglis, M., Mejia-Ramos, J. P., & Simpson, A. (in press). Modelling mathematical argumentation: The importance of qualification. *To appear in Educational Studies in Mathematics.*

Inglis, M. & Simpson, A. (2004). Mathematicians and the selection task. In M. Johnsen Høines & A. B. Fuglestad (Eds.), *Proceedings of the 28th International Conference on the Psychology of Mathematics Education* (Vol. 3, p. 89-96). Bergen, Norway: IGPME.

Inglis, M. & Simpson, A. (2005a). Characterising mathematical reasoning: Studies with the Wason selection task. In *Proceedings of the Fourth Congress of the European Society for Research in Mathematics Education.* Sant Feliu de Guíxols, Spain: ERME.

Inglis, M. & Simpson, A. (2005b). Heuristic biases in mathematical reasoning. In H. L. Chick & J. L. Vincent (Eds.), *Proceedings of the 29th International Conference on the Psychology of Mathematics Education* (Vol. 3, p. 177-184). Melbourne, Australia: IGPME.

Inglis, M. & Simpson, A. (2006). The role of mathematical context in evalu-

ating conditional statements. In J. Novotná, H. Moraová, M. Krátká, & N. Stehlíková (Eds.), *Proceedings of the 30th International Conference on the Psychology of Mathematics Education* (Vol. 3, p. 337-344). Prague, Czech Republic: IGPME.

Inhelder, B. & Piaget, J.(1958). *The growth of logical thinking from childhood to adolescence: an essay on the construction of formal operational structures.* New York: Basic Books.

Irwin, D. E. (2004). Fixation location and fixation duration as indices of cognitive processing. In J. M. Henderson & F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world* (p. 105-134). New York: Psychology Press.

Jackson, S. & Griggs, R. A.(1988). Education and the selection task. *Bulletin of the Psychonomic Society, 26*(4), 327-330.

Jackson, S. & Griggs, R. A. (1990). The elusive pragmatic reasoning schemas effect. *Quarterly Journal of Experimental Psychology, 42A*(2), 353-373.

Johnson, D. L.(1998). *Elements of logic via numbers and sets.* London: Springer-Verlag.

Johnson, R. H. (1999). The relation between formal and informal logic. *Argumentation, 13*, 265-274.

Johnson-Laird, P. N.(1999). Deductive reasoning. *Annual Review of Psychology, 50*, 109-135.

Johnson-Laird, P. N.(2001). Mental models and deduction. *Trends in Cognitive Science, 5*, 434-442.

Johnson-Laird, P. N.(2003). *Obituary: Peter Wason.* The Guardian, 25/04/03, [accessed 24/03/2004]: `http://education.guardian.co.uk/obituary/story/0,12212,943315,00.html`.

Johnson-Laird, P. N. & Byrne, R. M. J.(1991). *Deduction.* Hove, UK: Erlbaum.

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

Just, M. A. & Carpenter, P. A.(1980). A theory of reading: from eye fixations to comprehension. *Psychological Review, 87*, 329-354.

Kahneman, D.(2003). Maps of bounded rationality. In T. Frängsmyr (Ed.), *Les prix Nobel, the Nobel prizes 2002.* Stockholm: Nobel Foundation.

Kahneman, D. & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review, 103*, 582-591.

Kemeny, J. G.(1964). *Random essays on mathematics, education and computers.* Englewood Cliffs, NJ: Prentice Hall.

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

Kirby, K. N. (1994). Probabilities and utilities of fictional outcomes in Wason's four-card selection task. *Cognition, 51*, 1-28.

Klaczynski, P. A., Gelfand, H., & Reese, H. W. (1989). Transfer of conditional reasoning: effects of explanations and initial problem types. *Memory & Cognition, 17*(2), 208-220.

Knipping, C. (2003). Argumentation structures in classroom proving situations. In M. A. Mariotti (Ed.), *Proceedings of the Third Congress of the European Society for Research in Mathematics education.* Bellaria, Italy: ERME. Online Article [accessed 12/08/2005] `http://www.dm.unipi.it/~didattica/CERME3/proceedings/Groups/TG4/TG4_Knipping_cerme3.pdf`.

Knoblich, G., Ohlsson, S., & Raney, G. E. (2001). An eye movement study of insight problem solving. *Memory & Cognition, 29*, 1000-1009.

Knuth, E. (2002). Secondary school mathematics teachers conceptions of proof. *Journal for Research in Mathematics Education, 33*(5), 379-405.

