
This item was submitted to [Loughborough's Research Repository](#) by the author.
Items in Figshare are protected by copyright, with all rights reserved, unless otherwise indicated.

Advanced mathematics and deductive reasoning skills: testing the Theory of Formal Discipline

PLEASE CITE THE PUBLISHED VERSION

PUBLISHER

© Nina Attridge

LICENCE

CC BY-NC-ND 4.0

REPOSITORY RECORD

Attridge, Nina. 2019. "Advanced Mathematics and Deductive Reasoning Skills: Testing the Theory of Formal Discipline". figshare. <https://hdl.handle.net/2134/12166>.

This item was submitted to Loughborough University as a PhD thesis by the author and is made available in the Institutional Repository (<https://dspace.lboro.ac.uk/>) under the following Creative Commons Licence conditions.



For the full text of this licence, please go to:
<http://creativecommons.org/licenses/by-nc-nd/2.5/>

Advanced Mathematics and Deductive Reasoning Skills: Testing the Theory of Formal Discipline

Nina Attridge

A thesis submitted for the degree of
Doctor of Philosophy



Loughborough University

January 2013

©Nina Attridge

Contents

1	Introduction	1
2	Literature Review	4
2.1	Introduction	4
2.2	Defining reasoning	4
2.3	Rationality – what is ‘better’ reasoning?	8
2.4	A brief history of the Theory of Formal Discipline	15
2.5	Research into thinking skills and their relation to mathematics .	18
2.6	The psychology of reasoning	28
2.7	Current status of the Theory of Formal Discipline	47
3	Methodology	49
3.1	Introduction	49
3.2	Overview of longitudinal study	49
3.3	Research Ethics	49
3.3.1	Science and society	50
3.3.2	Welfare of participants	51
3.4	The Experimental Method	52
3.5	The Quasi-Experimental Method	55
3.6	Reliability and Validity	59
3.6.1	Reliability	60
3.6.2	Validity	61
4	Measures of reasoning	64
4.1	Introduction	64
4.2	Judgment and Decision Making	64
4.2.1	Heuristics and Biases tasks	65
4.3	Deductive Reasoning	69
4.3.1	Disjunction tasks	70
4.3.2	Conditional tasks	74
4.3.3	Syllogisms tasks	80

4.4	Summary	82
5	The development of reasoning skills in AS level mathematics students	84
5.1	Introduction	84
5.1.1	Testing the Theory of Formal Discipline	84
5.1.2	Summary	87
5.2	Pilot Studies	88
5.2.1	Pilot 1: Splitting the syllogisms task	88
5.2.2	Pilot 2: Disguising the Cognitive Reflection Test	91
5.2.3	Pilot 3: Duration and difficulty of measures	94
5.2.4	Summary	97
5.3	Main study	98
5.3.1	Method	98
5.3.2	Results	102
5.4	Discussion	117
5.4.1	Development of reasoning skills	117
5.4.2	Competency and biases in the Conditional Inference Task	119
5.4.3	Limitations	119
5.4.4	Revised status of the Theory of Formal Discipline	120
5.4.5	Summary of novel findings	121
6	The development of reasoning skills in undergraduate mathematics students	122
6.1	Introduction	122
6.1.1	Testing the Theory of Formal Discipline	122
6.1.2	Summary	124
6.2	Method	125
6.2.1	Design	125
6.2.2	Participants	125
6.2.3	Mathematics Syllabus	125
6.2.4	Measures	125
6.2.5	Procedure	127
6.3	Results	127
6.3.1	Preliminary analyses	128
6.3.2	Development of reasoning skills	130
6.4	Discussion	139
6.4.1	Summary of novel findings	141

7	Linguistic factors in mathematics students' conditional reasoning: 'if then' versus 'only if'	142
7.1	Introduction	142
7.2	Method	145
7.2.1	Participants	145
7.2.2	Design	145
7.2.3	Measures	145
7.2.4	Procedure	145
7.3	Results	145
7.3.1	Endorsement rates	146
7.3.2	Interpretations of the conditional	148
7.4	Discussion	148
7.4.1	Summary of novel findings	151
8	The role of the heuristic level in mathematics students' conditional reasoning	153
8.1	Introduction	153
8.2	Method	155
8.2.1	Design	155
8.2.2	Participants	156
8.2.3	Procedure	156
8.3	Results	156
8.3.1	Manipulation check and covariate assessment	156
8.3.2	Main analyses	158
8.4	Discussion	164
8.4.1	Summary of novel findings	167
9	The role of executive functions in conditional reasoning ability	169
9.1	Introduction	169
9.2	Method	173
9.2.1	Participants	173
9.2.2	Measures	173
9.2.3	Procedure	175
9.3	Results	175
9.3.1	Data cleaning	175
9.3.2	Task performance	175
9.3.3	Relationships between executive functions	177
9.3.4	Executive functions and Conditional Reasoning	178
9.4	Discussion	180
9.4.1	Summary of novel findings	182

10 Conclusions	183
10.1 Introduction	183
10.2 Overview of findings and interpretations	183
10.3 Future research	187
10.4 The Theory of Formal Discipline revisited	188
Publications List	189
A Ravens Advanced Progressive Matrices	203
B Conditional Inference Task	204
C Belief Bias Syllogisms Task	211
D Need for Cognition Scale	219
E AS level Mathematics Test	221
F Thematic Conditional Inference Task	224
G Undergraduate Mathematics Test	232
List of Abbreviations	235
Author Index	237

List of Figures

2.1	The basic structure of the four inferences.	5
2.2	The three positions on human rationality.	11
2.3	The Wason Selection Task.	17
2.4	An example of a reasoning task used by Owen et al (2010). . . .	19
2.5	An example of each type of item from the Belief Bias Syllogisms task.	22
2.6	The disjunctive Knights and Knaves problem used by Toplak and Stanovich (2002), adapted from Shafir (1994).	23
2.7	The causal schema conditional problem used in Lehman and Nisbett (1990).	25
2.8	A valid and unbelievable syllogism.	34
2.9	Evans's (2006) default-interventionist model.	36
2.10	Frederick's (2005) three item Cognitive Reflection Test.	37
2.11	An example of a) a congruent and b) an incongruent trial of the denominator neglect task used by Gillard (2009).	39
2.12	An illustration of g as the shared variance between cognitive abilities, such as school subjects.	42
2.13	An example item from Raven's Advanced Progressive Matrices. .	43
2.14	The Need for Cognition Scale.	46
3.1	Possible quasi-experimental design outcome 1.	57
3.2	Possible quasi-experimental design outcome 2.	58
3.3	Possible quasi-experimental design outcome 3.	59
3.4	Possible quasi-experimental design outcome 4.	60
4.1	Wason's abstract THOG problem.	70
4.2	The Pub problem, a contextualised version of the THOG task. .	71
4.3	Euler diagrams to represent 'if p then q ' under the material conditional and biconditional interpretations.	75
4.4	A Wason Selection Task example	76
4.5	Example Modus Ponens item from the Conditional Inference task.	78

4.6	Example items from the Conditional Inference task showing a) a Modus Tollens inference and b) a Denial of the Antecedent inference.	79
4.7	Example items from the Belief Bias Syllogisms task.	81
5.1	Example items from the Belief Bias Syllogisms task.	88
5.2	Believability ratings for each problem type by test half.	90
5.3	The three items from the Woodcock Johnson III Applied Problems subtest that were used in Pilot Study 2.	92
5.4	Number of intuitive responses in the mixed and non-mixed conditions.	93
5.5	Example item from the Conditional Inference Task.	101
5.6	Interaction between Group and Time on mathematics test scores.	105
5.7	Mean endorsement rates for each of the four inferences in each group at Time 1 and Time 2.	106
5.8	Mean interpretation index scores for each group at Time 1 and Time 2.	108
5.9	Proposed relationship between overall score and a bias index on a reasoning task.	110
5.10	Relationship between the Material Conditional Index and Negative Conclusion Index on the Conditional Inference Task when collating data from five studies.	111
5.11	Relationship between the Defective Conditional Index and Negative Conclusion Index on the Conditional Inference Task when collating data from five studies.	111
5.12	Mean syllogisms scores for each group at Time 1 and Time 2.	113
5.13	Mean belief bias index scores for each group and Time 1 and Time 2.	114
6.1	Example item from the Abstract Conditional Inference task.	126
6.2	Example item from the Thematic Conditional Inference task.	126
6.3	Interaction between Group and Time on CRT scores	130
6.4	Interaction between Group and Time on mathematics test scores	131
6.5	Endorsement rates for each of the four inferences in each group at Time 1 and Time 2 on the Abstract Conditional Inference Task	132
6.6	Interaction between Group and Time on Abstract Conditional Inference indices.	134
6.7	Endorsement rates for each of the four inferences in each group at Time 1 and Time 2 on the Thematic Conditional Inference Task.	136

6.8	Interaction between Group and Time on Thematic Conditional Inference interpretation indices.	138
7.1	Endorsement rates for each inference type in the IT and OI conditions.	147
7.2	Mean interpretation indices in the IT and OI conditions.	149
8.1	Endorsement rates for each inference type in the fast and slow conditions for non-mathematics participants.	159
8.2	Endorsement rates for valid and invalid inferences in the fast and slow conditions for the non-mathematics group.	160
8.3	Mean interpretation indices in each group in the slow condition. .	161
8.4	Consistency with each of the four interpretations for the mathematics and control group under the fast and slow conditions. . .	163
9.1	Example sequence of trials from the 2-back task	174
9.2	Distribution of scores on the shifting task.	176
9.3	Distribution of scores on the inhibition (Stroop) task.	176
9.4	Distribution of scores on the working memory (2-back) task. . . .	177
9.5	Correlation between MCI and working memory scores.	179