Kotov, A. (1971). *Think like a grandmaster.* London: Batsford.

Krantz, J. H. & Dalal, R. (2000). Validity of web-based psychological research. In M. H. Birnbaum (Ed.), *Psychological experiments on the internet* (p. 35-60). San Diego: Academic Press.

Krummheuer, G. (1995). The ethnology of argumentation. In P. Cobb & H. Bauersfeld (Eds.), *The emergence of mathematical meaning: Interaction in classroom cultures* (p. 229-269). Hillsdale: Erlbaum.

Küchemann, D. & Hoyles, C. (2002). Students' understanding of a logical implication and its converse. In A. D. Cockburn & E. Nardi (Eds.), *Proceedings of the 26th International Conference on the Psychology of Mathematics Education* (Vol. 3). Norwich, UK: IGPME.

Küchemann, D. & Hoyles, C. (2004). Year 10 students' proofs of a statement in number/algebra and their responses to related multiple choice items: Longitudinal and cross-secional comparisons. In O. McNamara (Ed.), *Proceedings of the British Society for Research into Learning Mathematics* (Vol. 24, p. 37-42). Kings College London: BSRLM.

Lakoff, G. & Johnson, M. (1980). *Metaphors we live by.* Chicago: Chicago University Press.

Laming, D. (1996). On the analysis of irrational data selection: A critique of Oaksford and Chater. *Psychological Review, 103*, 364-373.

Lave, J. (1988). *Cognition in practice: Mind, mathematics, and culture in everyday life.* Cambridge: CUP.

Lehman, D. R., Lempert, R. O., & Nisbett, R. E. (1988). The effects of graduate training on reasoning. *American Psychologist, 43*, 431-442.

Lehman, D. R. & Nisbett, R. E. (1990). A longitudinal study of the effects

of undergraduate training on reasoning. *Developmental Psychology, 26,* 952-960.

Leron, U. & Hazzan, O.(2006). The rationality debate: Application of cognitive psychology to mathematics education. *Educational Studies in Mathematics, 62,* 105-126.

Lewis, D. (1986). Probabilities of conditionals and conditional probablities. *Philosophical Review, 85,* 297-315.

Liberman, N. & Klar, Y. (1996). Hypothesis testing in Wason's selection task: social exchange cheating detection or task understanding. *Cognition, 58*(1), 127-156.

Liversedge, S. P., Paterson, K. B., & Pickering, M. (1998). Eye movements and measures of reading time. In G. Underwood (Ed.), *Eye guidance in reading and scene perception* (p. 55-76). Oxford: Elsevier Science.

Locke, J.(1706/1971). *Conduct of the understanding.* New York: Burt Franklin.

Lopes, L. L. (1991). The rhetoric of irrationality. *Theory and Psychology, 1,* 65-82.

Lucas, E. J. & Ball, L. J.(2005). Think-aloud protocols and the selection task: Evidence for relevance effects and rationalisation processes. *Thinking and Reasoning, 11,* 35-66.

Manktelow, K. I.(1999). *Reasoning and thinking.* Hove, UK: Psychology Press.

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

Manktelow, K. I. & Over, D. E.(1990). *Inference and understanding: a philosophical perspective.* London: Routledge.

Mariotti, M. A. (2006). Proof and proving in mathematics education. In A. Gutiérrez & P. Boero (Eds.), *Handbook of research on the psychology of mathematics education: Past, present and future* (p. 173-204). Rotterdam: Sense.

Markowitz, L. M. & Tweney, R. D.(1981a). *Confirmatory and disconfirmatory heuristics in mathematical reasoning.* (Unpublished manuscript)

Markowitz, L. M. & Tweney, R. D. (1981b, May). *An investigation of the behaviour of mathematicians engaged in testing a conjecture.* Presented at Midwestern Psychological Association, Detroit, MI.

Marton, F. (1981). Phenomenography – describing conceptions of the world around us. *Instructional Science, 10,* 177-200.

Marton, F. & Saljo, R.(1976). On qualitative differences in learning 1. *British Journal of Educational Psychology, 46,* 4-11.

Mason, J. (2002). *Researching your own practice: The discipline of noticing.* London: Routledge/Falmer.

Mason, J., Burton, L., & Stacey, K.(1982). *Thinking mathematically.* London:

Addison-Wesley.

McCulloch, W. & Pitts, W.(1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, *5*, 115-137.

Menger, K.(1943). What is dimension? *American Mathematical Monthly*, *50*, 2-7.

Mitchell, D.(1962). *An introduction to logic.* London: Hutchinson.

Mitchell, J. C.(1983). Case and situation analysis. *The Sociological Review*, *31*, 187-211.

Moore, E. H.(1900). On certain crinkly curves. *Transactions of the American Mathematical Society*, *1*, 72-90.

Moore, R.(1994). Making the transition to formal proof. *Educational Studies in Mathematics*, *27*, 249-266.

Mynatt, C. R., Doherty, M. E., & Tweney, R. D. (1978). Consequences of confirmation and disconfirmation in a simulated research environment. *Quarterly Journal of Experimental Psychology*, *30*, 395-406.

Nagel, E. & Newman, J. R.(2001). *Gödel's proof* (Revised ed.). London: New York UP.

Nagell, T. (1951). *Introduction to number theory.* New York: John Wiley & Sons.

Newstead, S. E., Charles Ellis, M., Evans, J. St. B. T., & Dennis, I. (1997). Conditional reasoning with realistic material. *Thinking and Reasoning*, *3*(1), 49-76.

Nisbett, R. E. & Wilson, T. D.(1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*, 231-295.

Oakhill, J. V., Johnson-Laird, P. N., & Garnham, A.(1989). Believability and syllogistic reasoning. *Cognition*, *31*, 117-140.

Oakley, C. O.(1946). Mathematics. *American Mathematical Monthly*, *56*, 19.

Oaksford, M. & Chater, N.(1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*, 608-631.

Oaksford, M. & Chater, N.(1995a). Information gain explains relevance which explains the selection task. *Cognition*, *57*, 97-108.

Oaksford, M. & Chater, N.(1995b). Theories of reasoning and the computational explanation of everyday inference. *Thinking and Reasoning*, *1*, 121-152.

Oaksford, M. & Chater, N. (2003). Optimal data selection: Revision, review and reevaluation. *Psychonomic Bulletin and Review*, *10*, 289-318.

O'Brien, T. C.(1973). Logical thinking in college students. *Educational Studies in Mathematics*, *5*(1), 71-79.

O'Brien, T. C., Shapiro, B. J., & Reali, N. C.(1971). Logical thinking – language and context. *Educational Studies in Mathematics*, *4*, 201-219.

Ormerod, T. C., Manktelow, K. I., & Jones, G. V. (1993). Reasoning with

three types of conditional: Biases and mental models. *Quarterly Journal of Experimental Psychology, 46A*, 653-677.

Osman, M. (2004). An evaluation of dual-process theories of reasoning. *Psychonomic Bulletin and Review, 11*, 988-1010.

Over, D. E. (2004). Psychology of conditionals. In K. I. Manktelow & M. Cheung Chung (Eds.), *Psychology of reasoning: Theoretical and historical perspectives* (p. 75-94). Hove: Psychology Press.

Over, D. E. & Evans, J. St. B. T. (2003). The probability of conditionals: The psychological evidence. *Mind and Language, 18*, 340-358.

Pan, B., Hembrooke, H. A., Gay, G. K., Granka, L. A., Feusner, M. K., & Newman, J. K. (2004). The determinants of web page viewing behavior: An eye-tracking study. In A. T. Duchowski & R. Vertegaal (Eds.), *Proceedings of the 2004 eye tracking research and applications symposium* (p. 147-154). New York: ACM Press.

Parsons, L. M. & Osherson, D. (2001). New evidence for distinct right and left brain systems for deductive versus probabilistic reasoning. *Cerebral Cortex, 11*, 945-965.

Pedemonte, B. (2003). What kind of proof can be constructed following an abductive argumentation? In M. A. Mariotti (Ed.), *Proceedings of the Third Congress of the European Society for Research in Mathematics education.* Bellaria, Italy: ERME. Online Article [accessed 12/08/2005] `http://www.dm.unipi.it/~didattica/CERME3/proceedings/Groups/TG4/TG4_Pedemonte_cerme3.pdf`.

Piaget, J. (1929). *The child's conception of the world.* London: Routledge & Kegan Paul.

Piattelli-Palmarini, M. (1994). *Inevitable illusions: How mistakes of reason rule our minds.* New York: John Wiley & Sons.

Pinker, S. (1997). *How the mind works.* London: Penguin Books.

Pinker, S. (2002). *The blank slate : the modern denial of human nature.* London: Allen Lane.