List of Tables

2.1	The four inferences and conditional statement types with and without negated premises and conclusions.	6
2.2	Truth table for the four interpretations of ‘if p then q ’.	7
2.3	Validity of the four inferences under each interpretation of the conditional.	8
2.4	Truth table for ‘if it rains then I take an umbrella’ assuming the material conditional.	29
3.1	Demonstration of complete and Latin Square counterbalancing for a set of three tasks.	53
4.1	Truth Table for the exclusive disjunction.	72
4.2	Truth Table for the inclusive disjunction.	72
4.3	Truth Table for the disjunction rule ‘I’ll order wine or water’. . .	73
4.4	Truth Table for ‘if p then q ’ where t = true and f = false. . . .	74
4.5	Truth Table for ‘if p then q ’.	77
4.6	Truth Table for the conditional rule ‘if it rains then I take an umbrella’.	78
4.7	The four inferences and conditional statement types with and without negated premises.	79
5.1	Duration information for each measure used in the test book. . .	96
5.2	Score information for measures in test book.	96
5.3	The four inferences and conditional statement types with and without negated premises and conclusions.	100
5.4	Mean number items endorsed by Inference type, Group and Time point	107
5.5	A hierarchical regression analysis predicting Time 2 Defective Conditional Index scores.	116
6.1	Number of participants in each group who completed each task at Time 2.	128

6.2	Mean scores on each task at Time for those who did and did not take part at Time 2	129
6.3	Number items endorsed on the Abstract Conditional Inference Task by Inference type, Group and Time point.	133
6.4	Index scores for each interpretation of the abstract conditional statement at Time 1 and Time 2 in each group.	133
6.5	Endorsement rates on the Thematic Conditional Inference Task by Inference type, Group and Time point	137
6.6	Index scores for each interpretation of the thematic conditional statement at Time 1 and Time 2 in each group.	137
6.7	Mean belief bias index scores for each group at Time 1 and Time 2.	138
7.1	Truth table for the conditional statements ' p only if q ' and 'if p then q '.	143
7.2	Percentage endorsement rates for each IT inference and the equivalent converted OI inference in Evans's (1977) study.	144
7.3	Mean number of items endorsed for each of the four inferences by condition.	146
7.4	Mean endorsement rates for each IT inference and the equivalent converted OI inference.	147
7.5	Percentage of items endorsed for each inference type across participants in four groups: the current maths IT group, the current maths OI group, Evans's (1977) IT non-maths group and Evans's OI non-maths group.	151
8.1	Mean index scores for each interpretation of the conditional statement in the fast and slow condition for the mathematics and non-mathematics group.	162
9.1	Correlations between Executive Functions.	177
9.2	Correlations between Executive Functions and interpretation indices.	178

Acknowledgements

I am grateful to many people for their help and support during my PhD. First and foremost I would like to thank Matthew Inglis for the huge amount of time and effort he has invested in my work and my development as a researcher. I am incredibly grateful to have had him as a supervisor.

I would also like to thank my second supervisor, Tony Croft, and the members of the midlands Mathematical Cognition Group for their valuable comments and suggestions throughout the course of my PhD. I am also grateful to the staff and students of the Mathematics Education Centre at Loughborough University for providing a friendly and encouraging environment in which to work.

This work would not have been possible without the help of my participants and their schools, who volunteered their time and effort for very little personal reward.

Finally, my thanks go out to the family and friends who have gotten me to this point in my life, in particular, my mum, Fay, my brother, Jack, and my grandparents, Pat and John.

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.

Abstract.

This thesis investigates the Theory of Formal Discipline (TFD): the idea that studying mathematics develops general reasoning skills. This belief has been held since the time of Plato (2003/375B.C), and has been cited in recent policy reports (Smith, 2004; Walport, 2010) as an argument for why mathematics should hold a privileged place in the UK's National Curriculum. However, there is no rigorous research evidence that justifies the claim. The research presented in this thesis aims to address this shortcoming.

Two questions are addressed in the investigation of the TFD: is studying advanced mathematics associated with development in reasoning skills, and if so, what might be the mechanism of this development? The primary type of reasoning measured is conditional inference validation (i.e. 'if p then q ; not p ; therefore not q '). In two longitudinal studies it is shown that the conditional reasoning behaviour of mathematics students at AS level and undergraduate level does change over time, but that it does not become straightforwardly more normative. Instead, mathematics students reason more in line with the 'defective' interpretation of the conditional, under which they assume p and reason about q . This leads to the assumption that not- p cases are irrelevant, which results in the rejection of two commonly-endorsed invalid inferences, but also in the rejection of the valid modus tollens inference. Mathematics students did not change in their reasoning behaviour on a thematic syllogisms task or a thematic version of the conditional inference task.

Next, it is shown that mathematics students reason significantly less in line with a defective interpretation of the conditional when it is phrased ' p only if q ' compared to when it is phrased 'if p then q ', despite the two forms being logically equivalent. This suggests that their performance is determined by linguistic features rather than the underlying logic. The final two studies investigated the heuristic and algorithmic levels of Stanovich's (2009a) tri-process model of cognition as potential mechanisms of the change in conditional reasoning skills. It is shown that mathematicians' defective interpretation of the conditional stems in part from heuristic level processing and in part from effortful processing, and that the executive function skills of inhibition and shifting at the algorithmic level are correlated with its adoption.

It is suggested that studying mathematics regularly exposes students to implicit 'if then' statements where they are expected to assume p and reason about q , and that this encourages them to adopt a defective interpretation of conditionals. It is concluded that the TFD is not supported by the evidence; while mathematics does seem to develop abstract conditional reasoning skills, the result is not more normative reasoning.

Chapter 1

Introduction

Since the time of Plato (375B.C./2003) it has been assumed that people can be taught to think more logically, and in particular, that mathematics is a useful tool for doing so. This is known as the Theory of Formal Discipline (TFD) and is exemplified by the philosopher John Locke’s suggestion 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).

Versions of the TFD are regularly cited in educational policy debates and curricula reform documents (e.g. National Research Council, 2001; Walport, 2010). For example, in a report to the UK government, Smith (2004) argued that mathematics education “disciplines the mind, develops logical and critical reasoning, and develops analytical and problem-solving skills to a high degree” (p.11). Smith went on to argue that mathematics students at university should receive tuition fees rebates, and that school mathematics teachers should benefit from higher salaries. The TFD is not readily accepted by all, however (e.g. Bramall & White, 2000). As of yet there is very little evidence to suggest that the TFD is accurate, and even less to inform us on what the nature and extent of any changes in logical thinking might be. Indeed, Kilpatrick (1983) argued that understanding whether (and which) reasoning skills are developed by studying advanced mathematics was one of the three most important open research questions in mathematics education, and it has been little addressed since that time.

The aim of this thesis is to investigate the veracity of the TFD: does studying advanced mathematics develop reasoning skills, and if so, how? Chapter 2 reviews the psychology of reasoning literature and previous research related to the TFD. Chapter 3 discusses research ethics, the experimental and quasi-experimental methods, and issues of reliability and validity. Chapter 4 intro-

duces the most commonly used tasks to measure reasoning skills, and presents an argument for using Evans, Clibbens and Rood's (1995) Conditional Inference Task as the primary measure of reasoning skills for this thesis.

Chapter 5 presents a longitudinal quasi-experimental study investigating the development of conditional and syllogistic reasoning skills in AS level mathematics students compared to AS level English literature students. It is shown that while the groups do not differ in reasoning behaviour at the start of post-compulsory education, the mathematics students hold a significantly changed interpretation of the conditional after a year of study, while the English literature students do not. However, the students did not become straightforwardly more normative (i.e. correct according to formal logic) in their reasoning. Instead they increasingly adopted the so-called 'defective' interpretation of the conditional, which despite its name can be seen as an improvement on the biconditional interpretation that was more widely adopted at the start. Under a defective interpretation of the conditional statement 'if p then q ', not- p cases are considered irrelevant to the conditional. This leads to the failure to endorse the Modus Tollens inference, which is considered valid under the normative model of conditional logic (the different models of the conditional are discussed in Chapter 2). The same pattern of change was found in undergraduate mathematics students in Chapter 6, but the study suffered from low power rendering the effect non-significant.

Chapter 7 presents an experimental study in which mathematics undergraduate students are given the Conditional Inference Task with the conditional statements phrased as either 'if p then q ' or the logically equivalent ' p only if q '. The 'only if' group performed significantly less in line with the defective conditional than the 'if then' group, suggesting that the change in interpretation found in the AS level study is specific to the linguistic phrasing commonly encountered in mathematics – 'if p then q '. This is contrary to the TFD assumption that studying mathematics brings about a broad change in reasoning behaviour.

Chapter 8 demonstrated, with a time-limited version of the Conditional Inference Task, that the difference in mathematicians' and non-mathematicians' adoption of the defective conditional is partly, but not entirely, due to automatic cognitive processing.

Finally, Chapter 9 examined the relationship between conditional reasoning behaviour and executive functions – the skills that allow us to control our attention and cognitive effort – and suggested that inhibition and shifting skills may play a role in the adoption of a defective conditional interpretation over a biconditional interpretation.

The thesis concludes (Chapter 10) with a proposal that studying mathematics exposes students to implicit statements of the form 'if p then q ', where they

are expected to assume that p is true and deduce something about q (Houston, 2009). This appears to foster a defective interpretation of the conditional that is limited to abstract conditional reasoning of the form ‘if p then q ’. The relationship between studying mathematics and changes to reasoning skills may, therefore, be far more limited than previously thought.

Chapter 2

Literature Review

2.1 Introduction

This research aims to test the claim known as the Theory of Formal Discipline (TFD): that studying mathematics improves general reasoning skills to a greater extent than other disciplines. The research into this issue falls into two strands:

1. Does studying mathematics at advanced levels improve general reasoning skills to a greater extent than other subjects?
2. What cognitive mechanisms are associated with ‘better’ reasoning skills (see Section 2.3 for a discussion of what constitutes ‘better’ reasoning) and are these responsible for any improvement found in strand 1?