Plantinga, A. (1993). *Warrant and proper function.* New York: Oxford University Press.

Poincaré, H. (1905). *Science and hypothesis.* London: Walter Scott Publishing.

Porteous, K. (1990). What do children really believe? *Educational Studies in Mathematics, 21*, 589-598.

QAA. (2002). *Mathematics, statistics and operational research subject benchmark standards.* Online article [accessed 15/07/2005]: `http://www.qaa.ac.uk/academicinfrastructure/benchmark/honours/mathematics.pdf`.

Quine, W. V. O. (1966). *Methods of logic* (Second (revised & corrected) ed.).

London: Routledge & Kegan Paul.

Ramsey, F. P. (1931/1990). General propositions and causality. In D. H. Mellor (Ed.), *Philosophical papers* (p. 145-163). Cambridge: CUP.

Rav, Y. (1999). Why do we prove theorems? *Philosophia Mathematica, 7*, 5-41.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*, 372-422.

Recio, A. & Godino, J. (2001). Institutional and personal meanings of mathematical proof. *Educational Studies in Mathematics, 48*, 83-99.

Reid, D. & Inglis, M. (2005). Talking about logic. *For the Learning of Mathematics, 25*(2), 24-25.

Reips, U.-D. (2000). The web experiment method: Advantages, disadvantages, and solutions. In M. H. Birnbaum (Ed.), *Psychological experiments on the internet* (p. 89-117). San Diego: Academic Press.

Rips, L. (1989). The psychology of knights and knaves. *Cognition, 31*(2), 85-116.

Rips, L. (1994). *The psychology of proof : deductive reasoning in human thinking.* Cambridge, Mass: MIT Press.

Roberts, M. J. (1998a). How should relevance be defined? What does inspection time measure? A reply to Evans. *Quarterly Journal of Experimental Psychology, 51A*, 815-817.

Roberts, M. J. (1998b). Inspection times and the selection task: are they relevant? *Quarterly Journal of Experimental Psychology, 51A*, 781-810.

Roberts, M. J. (2002). The elusive matching bias effect in the disjunctive selection task. *Experimental Psychology, 49*(2), 89-97.

Roberts, M. J. & Newton, E. J. (2001). Inspection times, the change task, and the rapid-response selection task. *Quarterly Journal of Experimental Psychology, 54A*, 1031-1048.

Robson, C. (1993). *Real world research.* Oxford: Blackwell.

Rodd, M. M. (2000). On mathematical warrants: Proof does not always warrant, and a warrant may be other than a proof. *Mathematical Thinking and Learning, 2*, 221-244.

Ross, M. & Sicoly, F. (1979). Egocentric biases in availability and attribution. *Journal of Personality and Social Psychology, 37*, 322-336.

Rousseau, J.-J. (1762/1997). *The social contract and other later political writings.* Cambridge: CUP.

Russell, B. (1961). *The basic writings of bertrand russell, 1903-1959.* London: Allen and Unwin.

Sadler, D. R. (1981). Intuitive data processing as a potential source of bias in naturalistic evaluations. *Educational Evaluation and Policy Analysis, 3*, 25-31.

Schmalz, R. (1988). The role of intuition in doing mathematics. *Journal of Mathematical Behavior*, *7*, 33-44.

Schoenfeld, A. H. (1992). Learning to think mathematically: problem solving, metacognition and sense making in mathematics. In D. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (p. 334-370). New York: Macmillan.

Schroyens, W., Schaeken, W., Fias, W., & d'Ydewalle, G. (2000). Heuristic and analytic processes in propositional reasoning with negatives. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *26*, 1713-1734.

Sclingensiepen, K. H., Campbell, F., Legge, G. E., & Walker, T. D.(1986). The importance of eye movements in the analysis of simple patterns. *Vision Research*, *26*, 1111-1117.

Selden, A. & Selden, J. (2003). Validations of proofs considered as texts: can undergraduates tell whether an argument proves a theorem? *Journal for Research in Mathematics Education*, *34*(1), 4-36.

Simon, H. A.(1983). *Reason in human affairs.* Stanford, CA: Stanford University Press.

Simosi, M. (2003). Using Toulmin's framework for the analysis of everyday argumentation: Some methodological considerations. *Argumentation*, *17*, 185-202.

Simpson, A. (1995). Focusing on student attitudes to proof. *Teaching and Learning Undergraduate Mathematics Newsletter*, *3*.

Singh, S.(1997). *Fermat's last theorem.* London: Fourth Estate.

Skemp, R. R.(1979). *Intelligence, learning and action.* Chichester: John Wiley & Sons.

Sloman, S. A.(1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*, 3-22.

Smagorinsky, P.(1989). The reliability and validity of protocol analysis. *Written Communication*, *63*, 463-479.

Sperber, D., Cara, F., & Girotto, V. (1995). Relevance theory explains the selection task. *Cognition*, *57*, 31-95.

Sperber, D. & Girotto, V.(2002). Use or misuse of the selection task? Rejoinder to Fiddink, Cosmides and Tooby. *Cognition*, *85*, 277-290.

Sperber, D., Premack, D., & Premack, A. J. (Eds.). (1995). *Causal cognition: a multidisciplinary debate.* Oxford: OUP.

Sperber, D. & Wilson, D. (1986). *Relevance: Communication and cognition.* London: Blackwell.

Stalnaker, R.(1968). A theory of conditionals. *American Philosophical Quarterly Monograph Series*, *2*, 98-112.

Stanovich, K. E. (1999). *Who is rational? studies of individual differences in reasoning.* Mahwah, NJ: Lawrence Erlbaum.

Stanovich, K. E. (2003). The fundamental computational biases of human cognition: Heuristics that (sometimes) impair decision making and problem solving. In J. E. Davidson & R. J. Sternberg (Eds.), *The psychology of problem solving* (p. 291-342). New York: CUP.

Stanovich, K. E. (2004). *The robot's rebellion: Finding meaning in the age of darwin.* Chicago: Chicago University Press.

Stanovich, K. E. & West, R. F. (1998). Cognitive ability and variation in selection task performance. *Thinking and Reasoning, 4*, 193-230.

Stanovich, K. E. & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioural and Brain Sciences, 23*, 645-726.

Stavy, R. & Tirosh, D. (1996). Intuitive rules in science and mathematics: The case of more a - more b. *International Journal of Science Education, 18*, 653-667.

Stenning, K. & van Lambalgen, M. (2001). Semantics as a foundation for psychology: A case study of Wason's selection task. *Journal of Logic, Language and Information, 10*, 273-317.

Stenning, K. & van Lambalgen, M. (2004). The natural history of hypotheses about the selection task: towards a philosophy of science for investigating human reasoning. In K. I. Manktelow & M. Cheung Chung (Eds.), *Psychology of reasoning: Theoretical and historical perspectives.* Hove: Psychology Press.

Sternberg, R. J. (2000). The ability is not general, and neither are the conclusions. *Behavioural and Brain Sciences, 23*, 697-698.

Stewart, I. N. & Tall, D. O. (1977). *The foundations of mathematics.* Oxford: Oxford University Press.

Stylianedes, A. J., Stylianedes, G. J., & Philippou, G. N. (2004). Undergraduate students' understanding of the contraposition equivalence rule in symbolic and verbal contexts. *Educational Studies in Mathematics, 55*, 133-162.

Swanson, D., Schwartz, R., Ginsburg, H., & Kossan, N. (1981). The clinical interview: Validity, reliability and diagnosis. *For the Learning of Mathematics, 2*(2), 31-38.

Tall, D. O. (1979). Cognitive aspects of proof, with special reference to the irrationality of $\sqrt{2}$. In *Proceedings of the 3rd International Conference on the Psychology of Mathematics Education* (p. 203-205). Warwick, UK: IGPME.

Tall, D. O. (1980). The anatomy of a discovery in mathematics research. *For the Learning of Mathematics, 1*(2), 25-34.

Tall, D. O. (1989). The nature of mathematical proof. *Mathematics Teaching, 127*, 28-32.

Tall, D. O. (1995). Cognitive development, representations and proof. In *Proceedings of justifying and proving in school mathematics* (p. 27-38). London: IoE.

Tall, D. O. (1997). From school to university: the effects of learning styles in the transition from elementary to advanced mathematical thinking. In M. Thomas (Ed.), *Proceedings of the Seventh Annual Australasian Bridging Network Mathematics Conference* (p. 9-26). Auckland, New Zealand: University of Auckland.

Tall, D. O. (2004). Building theories: The three worlds of mathematics: A comment on Inglis. *For the Learning of Mathematics, 23*(3), 29-32.

Tall, D. O. & Vinner, S. (1981). Concept image and concept definition in mathematics with particular reference to limits and continuity. *Educational Studies in Mathematics, 12*, 151-169.