The literature review presented below will define reasoning, discuss some perspectives on rationality, review the history of the TFD and evidence relating to its claims, discuss some theories of reasoning and how it might be improved, and finally, it will summarise the current status of the TFD and why it is important that it be investigated.

2.2 Defining reasoning

Before discussing the literature on how we reason and its relation to mathematics, it is important to clarify what reasoning is. Reasoning can be seen as the cognitive process of inferring new information from given information, and there are broadly two forms: deductive and inductive. In deductive reasoning, a conclusion must be true when the premises are true, for example, ‘if p then q ; p ; therefore q ’. With inductive reasoning, a conclusion is probably, but not necessarily, true when the premises are true, for example, ‘all ps seen so far

If p then q	If p then q
p	not p
Conclusion: q	Conclusion: not q
a) Modus Ponens (MP)	b) Denial of the antecedent (DA)
If p then q	If p then q
q	not q
Conclusion: p	Conclusion: not p
c) Affirmation of the consequent (AC)	d) Modus Tollens (MT)

Figure 2.1: The basic structure of the four inferences: Modus Ponens, Denial of the Antecedent, Affirmation of the Consequent, and Modus Tollens.

are qs ; therefore; all ps are qs '. Deductive reasoning is therefore more rigorous than inductive reasoning, and it is also the form used in mathematical proof. In fact, Polya (1954, p. V) stated that "[a] mathematical proof is [deductive] reasoning". Deductive reasoning can be sub-divided into several forms, such as conditional, disjunctive and syllogistic reasoning, and these are discussed in more detail in Chapter 4. However, conditional reasoning will be central to the thesis so it is worth elaborating on here.

Conditional reasoning is the process of drawing conclusions from a conditional statement such as 'if p then q '. It is considered central to logic (Anderson & Belnap, 1975; Braine, 1978; Inglis & Simpson, 2008) and to mathematics (Houston, 2009). There are four inferences commonly drawn from a conditional statement: modus ponens (MP), modus tollens (MT), denial of the antecedent (DA) and affirmation of the consequent (AC). The structure of each of these inferences is shown in Figure 2.1.

There are also four forms of conditional statement created by the presence and absence of negations: 'if p then q ', 'if p then not q ', 'if not p then q ', and 'if not p then not q '. Table 2.1 shows the premises and conclusions of each combination of conditional statement and inference type. To elaborate on an example from the table, an AC inference from an 'if not p then q ' statement would read:

Rule: If not p then q
 Premise: q
 Conclusion: Therefore, not p

Conditional	MP		DA		AC		MT	
	Pr	Con	Pr	Con	Pr	Con	Pr	Con
if p then q	p	q	$\neg p$	$\neg q$	q	p	$\neg q$	$\neg p$
if p then $\neg q$	p	$\neg q$	$\neg p$	q	$\neg q$	p	q	$\neg p$
if $\neg p$ then q	$\neg p$	q	p	$\neg q$	q	$\neg p$	$\neg q$	p
if $\neg p$ then $\neg q$	$\neg p$	$\neg q$	p	q	$\neg q$	$\neg p$	q	p

Table 2.1: The four inferences and conditional statement types with and without negated premises (Pr) and conclusions (Con). The symbol \neg should read ‘not’.

The validity of each of the four inferences depends on the way in which one interprets the conditional statement. There are four common interpretations of a conditional: the material conditional, biconditional, defective conditional and conjunctive conditional readings (Evans, Handley, Neilens & Over, 2007). Each of these is demonstrated in truth table form in Table 2.2.

Under a material conditional reading, which is considered correct by logicians, the statement ‘if p , then q ’ means that p is sufficient for q , and q is necessary for p . In other words, the conditional means ‘ q or not- p ’. The conditional is true in all cases except when p is true and q is false. It is possible for p to be false and q to be true, because p is not necessarily the only determinant of q .

Under a biconditional reading, the conditional is treated as ‘ p if and only if q ’. In other words, p implies q and q implies p , so that both are necessary and sufficient for each other. As such, the statement ‘if p , then q ’ is not true when p is false and q is true as well as when p is true and q is false – p and q must both be true or both be false.

Under a defective reading of the conditional, p is considered necessary and sufficient for q , so only p q cases are considered to make the rule true. However, rather than not- p cases making the rule false, they are considered irrelevant. The rule is considered true when both p and q are true, false when p is true and q is false, and irrelevant when p is not true. This leads to a similar pattern of inferences being deduced as in the case of the material conditional, except for MT. An example of an MT inference is as follows:

If p then q
not- q
Therefore not- p

MT inferences are considered valid under a material interpretation but invalid

p	q	Material	Biconditional	Defective	Conjunctive
t	t	t	t	t	t
t	f	f	f	f	f
f	t	t	f	i	f
f	f	t	t	i	f

Table 2.2: Truth table for the four interpretations of ‘if p then q ’ where t = true, f = false, and i = irrelevant.

under a defective interpretation. This is because the minor premise does not affirm p and is therefore considered irrelevant to the conditional meaning that no necessary conclusions can be drawn. However, if the reasoner is able to construct a contradiction proof then they may be able to deduce not- p from the premise not- q and accept the MT inference. For example “assume for contradiction p , deduce q , but I know not- q , so the assumption p must be false”. This process requires a high level of working memory and so may only be available to the most able reasoners. For this reason, MT is generally considered invalid under a defective interpretation.

Finally, under a conjunctive reading, the conditional is interpreted to mean ‘ p and q ’, so that the conditional is only true when both p and q are true, and false in all other cases.

Consider the example ‘if America is in Europe, then China is in Asia’. We know that the antecedent is false and the consequent is true, and we can see from Table 2.2 that a false antecedent true consequent case is considered irrelevant to the rule under a defective interpretation, but it is considered to make the rule true under the material conditional and false under the biconditional and conjunctive interpretations.

Table 2.3 shows whether each of the four inferences (MP, MT, DA, AC) is considered valid under each of the four interpretations of the conditional. Evans et al. (2007) found that MP was almost universally accepted (97.5%), followed by AC (74%), MT (50%) and finally DA (38.5%). It is the material conditional that is considered correct by logicians, but Evans et al.’s participants clearly did not hold this interpretation. Conditionals in everyday language tend to assume a biconditional or defective interpretation (Cummins, 1995; Markovits, 1985; Venet & Markovits, 2001; Zepp, 1987; Verschueren, Schaeken & d’Ydewalle, 2005) and so on the whole, people are not well versed with the material conditional, as reflected in the endorsement rates of Evans et al.’s (2007) participants. An interesting possibility in the context of this thesis is that the study of mathematics increases one’s familiarity and competence with the

Inference	Material	Biconditional	Defective	Conjunctive
MP	Valid	Valid	Valid	Valid
DA	Invalid	Valid	Invalid	Invalid
AC	Invalid	Valid	Invalid	Valid
MT	Valid	Valid	Invalid	Invalid

Table 2.3: Validity of the four inferences under each interpretation of the conditional.

material conditional, leading to more normative performance on the Conditional Inference Task.

Conditional inference ability will be the focus of this thesis for reasons discussed in Chapter 4, but the literature review will consider relevant research into any form of reasoning because, as is demonstrated below, the TFD claims are vague as to exactly which skills are improved by studying mathematics. Each further type of reasoning discussed will be defined as it is introduced.

2.3 Rationality – what is ‘better’ reasoning?

This thesis is concerned with whether reasoning can be improved, so an obvious issue is what should be considered ‘better’ and ‘worse’ reasoning. It is widely evidenced that people’s reasoning behaviour is biased (i.e. it deviates from normative models of reasoning, Kahneman & Tversky, 1972; Stanovich, 2009b), but the issue is whether this demonstrates irrationality or whether it can be explained and excused in some other way. Some argue that the biases we are prone to are not evidence of bad reasoning at all, while of course others disagree and argue that we can do better. Before elaborating on these perspectives, some of the common reasoning biases are discussed in order to clarify the problem that these different views conflict over.

Heuristics and Biases

There is a huge literature on heuristics and biases in reasoning behaviour, and this section will review just a few of the common issues. The term heuristic refers to a cognitive ‘shortcut’, which is used in place of a longer process of thinking, and which may be conscious or unconscious (Gigerenzer & Gaissmaier, 2011). Heuristics save a great deal of mental effort and usually they are very effective, but in some cases they may lead to biases. A bias is a systematic deviation from normative models of rationality: an error that tends to be shown repeatedly

within and between reasoners (Kahneman & Tversky, 1972). Biases are deemed to be errors based on the assumption that normative models maximise the outcomes of the individual, and so if one does not follow the normative model of a given reasoning situation, it is to their own detriment (Stanovich, 1999). Chapter 4 provides an introduction to the most commonly used tasks to measure reasoning behaviour, and in doing so it also provides an overview of many of the biases commonly observed. Here, only a few biases are introduced in order to demonstrate that humans do not always behave in line with normative models.

Beginning with the Conditional Inference Task, two biases that are commonly observed are the negative conclusion bias (NCB) and affirmative premise bias (APB). The NCB refers to the tendency for participants to accept more inferences that result in negative conclusions than affirmative conclusions. For example, the DA inference ‘if p then q , not p , therefore not q ’ would be accepted more often than the DA inference ‘if p then not q , not p , therefore q ’ despite being logically equivalent (Evans et al., 1995; Evans & Handley, 1999). This effect is observed on both denial inferences (MT and DA) but only occasionally and weakly on AC, and never on MP (Schroyens, Schaeken & d’Ydewalle, 2001). There are two popular explanations for NCB. One suggestion is that people assume that not- p is more common than p in the real world (e.g. there are more not-blue things than there are blue things), and are therefore more willing to conclude not- p than p (Pollard & Evans, 1980; Oaksford, Chater & Larkin, 2000). An alternative account is that NCB stems from a problem with processing double negations. NCB is most commonly observed on MT and DA inferences, where an affirmative conclusion results from a double negation, for example ‘if not A then 3; not 3; therefore not (not A)’. The reduced endorsement rates in these problems may be due to reasoners failing to convert ‘not (not A)’ into ‘A’.

The APB refers to the observation that participants are more likely to accept inferences with an affirmative premise than a denial premise, particularly when the denial is implicit (such as ‘7’ instead of ‘not 3’, Evans et al., 1995). For example, the inference ‘if not A then 5, 5, therefore not A’ would be accepted more often than the inference ‘if A then not 5, 8, therefore A’, although both are invalid AC inferences. This has been explained as a matching bias (see more below), whereby the premise ‘5’ is more obviously related to the conditional than is the premise ‘8’ (Evans & Handley, 1999).

Another bias which is particularly relevant to this thesis is belief bias. This is the tendency for participants to reason according to their prior beliefs rather than the information at hand (Evans, Barston & Pollard, 1983; Sá, West & Stanovich, 1999). For example, when faced with the syllogism “all things with four legs are dangerous; poodles are not dangerous; therefore, poodles do not

have four legs”, a person biased by their prior beliefs that poodles do have four legs would answer that the syllogism is invalid, whereas a person basing their judgement on the information at hand would answer that it is valid. It is only on items where validity and believability are in conflict (i.e. valid and unbelievable or invalid and believable problems) where belief bias may be shown.

This section has described just a small selection of the biases that have been empirically demonstrated in the reasoning literature, but they are sufficient to show that people do not always reason according to normative standards. The following section discusses this discrepancy in terms of various views on rationality and whether we can expect people to be capable of ‘doing better’ by reasoning according to the normative models.

Are humans irrational?

As demonstrated above, humans often show biases when reasoning. The issue at hand is whether these biases demonstrate irrationality and whether it is reasonable to expect people to do better, or at least to be *able* to do better. This is a vital question for the thesis, because if we cannot expect people to do better, then the question of whether mathematics specifically makes people better at reasoning would be redundant.

There are three positions on human rationality which differ in their perspective on the relationship between descriptive, prescriptive and normative models of human reasoning (Stanovich, 1999). Descriptive models describe and theorise on the reasoning patterns of human beings. Normative models set the ideal standards for reasoning, which if achieved would maximise outcomes for the individual. However, as Harman (1995) and Stich (1990) argued, humans have limited intelligence and limited time, and it would be unreasonable to expect us to act in a normatively rational way given these restrictions. It is this argument that leads to the idea of a prescriptive model. Prescriptive models specify the best we can hope to achieve when reasoning, given the cognitive and often situational limitations that the reasoner must work within. This then replaces the normative model as the standard we should hope to achieve and with which we should compare descriptive models.

As mentioned above, there are three positions on the relationship between descriptive, prescriptive and normative models of reasoning behaviour, and these are shown in Figure 2.2. ‘Panglossians’ do not see that there is any substantial gap between the three models (Stanovich, 1999). They argue that humans reason as well as they can and as well as they should, and that human irrationality is therefore not an issue. The biases described above are explained as either random performance errors, incorrect norm application on the exper-

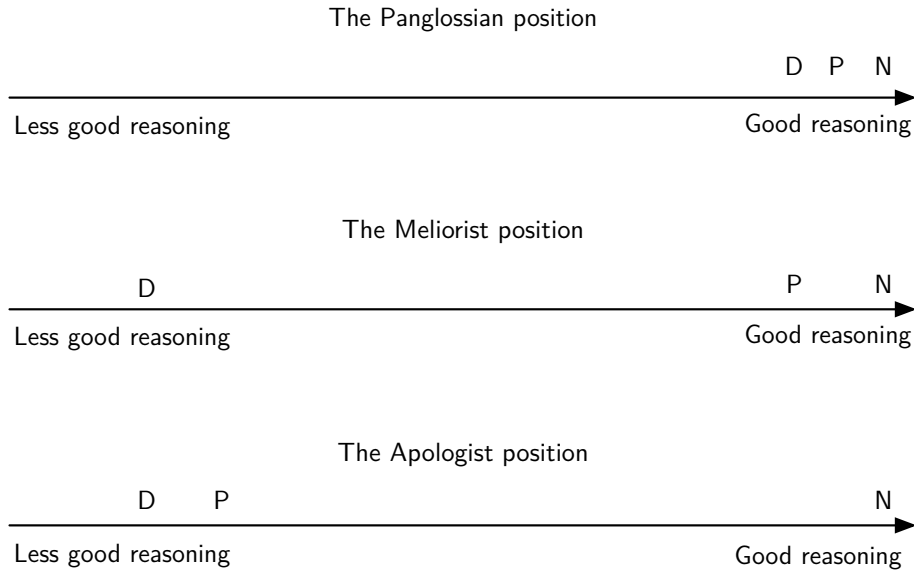


Figure 2.2: The three positions on human rationality. N = normative model, P = prescriptive model, and D = descriptive model.

imenter’s side, or a misunderstanding of the problem by the participant, due to the experimenter being unclear (Stanovich & West, 2000). These explanations are elaborated below. Under no circumstances are the biases considered demonstrations of irrationality in the participants.

‘Meliorists’, on the other hand, argue that the way we actually reason is far from the standard by which we could and should reason. They see the prescriptive model as close to the normative model, while the descriptive model falls somewhat short of the other two. The gap between the descriptive and prescriptive models can therefore be deemed irrationality, because we are failing to do as well as we could reasonably be expected to do (Stanovich, 1999).

The final position is that of the ‘Apologist’. The Apologist position agrees with the Meliorists in that the descriptive model falls short of the normative model, but it differs in that it places the prescriptive model much closer to the descriptive. It argues, then, that while we are not reasoning perfectly, we are doing the best we could hope to do given our cognitive limitations (Stanovich, 1999). Like the Panglossians, the Apologists believe that we cannot accuse humans of irrationality, because we are just about as rational as we could possibly be. The difference between the two is that the Apologists recognise that this is not up to ideal/normative standards, whilst the Panglossians argue that it is.

Throughout his book ‘Who is Rational?’, Stanovich (1999) used data on individual differences in reasoning behaviour to investigate the possible reasons

for the apparent gap between the descriptive and normative models. The possible reasons, as alluded to above, are: random performance errors, incorrect norm application, alternative task construal, computational limitations, and systematic irrationality.

A Panglossian argument used to preserve human rationality is that the apparent gap between normative and descriptive models is due to random performance errors by the participants – perhaps they were momentarily distracted, for example (Stanovich & West, 2000). The problem that Stanovich (1999) identified with this account is that errors are systematic, not random. People consistently make the same mistakes, and the extent of their errors on one task predicts the extent of their errors on another task. Performance is also related to cognitive and personality variables, such as intelligence (or cognitive capacity), which is contradictory to a random error view. Whilst all measures used in psychological research will inevitably be subject to some random errors, or ‘noise’, this is not a sufficient explanation to account for the substantial gap between descriptive and normative models of reasoning.

The incorrect norm argument places the blame for the normative/descriptive gap with the experimenter (Stanovich, 1999; Stanovich & West, 2000). In the heuristics and biases literature, performance is usually compared to a statistical or logical norm, and it can be argued that this is inappropriate. Some would suggest that instead of using norms derived from statistics or logic, the responses given most often by participants should in fact be considered the norm, because people are essentially rational (Stanovich, 1999). Stanovich, however, argued that if we are to use participants’ performance to determine the norm, we should be looking at the responses of the most intelligent.

Intelligence can be considered a consistent ability for effectiveness in different environments and situations (Larrick, Nisbett & Morgan, 1993, also see discussion in Section 2.6). This suggests that the behaviour of the most intelligent individuals reflects ‘better’ or more effective behaviour than that of less intelligent individuals. When the behaviour of the most intelligent individuals is in line with the normative model set by experts, perhaps this is evidence that the normative model is in fact appropriate. If, on the other hand, the most intelligent individuals give a non-normative response, then maybe we should consider revising the normative model as some suggest.

Stanovich (1999) argued that the first scenario is usually the case. He showed that across a variety of tasks, participants were consistent in whether they gave the normative or non-normative response and that there were significant correlations between normative performance and general intelligence or thinking disposition. He argued that if experts and more capable participants agree that the normative response is the correct response, then it very likely *is* the correct

response, and this was the case in the majority of tasks he examined. However, there were a small minority of tasks where this was not the case and in such cases, perhaps there is room for the incorrect norm argument to explain the discrepancy between normative and descriptive models.

Another explanation proposed to preserve human rationality is that participants misconstrue the tasks given to them. The experimenter may be applying the correct norm for the task they intend, and the participants may be processing without error, yet there is still a problem because the participant is responding correctly to a different interpretation of the task. Cosmides and Tooby (1992) have argued that humans have inbuilt algorithms for solving reasoning tasks that have been encountered through our evolutionary history, and that when problems are framed to elicit these algorithms, we actually reason quite well (Cosmides, 1989; Cosmides & Tooby, 1992).

However, Stanovich (1999) used the same argument as for the incorrect norm issue to show that the normative task construal, in most cases, should be considered the correct one. Those of higher general cognitive ability tended to construe tasks in the manner intended by the experimenter. Again, this was not the case for every task so there is some room for the argument that tasks are presented in an ambiguous way, but it is the case for the majority of tasks.

One could argue that it is incorrect to use the behaviour of high ability individuals to justify traditional norm applications and task interpretations. The argument seems somewhat circular – high ability individuals are considered high ability because they tend to give normative responses on intelligence tests, and we therefore assume that because they give normative responses to other tasks as well, the norms must be correct. However, there are very compelling reasons to consider high cognitive ability individuals’ responses as ‘better’ – these people get better degrees (Farsides & Woodfield, 2003), have higher job performance and success (Deary, 2001; Judge, Higgins, Thoresen & Barrick, 1999), higher incomes (Ashenfelter & Rouse, 1999), and, surprisingly, they live longer (Deary, 2008). As Frederick (2005) noted, this issue could be clarified by giving the responses and intelligence scores of previous participants to new participants. If the new participants were influenced by the responses of high intelligence individuals, it would indicate some degree of consensus in the idea that intelligent individuals tend to make better choices and that we want to do as they do.