Thurston, W. P. (1994). On proof and progress in mathematics. *Bulletin of the American Mathematical Society, 30*, 161-177.

Tirosh, D. & Stavy, R. (1999). Intuitive rules: A way to explain and predict students' reasoning. *Educational Studies in Mathematics, 38*, 51-66.

Toates, F. (1998). The interaction of cognitive and stimulus-response processes in the control of behaviour. *Neuroscience and Biobehavioural Reviews, 22*, 59-83.

Toulmin, S. (1958). *The uses of argument.* Cambridge: CUP.

Toulmin, S. (2001). *Return to reason.* Cambridge, MA: Harvard University Press.

Toulmin, S., Rieke, R., & Janik, A. (1984). *An introduction to reasoning* (Second ed.). New York: Macmillan.

Tsamir, P. (2003). Using the intuitive rule more A—more B for predicting and analysing students' solutions in geometry. *International Journal of Mathematical Education in Science and Technology, 34*, 639-650.

Tversky, A. & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5*, 207-232.

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

Van Dooren, W., de Bock, D., Weyers, D., & Verschaffel, L. (2004). The predictive power of intuitive rules: A critical analysis of 'more A–more B' and 'same A–same B'. *Educational Studies in Mathematics, 56*, 179-207.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology, 12*, 129-140.

Wason, P. C. (1966). Reasoning. In B. Foss (Ed.), *New horizons in psychology* (p. 135-151). Harmondsworth: Penguin Books.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology, 20*, 273-281.

Wason, P. C. (1969). Regression in reasoning? *British Journal of Psychology, 60*(4), 471-480.

Wason, P. C. (1981). The importance of cognitive illusions. *Behavioural and Brain Sciences, 4*, 356.

Wason, P. C. (1983). Realism and rationality in the selection task. In J. St. B. T. Evans (Ed.), *Thinking and reasoning: Psychological approaches* (p. 44-75). London: Routledge & Kegan Paul.

Wason, P. C. & Evans, J. St. B. T. (1975). Dual processes in reasoning? *Cognition, 3*, 141-154.

Wason, P. C. & Green, D. (1984). Reasoning and mental representation. *Quarterly Journal of Experimental Psychology, 36A*, 598–611.

Wason, P. C. & Johnson-Laird, P. N. (1969). Proving a disjunctive rule. *Quarterly Journal of Experimental Psychology, 21*, 14-20.

Wason, P. C. & Johnson-Laird, P. N. (1970). A conflict between selecting and evaluating information in an inferential task. *British Journal of Psychology, 61*, 509-515.

Wason, P. C. & Johnson-Laird, P. N. (1972). *Psychology of reasoning.* London: B.T.Batsford.

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

Weber, K. (2001). Student difficulty in constructing proofs: the need for strategic knowledge. *Educational Studies in Mathematics, 48*, 101-119.

Weber, K. (2003). *Students' difficulties with proof.* MAA Research Sampler 8, available at www.maa.org.

Weber, K. & Alcock, L. (2004). How do mathematicians validate proofs? In D. E. McDougall & J. A. Ross (Eds.), *Proceedings of the 26th Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (Vol. 2, p. 621-628). Toronto, Canada: PMENA.

Weber, K. & Alcock, L. (2005). Using warranted implications to understand and validate proofs. *For the Learning of Mathematics, 25*(1), 34-38.

Whyburn, G. (1942). What is a curve? *American Mathematical Monthly, 49*, 493-497.

Wilson, D. & Sperber, D. (2004). Relevance theory. *UCL Working Papers in Linguistics, 14*, 249-287.

Yackel, E. (2001). Explanation, justification and argumentation in mathematics classrooms. In M. van den Heuvel-Panhuizen (Ed.), *Proceedings of the 25th International Conference on the Psychology of Mathematics Education* (Vol. 1, p. 9-23). Utrecht, Holland: IGPME.

Zazkis, R. (1999). Intuitive rules in number theory: Examples of 'the more of A, the more of B' rule implementation. *Educational Studies in Mathematics, 40*, 197-209.

Author Index

This thesis has been typeset using LaTeX 2_ε. I would like to thank the authors of TAMSAnalyzer, TeXShop, BibDesk, APAcite.sty, flow.c, dialogue.sty and the MacTeX distribution for making their work freely available to the academic community.

Matthew Inglis
Institute of Education
University of Warwick
m.j.inglis@warwick.ac.uk

Compiled on 28th November 2006.