The idea of cognitive limitations being the cause of the reasoning gap has some support, although it is not a complete explanation. Intelligence is limited and related to reasoning performance (Evans et al., 2007; Sá et al., 1999; Stanovich & West, 2008). However, in Stanovich’s (1999) analyses, once intelligence was controlled for there was substantial remaining variance in reasoning per-

formance, and this was related across tasks and predictable from individuals' thinking dispositions (West, Toplak & Stanovich, 2008).

Thinking dispositions refer to an individual's cognitive style – traits such as the extent to which they are willing to put effort into solving a task, or to change their beliefs based on new evidence or a lack of coherence. The evidence suggests that while there is some blame for the descriptive/normative gap to be put on cognitive limitations, there is also a role for dispositions (Stanovich, 1999, also see Section 2.6). This leaves us with the final explanation for the normative/descriptive gap – systematic irrationality. If it is the case that individuals vary in terms of how much effort we are willing to put into a task, and that the amount of effort we put in determines, in part, the quality of our reasoning, then it could be said that we are irrational when we do not choose to put all of our cognitive resources into computing a response that will benefit us.

Stanovich (1999) demonstrated that none of the excuses given by Panglossians or Apologists for the normative/descriptive gap were sufficient. Random performance errors, incorrect norm application, alternative task construal, and computational limitations were all shown to be either insufficient or inapplicable to the majority of tasks where biases are shown. This leaves only the Meliorist explanation – systematic human irrationality. Even when we are capable of reasoning normatively, we do not always do so. Although this may seem to be a very negative view of human rationality, it does leave open the optimistic possibility that we can do better. There is a real gap between the descriptive model of behaviour and the prescriptive/normative model, at least for most individuals on most tasks. It is possible, then, that the study of mathematics might be one way of reducing the gap.

To refer back to the original issue set out in this section, 'better' reasoning when mentioned in this thesis should be taken to mean reasoning closer to the normative (and prescriptive) models, in line with the Meliorist position. In conditional reasoning, a material interpretation could be considered 'best' because it is the normative model, considered correct by logicians. Of the three other interpretations, the defective conditional could be considered second best because although MT is not accepted when it should be (according to the material interpretation) the invalid inferences are also not accepted. Under a biconditional interpretation, all of the inferences are accepted, two valid and two invalid, and under a conjunctive interpretation MP and AC are the only ones accepted, one valid and one invalid. Therefore the biconditional and conjunctive interpretations could be considered equal and 'worse' than the defective interpretation.

Section summary

- Conditional ('if, then') logic is a central aspect of deductive reasoning.
- Humans demonstrate a vast array of heuristics and biases in reasoning, including with conditional problems.
- Stanovich's (2000) individual differences data suggested that a Meliorist perspective on rationality is accurate – on the whole, the standard normative models of reasoning are appropriate competence models and because humans do not always reason to these standards, they thus show systematic irrationality.
- When mentioned in this thesis, 'better' reasoning should be taken to mean reasoning closer to the relevant normative model.

2.4 A brief history of the Theory of Formal Discipline

For millennia it has been assumed that people can be taught thinking skills. Plato was a holder of this belief:

“Those who have a natural talent for calculation are generally quick at every other kind of knowledge; and even the dull, if they have had an arithmetical training [...] become much quicker than they would otherwise have been [...] We must endeavour to persuade those who are to be the principal men of our state to go and learn arithmetic.”

Plato (375B.C/2003, p. 256)

Nisbett (2009) provided a brief history of the Theory of Formal Discipline, describing how the curriculum was extended from Plato's arithmetic to include grammar, logic, Latin and Greek throughout the Roman, medieval and Renaissance periods. This curriculum lasted for centuries and eventually resulted in the English school system of the nineteenth century.

The first challenge to the TFD came at the beginning of the twentieth century with the development of psychology as an academic discipline. Nisbett summarised the harsh rejection of the TFD well by stating that “William James ridiculed the idea that the mind had muscles that could be exercised by arithmetic or Latin” (p.29). The behaviourist movement conjectured that learning was very much limited to stimulus-response links, learned through such mechanisms as positive and negative reinforcement of behaviours (Hergenhahn &

Olson, 2004; Thorndike & Woodworth, 1901; Skinner, 1938), so the idea of skills transfer from one subject to another or to thinking overall was a complete contradiction. One of the first authors to dispute the behaviourists' extreme domain-specific view of learning was Piaget.

Piaget's theory of cognitive development suggested that we develop domain-general skills and knowledge through four stages (Beth & Piaget, 1966; Inhelder & Piaget, 1958; Piaget, 1928; Piaget, 1960; Piaget, 1970). The sensorimotor stage occurs between birth and 2 years of age and is characterised by infants learning about the physical world through their interactions with and sensory experiences of it. The preoperational stage occurs between the ages of approximately 2 and 7, and is characterised by sparse and logically inadequate mental operations. The child can represent objects through words and drawings but cannot perform complex mental operations and cannot take the viewpoint of others.

The concrete operational stage occurs between 7 and 11 years of age and consists of proper use of logical reasoning. Children are able to sort objects by categories, understand relationships between objects, and understand others' points of view, to name a few examples. In the final stage, the formal operational stage, individuals from around 11 years into adulthood become able to think abstractly as well as concretely. This means there is further development of logical reasoning, such as with hypothetical situations. Piaget believed that only formal operational thinkers are able to distinguish the material conditional from the biconditional (Beth & Piaget, 1966).

Each of Piaget's proposed stages is domain-general, and the logical skills developed are applicable across contexts. Inhelder and Piaget (1958) highlighted this belief by stating that "reasoning is nothing more than the propositional calculus itself" (p. 305, translated in Johnson-Laird, Byrne & Schaeken, 1992).

In the 1950s and 1960s came the cognitive revolution (see Neisser, 1967), which argued against behaviourism that in fact it is possible and useful to infer mental processes rather than just the behaviour resulting from them. This meant there was a shift towards investigating the processes behind reasoning. What it also meant was that there was scope for investigating what makes better or worse reasoning.

According to Piaget's theory of cognitive development, everybody goes through the same stages acquiring the same skills (Inhelder & Piaget, 1958), and again, according to behaviourism, everybody would learn the same reasoning behaviour from the same stimulus, and so there is little room for variation in quality of reasoning between individuals or for strategies to improve reasoning beyond the standard stages or stimuli. What cognitivism allows is the possibility that there are different types of processes used in reasoning, that some of these pro-

Each of the cards below has a letter on one side and a number on the other side. Select all those cards, but only those cards, which would have to be turned over in order to discover whether the rule is true.

Rule: If there is an A on one side of the card, there is a 3 on the other side.

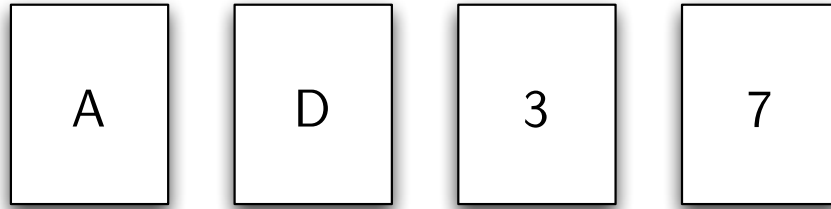


Figure 2.3: The Wason Selection Task.

cesses are more effective than others, and that by investigating these processes it might be possible to alter an individual's reasoning.

Wason was one of the first psychologists to demonstrate how flawed human reasoning can be. He disputed Piaget's idea of the development of formal logical reasoning by showing experimentally that, on the whole, humans are not very logical (Wason, 1968). He also attempted to explain our errors by reference to the cognitive processing responsible for solving reasoning tasks (Johnson-Laird & Wason, 1970; Wason, 1966). Probably the most famous task in the reasoning literature is Wason's four card selection task (Wason, 1966, 1968, see Figure 2.3.) Participants are asked which cards they need to turn over to check whether the rule 'if there is an A on one side of the card, then there is a 3 on the other side' is true. The correct answer is to choose the A and 7 cards, but only around 10% of participants gave this response (Wason, 1968). The majority answered either A only, or both A and 3.

Johnson-Laird & Wason (1970) proposed a detailed model of the cognitive processing involved in dealing with the selection task, including some explanations of where people go wrong. For example, they suggested that reasoners initially focus only on the items mentioned in the rule, A and 3, and consider items not mentioned irrelevant. This leads to the common and incorrect response 'A and 3'. They also proposed that subjects have either no insight, partial insight, or complete insight, and that this determines whether they will select, respectively, only their first intuitions (A and 3), select the cards that can falsify the rule as well (A, 3 and 7) or select *only* cards that can falsify the rule (A and 7). This is a level of theorising about cognitive processes and

individual differences that would not have occurred before the cognitive revolution. As such, this type of work represents a significant step forward in our understanding of human logical reasoning.

Following on from Wason’s groundbreaking work, other psychologists such as Philip Johnson-Laird, Jonathan Evans, Keith Stanovich, Daniel Kahneman and Amos Tversky have described a wide range of other biases in human reasoning (see Section 4.2.1) and provided theories on the cognitive processing that underlies both successful and flawed responses. Current theories on reasoning are described below in Section 2.6. What is important here is that current theories do allow for differences in processing of reasoning problems, and hence the possibility of improving reasoning.

2.5 Research into thinking skills and their relation to mathematics

The TFD makes two claims: 1) thinking skills are transferable across contexts, and 2) the study of mathematics improves these skills to a greater extent than other disciplines. Research related to each of these assumptions will be reviewed separately, beginning with the transferability of thinking skills.

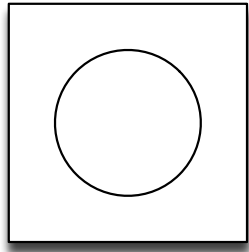
Are thinking skills transferable across contexts?

One of the first tests of the TFD was published by Thorndike and Woodworth (1901), when behavioural psychologists began to doubt the generality of learned skills, advocating instead specific stimulus-response links. Thorndike & Woodworth reported an experiment in which participants were trained in estimating the area of rectangles, and improvement in estimating the area of other shapes was measured.

The results showed that improvement in estimating the area of rectangles was not paralleled by improvement in estimating the area of different shapes. Thorndike and Woodworth (1901) suggested that the estimation of areas is a group of functions rather than a single function, with the specific data (e.g. shape and size) determining which function is necessary. In this view, functions (or skills) are not transferable because they are closely tied to the stimulus, and any change to the stimulus renders the function redundant.

This conclusion was supported in a much more recent and large scale study published by Owen et al (2010). In the study, 11,430 viewers of the BBC programme ‘Bang Goes The Theory’ received six weeks’ online training in reasoning, memory, planning, visuospatial skills and attention. Pre- and post-intervention reasoning ability was assessed by a speeded grammar task, where

a) The circle is not smaller
than the square



b) The square is in the circle

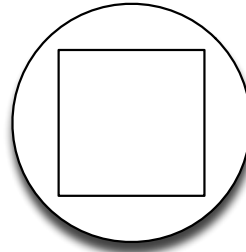


Figure 2.4: An example of a reasoning task used by Owen et al (2010) showing (a) false and (b) true sentences about shape pairs.

participants read a statement about a pair of shapes and judged whether the statement was true. For example, when shown a small circle within a larger square, participants judged whether the sentence ‘The circle is not smaller than the square’ was true (see Figure 2.4).

During the intervention participants practiced three different reasoning tasks, as well as memory, planning and visuospatial tasks, with feedback. In one reasoning task, participants were required to infer weight relationships from images of see-saws with objects at each end in order to select the heaviest object from a choice of three presented separately. In the second reasoning task, participants were required to select the odd one out of four shapes that varied in shape, colour, and fill. The final reasoning task required participants to move crates from a pile with reference to the overall pattern of crates. Although there were improvements in every one of the tasks trained, there was no evidence of transfer to the untrained grammatical reasoning task, which the authors considered to be closely related (Owen et al., 2010). This seems to be evidence that thinking skills are closely tied to context, as Thorndike and Woodworth (1901) proposed.

In another blow to the idea of training thinking skills, Cheng, Holyoak, Nisbett and Oliver (1986) found that even a one semester course in standard logic did not lead to improved performance on four Wason Selection Task problems. The course consisted of around 40 hours of teaching and covered modus ponens, modus tollens, denial of the antecedent, affirmation of the consequent and the distinction between the material conditional and biconditional statements. Students were also taught with both contextualised sentences and formal truth tables. It seems reasonable to expect that after such training participants should be very competent at dealing with conditional reasoning problems, and it is even difficult to imagine a more promising way to improve a students’ logical com-

petency. Nonetheless, there was a non-significant decrease in errors of only 3%. However, this could be due to the measure used. It has recently been suggested that Selection Tasks may not actually measure conditional reasoning ability, particularly thematic (situated in a real-world context) versions (Sperber, Cara & Girotto, 1995; Sperber & Girotto, 2002).

Sperber et al. (1995) have suggested that Selection Task performance is highly influenced by relevance mechanisms that pre-empt any reasoning mechanisms. When faced with a reasoning problem, or any other text, we need to comprehend the meaning intended by the author. In the case of the selection task, participants are asked to judge the relevance of each of the cards to the rule so the judgments of relevance that come from the comprehension process provide an intuitive answer to the problem and there is no explicit need to engage in any further reasoning. This may be the source of the pervasive matching bias described above. This account has been supported in several studies (Sperber et al., 1995; Sperber & Girotto, 2002) and it implies that Selection Tasks do not actually measure reasoning processes at all. Sperber's argument is further elaborated on in Chapter 4.

In a similar study on teaching reasoning, White (1936) investigated the effect of logic training on 12 year old boys' reasoning ability. One class spent an hour per week for three months being taught logic, including deduction, induction and syllogisms, while another class did not. At the end of the three months the students were given a reasoning test that measured, among other things, syllogism validation. The class that had been taught logic scored significantly higher on the reasoning test than the control class. The authors concluded, conversely to Cheng et al. (1986), that logical thinking *can* be taught. Perhaps the difference between the findings of Cheng et al. (1986) and White (1936) is due to the age of the participants, it could be that children's reasoning strategies are more malleable than adults'. Another possibility is that the measure used by Cheng et al. (1986) was not appropriate for measuring development, as elaborated on above (Sperber et al., 1995; Sperber & Girotto, 2002).

Further supporting the idea that reasoning can be taught, Lehman, Lempert and Nisbett (1988) found that the extent to which conditional reasoning was improved in graduate students was a function of their degree studies. Students in psychology, medicine, law and chemistry were tested on conditional reasoning, statistical and methodological reasoning, and verbal reasoning in their first and third years of study. The medical students improved in verbal reasoning and both the psychology and medical students improved in statistical and methodological reasoning. Most relevant here is that the psychology, law and medical students all improved in conditional reasoning as tested by one abstract, one causal framed, and one permission framed Selection Task and one biconditional

Selection Task. However, as mentioned above, it is unclear what skills contextualised Selection Tasks are actually measuring (Sperber et al., 1995; Sperber & Girotto, 2002). The findings should therefore be treated with caution.

Kosonen and Winne (1995) found evidence for transfer after training participants in statistical reasoning. Training in the ‘law of large numbers’ (which states that the greater the sample size, the more representative it is) improved students’ ability to reason with everyday problems that required the law of large numbers in a wide range of domains. Before training, participants were more likely to reason statistically about problems that explicitly referred to probability than those which did not, but this equalled out post-training.

Because their undergraduate and teenage participants did use statistical rules before instruction, although infrequently and without sophistication, Kosonen and Winne (1995) suggested that informal reasoning rules are learned through experience and that these were recruited and refined with the training given, making them transferable. This was also the conclusion of Fong, Krantz and Nisbett (1986), who found that training in the law of large numbers improved participants’ performance even when they were tested with a problem context completely different to that of the training. Furthermore, Fong and Nisbett (1991) found that the transfer was still evident after a two week delay.

In order to test the generality of critical thinking, Sá et al. (1999) measured participants’ belief bias in a verbal (syllogism) task and a non-verbal (height judgement) task. The syllogism task consisted of 24 syllogisms with real-world content, such as ‘all living things need water; roses need water; therefore, roses are living things’ (see Figure 2.5). In the height judgement task participants were shown full length photographs of seated males and females and asked to estimate their heights. In an ecological version, the heights shown reflected actual male-female height differences. In a matched version, the pictures shown were of males and females matched for height, and participants were informed of this and that they should ignore the gender cue when making their judgements.

In each task, participants’ responses could be influenced by prior beliefs and/or logical validity. Belief bias is the extent to which judgements are influenced by prior beliefs over validity when the two are in conflict. The authors found a significant correlation ($r=.209$) between belief bias in the syllogism task and belief bias in the matched height judgement task and there was an even larger correlation between belief bias in the two versions of the height judgement task ($r=.526$). While the significant correlations conflict with an extreme domain-specificity view, like that of Thorndike and Woodworth (1901), they do support the idea that generality is related to similarity between task stimuli; the correlation was stronger between the two non-verbal height judgement tasks than between either of those and the verbal syllogism task.

Believable, valid:	Believable, invalid:
Premises: All fish can swim.	Premises: All living things need water.
Tuna are fish.	Roses need water.
Conclusion: Tuna can swim.	Conclusion: Roses are living things.
Unbelievable, valid:	Unbelievable, invalid:
Premises: All things with four legs are dangerous.	Premises: All guns are dangerous.
Poodles are not dangerous.	Rattlesnakes are dangerous.
Conclusion: Poodles do not have four legs.	Conclusion: Rattlesnakes are guns.
Neutral, valid:	Neutral, invalid:
Premises: All ramadions taste delicious.	Premises: All lapitars wear clothes.
Gumthorps are ramadions.	Podips wear clothes.
Conclusion: Gumthorps taste delicious.	Conclusion: Podips are lapitars.

Figure 2.5: An example of each type of item from the Belief Bias Syllogisms task.

On a similar theme, Toplak and Stanovich (2002) investigated the domain generality of disjunctive (‘either, or’) reasoning. They gave participants nine tasks that all relied on an exhaustive consideration of all possible states of the world given the ‘either, or’ rule (such as the Knights and Knaves problem shown in Figure 2.6). If this is a single skill that generalises to all of the tasks, there should have been large correlations between performance on them. In fact, performance on the tasks showed specificity. This led the authors to concentrate on five of the tasks which indisputably require disjunctive reasoning. Between these tasks, five of the ten correlations were statistically significant, still indicating a considerable degree of domain specificity between tasks which are confidently thought to use the same skill, although within different contexts. Again, this appears to support Thorndike and Woodworth’s (1901) argument that any change in stimuli prevents a skill suitable for another context from being properly applied.

There are two final studies that can inform the question of whether thinking skills are transferable, although both have significant impediments that mean they should not be considered definitive. Sanz de Acedo Lizarraga, Sanz de Acedo Baquedano and Soria Oliver (2010) looked at the effect of a year long teaching intervention on thinking skills in two Spanish schools and Lehmann (1963) looked at the development of critical thinking over four years of college education.

In the Sanz de Acedo Lizarraga et al. (2010) study, the participants were aged 11 to 13 years old and were from two schools in Spain. One school was randomly

Imagine that there are three inhabitants of a fictitious country, A, B and C, each of whom is either a knight or a knave. Knights always tell the truth. Knaves always lie. 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?

Figure 2.6: The disjunctive Knights and Knaves problem used by Toplak and Stanovich (2002), adapted from Shafir (1994).

assigned to the experimental condition and one to the control condition. The experimental condition involved the teaching method ‘Thinking Actively in an Academic Context’, which is designed to promote thinking skills in students. The intervention lasted for one academic year and was implemented for twelve hours per week. Participants completed pre- and post-intervention tests that measured a range of cognitive abilities.

The intervention was shown to improve verbal, abstract and numerical reasoning, creativity, and academic achievement to a greater extent than the conventional teaching method used with the control group. While this may support the generality of thinking skills, there are important drawbacks to the design of the study. The intervention was assigned randomly at the school level rather than the participant level, meaning any differences between the schools that may have influenced the results, such as general ability of the students, were not controlled for. It was also not assigned blindly. This issue is particularly important because the study involved an entirely novel teaching method, both for the teachers and students, which they were told would have the intended benefits. This may have led to increased interest, enthusiasm and expectations of those involved, and this so-called Hawthorne effect¹ may be what was actually responsible for the effects observed.

Lehmann’s (1963) study investigated changes in American college students’ critical thinking and stereotypical beliefs over the course of their higher education studies. Between their first and fourth years, he found a significant improve-

¹The Hawthorne effect refers to the observation that participants in research studies may modify their behaviour simply as a result of being observed, regardless of the manipulation. This was first identified by Henry Landsberger who conducted a study at the Hawthorne Works factory into the effect of lighting levels on productivity. An increase in productivity occurred while the study was being conducted and a decrease occurred when the study ceased. It was suggested that the workers became more productive simply because they were being observed (Landsberger, 1958).

ment in critical thinking, as measured by the American Council on Education's Test of Critical Thinking (American Council on Education, 1953), and a decrease in stereotypical beliefs (i.e. the participants became more flexible and less authoritarian). The students were not specifically trained for these effects, so it might be that the changes came about as a generalisation of something from their education. Alternatively, as there was no control group, it might be the case that these changes occur in all college-aged people regardless of whether they are in education or not. Nevertheless, these two education studies provide some hope for the generality of critical thinking.

In sum, the evidence for the first claim made by the TFD, that thinking skills are transferable across contexts, is inconsistent and therefore fairly weak. While some studies have found some degree of generality of thinking skills, others have found complete specificity.

Perhaps this is because each study has looked at different aspects of thinking and it may be the case that each skill has a different position on a specificity-generality continuum. Statistical reasoning (Kosonen & Winne, 1995) and susceptibility to belief bias (Sá et al., 1999) would appear to be further towards the generality end of the spectrum than disjunctive reasoning (Toplak & Stanovich, 2002) and reasoning about shape area (Thorndike & Woodworth, 1901). A wide range is taught in the 'Thinking Actively in an Academic Context' teaching method and in higher education, so it is unclear exactly what skills are being generalised in the cases of Sanz de Acedo Lizarraga et al. (2010) and Lehmann (1963).

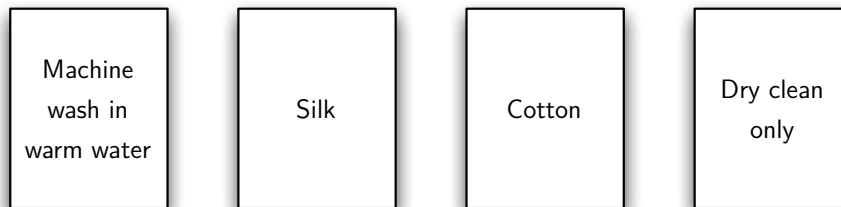
This interpretation of the inconsistencies is in line with Nisbett and his colleagues' suggestion that the difficulty of teaching a type of reasoning depends on the extent to which people already have an intuitive understanding of the rules (Nisbett, Fong, Lehman & Cheng, 1987; Nisbett, 2009). As Kosonen and Winne (1995) also suggested, people may have a basic understanding of statistical laws inducted through experiences, which makes them available for training. Perhaps disjunctive and geometrical (shape area) reasoning are not already within the cognitive repertoire of rudimentary rules and are therefore more difficult to teach in a transferable way.

Does the study of mathematics improve 'thinking skills'?

Moving on to the second assumption of the TFD, that mathematics improves thinking skills to a greater extent than other subjects, the evidence is again mixed. As with the first assumption, Thorndike (1924) published one of the first studies that tested the influence of school subjects on reasoning abilities.

Thorndike (1924) used a pre-test/intervention/post-test design, where the

As part of your job as quality control inspector at a shirt factory, you have the task of checking fabric and washing instruction labels to make sure they are correctly paired. Fabric and washing instruction labels are sewn back to back. Your task is to make sure that all silk labels have the 'dry clean only' label on the other side.



You must only turn over those labels you need to check to make sure the labels are correct.

Figure 2.7: The causal schema conditional problem used in Lehman and Nisbett (1990).

intervention was one year of school education with varying subjects and the pre- and post-tests measured general intelligence (ability on tests of various school-type subjects, such as arithmetic and word analogies and opposites). He found that the subjects taken by students had only a minimal influence on scores on general intelligence tests. French, chemistry and trigonometry were associated with the largest, albeit small, improvements, while arithmetic, geometry and algebra were associated with improvements barely above zero. So again, Thorndike's research paints a bleak picture for the TFD.

However, Lehman and Nisbett (1990) did find some support for a version of the TFD. They tested US undergraduates in their first and fourth years on statistical and methodological reasoning, conditional reasoning and verbal reasoning. They compared students studying natural sciences, humanities, social sciences, and psychology. Although there were no between group differences on any of the measures at the first test or on SAT scores, there were some post-test effects of discipline studied. The social science and psychology students improved dramatically on statistical and methodological reasoning, while the natural science and humanities students improved significantly on conditional reasoning, becoming more material on one abstract, one causal framed (Figure 2.7), and one permission framed Selection Task, and one biconditional Selection Task.

Most importantly, there was a correlation between change in material conditional reasoning and number of mathematics courses taken (Lehman & Nisbett,

1990). As has already been stated, conditional reasoning is an important component of logical reasoning (Anderson & Belnap, 1975; Braine, 1978; Inglis & Simpson, 2008). It falls within the bracket of deductive reasoning, which Polya (1954) considered very important in mathematical proof. The correlation with number of mathematics modules taken was significant over all majors ($r=.31$), but more strongly so when focusing on natural science majors ($r=.66$), who took the most mathematics courses.

The authors concluded that reasoning can be taught, and that different disciplines teach different types of reasoning (Lehman & Nisbett, 1990). This is very promising for the TFD, especially considering how pessimistic some of the findings discussed up to this point have been. However, as mentioned above in the discussion of the Lehman et al. (1988) study, it is unclear what skills Selection Tasks are really measuring (Sperber et al., 1995; Sperber & Girotto, 2002). Again, then, the results must be considered with caution.

Conditional reasoning ability was also investigated by Inglis and Simpson (2008), who compared mathematics and arts undergraduates. They gave the undergraduates a 32 item abstract Conditional Inference Task and observed that mathematics undergraduates performed significantly more in line with the material conditional than the arts undergraduates. Again, this seems to support the TFD, but there were two problems with this study: there was no measure of intelligence, which may have differed between groups and been responsible for the difference in conditional inference scores, and there was no longitudinal component, which makes it impossible to infer development. These problems were addressed in a second study.

Inglis and Simpson (2009a) matched a group of mathematics undergraduates with a group of comparison undergraduates on intelligence scores determined by the AH5 test (Heim, 1969), and again gave them the Conditional Inference Task. Intelligence was positively related to a material conditional reasoning scores and was higher in the mathematics students than the non-mathematics students before the matched sub-groups were selected. Nevertheless, matching the groups for intelligence still left a significant difference between them on the material conditional score, with mathematics students again outperforming the comparison group. In other words, the mathematics group was more normative in their conditional reasoning over and above their higher intelligence scores.

The mathematics students were re-tested at the end of their first year of study to look for development in conditional reasoning ability. There was an average change of only 1.8% towards a material interpretation, which did not approach significance. The lack of improvement leaves two possible explanations for the initial difference between groups on entry to university: either post-compulsory but pre-university study of mathematics is responsible, i.e. A

level²; or those who are more normative in their conditional reasoning are disproportionately filtered into studying university level mathematics. Distinguishing these two possibilities is in essence the key aim of this thesis.

Assuming that the results found by Inglis and Simpson (2008, 2009a) are sufficient to say that mathematics undergraduates are more normative in their conditional reasoning than other undergraduates on entry to university, then the issue to resolve is not whether mathematicians are better at reasoning, but whether mathematics *made* them better at reasoning. Inglis and Simpson's mathematics participants may have outperformed their comparison participants because studying mathematics at A level *changed* their reasoning ability to get them to that point, as the TFD suggests, or because they were always more normative in their reasoning, independently of any mathematics studied, and were filtered into studying mathematics at advanced levels because of this inherent trait. The way to distinguish these possibilities is with a longitudinal study looking at changes in reasoning ability throughout an A level (the first post-compulsory stage of education), where mathematics students are compared to non-mathematics students. Although it is not possible to determine causation without an experimental design (see Chapter 3), it is possible to distinguish the development hypothesis from the pre-disposition hypothesis.

In conclusion, the research reviewed here is unclear on whether the assumptions of the TFD are justified or not. Although it appears that some thinking skills cannot be transferred from the original context in which they were learned, there is some evidence that mathematicians reason more normatively on the Conditional Inference Task. In order to effectively test the TFD, a longitudinal study is required in which sixth form and university students of mathematics are compared to students of another discipline, and each group's development is measured and contrasted. Only then will it be possible to effectively test the two claims of the TFD. The two studies by Inglis and Simpson (2008, 2009), along with the Lehman and Nisbett (1990) study, provide a promising line of investigation in further studies: if the TFD is correct, it appears that conditional reasoning may be a thinking skill that mathematics develops.

Section summary

- The question of whether thinking skills are transferable across contexts or domains has been hotly debated and there is evidence both for and against it.

²A levels are optional two year courses taken after compulsory education in the UK. Students usually take four A levels in chosen subjects and the results are used by universities to select incoming undergraduates. The first year of an A level is called the Advanced Subsidiary (AS) level, and is a qualification in itself.

- There is some evidence that studying mathematics may be related to conditional reasoning ability (Inglis & Simpson, 2008) but not necessarily improvement in it (Inglis & Simpson, 2009a) or improvement in statistical, methodological or verbal reasoning (Lehman & Nisbett, 1990).

2.6 The psychology of reasoning

The TFD suggests that studying mathematics improves reasoning, but it does not elaborate on what cognitive change might actually be taking place that could cause this. However, a huge amount of research has been conducted into the psychology of reasoning independently of any relation to mathematics, and there are currently several dominant theories that may help to identify possible mechanisms. The theories are reviewed in three sections depending on how many reasoning systems or process-types they posit: single-process, dual-process and tri-process theories.

Single-process theories

It is worth noting here that single-process theories are not necessarily in competition with each other or with multiple-process theories. At least some single processes theories only attempt to explain certain types of reasoning rather than reasoning as a whole, and they could also be considered explanations of one of the process-types in multiple-process theories. This should become more clear as each of the theories is discussed below.

Mental Models Theory

The mental models theory of reasoning (Johnson-Laird & Byrne, 1991; Johnson-Laird & Byrne, 2002; Johnson-Laird, 2008) suggests that participants create in their mind a model of the task premises as they understand them, and they reason from their model of the task. Mental models represent possible states of the world given that the major premise is true. For example, when given the conditional statement ‘if it rains then I take an umbrella’, the initial model might be:

rain take umbrella
...

The first line indicates a model in which both ‘rain’ and ‘take an umbrella’ are true. The ellipsis indicates an implicit model that reminds the reasoner that *p* may not be true, without actually building the model explicitly. This oc-

raining	take umbrella	‘if it rains then I take an umbrella’
t	t	t
t	f	f
f	t	t
f	f	t

Table 2.4: Truth table for ‘if it rains then I take an umbrella’ where t = true and f = false, assuming the material conditional.

curs because building mental models requires working memory, which is limited (Johnson-Laird & Byrne, 1991; Johnson-Laird & Byrne, 2002; Johnson-Laird, 2008).

Each mental model represents a ‘true’ row in a truth table, although as already noted the models may not represent all of the information in the corresponding row, and there may be rows which are not represented explicitly at all (Johnson-Laird & Byrne, 2002). The full material truth table for this conditional statement is shown in Table 2.4.

Reasoners may construct multiple models from the premises to represent all possibilities as they see them, and from these models they draw a conclusion that holds true in all of them. For example, a reasoner may see the possibilities as:

rain *take umbrella*
no rain *don’t take umbrella*
...

If the minor premise presented is ‘take umbrella’ then based on their mental models the reasoner may conclude that ‘rain’ necessarily follows, because there are no models in which that conclusion is not true, but this would be an invalid affirmation of the consequent deduction.

The last step in the reasoning process proposed by mental models theory is that the reasoner attempts to think of counterexamples to their models and conclusion. If they do not find any counterexamples they accept their conclusion, otherwise they may reconstruct their models to include any counterexamples and then draw a new conclusion.

To further demonstrate this theory, we can imagine how a person might deal with the abstract conditional statement ‘if p then q ’. The initial model might be:

p q

...

The first line indicates a model in which both p and q are true and the ellipsis indicates an implicit model that reminds the reasoner that p may not be true.

If the minor premise following ‘if p then q ’ is the modus ponens inference ‘ p ’, then participants can easily draw the conclusion ‘ q ’ from their initial model. Similarly, if the minor premise is the affirmation of the consequent inference ‘ q ’, then it would be easy for participants to incorrectly conclude ‘ p ’. In the case of modus tollens, though, the initial premise is not sufficient to draw a conclusion. When the minor premise reads ‘ $\text{not } q$ ’, the reasoner must flesh out their implicit model in order to draw any conclusions. If a reasoner completely builds all models that are possible under the conditional rule, they will end up with:

p	q
$\text{not } p$	q
$\text{not } p$	$\text{not } q$

In this case, the premise ‘ $\text{not } q$ ’ is true in the last model, in which ‘ $\text{not } p$ ’ is also true, and so this is the conclusion drawn. The participants would then search for a counterexample to this conclusion, a model in which the conclusion could not be true. If they fail to find such a model, they accept the conclusion.

Errors in reasoning occur when participants either fail to flesh out their implicit mental models, or when they miss a valid counterexample (Johnson-Laird & Byrne, 1991; Johnson-Laird & Byrne, 2002; Johnson-Laird, 2008). This may occur because of working memory limitations and/or lack of effort. As such, there are ways to improve reasoning under the mental models theory: increase working memory capacity, aid working memory with task design, or increase the effort participants put into their reasoning. In terms of how mathematics specifically might improve reasoning ability, an increase in working memory capacity or a change in thinking disposition (such as increased enjoyment of effortful thinking), might result from studying mathematics.

Mental Logic Theory

Mental logic theory (Rips, 1989; O’Brien, 2009; O’Brien & Manfrinati, 2010), also known as mental rules theory, inference rules theory, and natural deduction theory, suggests that humans have an inbuilt logical rule system. This is somewhat similar to Piaget’s proposition (Inhelder & Piaget, 1958). When we solve a reasoning task, according to mental logic theory, we follow a series of logical steps to reach a conclusion. This is the case for any type of reasoning task, whether it be contextualised or abstract, familiar or novel. The system

is innate, as are the rules that it uses. However, whilst we follow a pattern of formal logical deduction, our system and its rules are not perfect and there is much room for error, which is why humans display as many biases as we do. For example, we may have innate access to the modus ponens deduction but not the modus tollens deduction.

According to the mental logic theory, the way in which reasoning may be improved would be to work on the accuracy of the natural deduction system, presumably through the study of formal logic since it is of the same structure as our innate system. This suggests that mathematics may actually be a fairly good way to improve generalised reasoning, because it exposes students to logic in the form of mathematical proofs, and in some instances it may even teach logic directly.

Pragmatic Reasoning Schemas Theory

Cheng and Holyoak (1985) proposed that we reason based on pragmatic reasoning schemas. These are context-sensitive knowledge structures induced from everyday experiences. They relate to rule sets such as permissions, obligations, and causations and they allow us to deal effectively with the sorts of reasoning situation we encounter daily. For example, if somebody lends us money, our obligation schema is invoked to tell us that we are obliged to pay them back.

Rules not related to our previous experiences, including abstract rules and rules set in a context with which we are not familiar, will not invoke a schema and will therefore be very difficult to solve. If we do not have a relevant schema we must rely on logical rules, which only few of us are said to be competent with (Cheng & Holyoak, 1985). This is supported by the finding that 81% of participants could correctly solve a thematic version of the Selection Task, while only 15% could solve an abstract version (Johnson-Laird, Legrenzi & Legrenzi, 1972). However, Sperber et al. (1995) showed that this is not always the case, and at the very least the pragmatic reasoning schemas theory is not a comprehensive theory of reasoning if it only attempts to explain reasoning in specific contexts and defers to theories such as mental logic to explain reasoning in non-schema-eliciting contexts.

An important feature of pragmatic reasoning schemas is that they are not always logically valid, but are instead heuristics that help us to solve most reasoning tasks we would come across in day-to-day life. For example, a permission schema provides these four rules:

- 1) If an action is to be taken, then the precondition must be satisfied.
- 2) If an action is not to be taken, 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, then the action must not be taken.
(Cheng & Holyoak, 1985, p. 397)

This schema would be elicited given the conditional statement ‘if someone drinks alcohol, then they must be over 18.’ However, rule 3 is not logically consistent with rule 1. It may be the case that if someone is to drink alcohol they must be over 18, but it is not necessarily the case that if someone is over 18 they may drink alcohol. It might be the case that they are pregnant or about to drive a car, in which case the precondition of being over 18 would be necessary but not sufficient for the action to be taken. The problem occurs because schemas involve words such as ‘may’, whereas conditional logic is concrete.

Pragmatic reasoning schemas theory explains poor reasoning performance in terms of lack of an appropriate schema for the task. Abstract tasks such as the Conditional Inference Task are too far removed from everyday experience for participants to be able to deal with them by applying a schema.

This is a problem both for the theory and for the purpose of finding a mechanism for the TFD: firstly, some people can do the Conditional Inference Task very well, even those with no formal training in conditional logic can score at above chance levels (Inglis & Simpson, 2009a), and secondly, the theory doesn’t provide any potential mechanism by which abstract reasoning could improve. The latter issue isn’t necessarily a problem for the theory itself – perhaps it is not possible to improve in abstract reasoning – but it does mean that the theory is not very useful in the current context.

Dual-process theories

Dual-process theories posit that humans have two types of cognitive processes: heuristic based Type 1 processes and analytic based Type 2 processes (Evans, 2003; Evans, 2007; Evans, Handley & Bacon, 2009). Type 1 processes are automatic, fast, effortless, non-conscious and evolutionarily old. They are shared with other animals and among many things, they allow us to navigate our environment and filter out all of the irrelevant information we are surrounded with (Evans, 2003). For example, they allow us to walk through a room avoiding obstacles without consciously processing every detail of them. They are sometimes referred to as The Autonomous Set of Systems (TASS, Stanovich, 2004) which emphasises the important point that many different systems are encompassed as being Type 1 process based.

Type 2 processes on the other hand are deliberate, slow, demanding of working memory, conscious and evolutionarily more recent. They allow us to com-

plete complex and evolutionarily novel tasks, such as, for example, setting up a DVD player or thinking hypothetically (Evans, 2003). Here, the processes will be referred to as Type 1 and Type 2.³

Dual-process theories are not specific to reasoning, but rather they describe two types of processing that underlie all cognition. As such, various dual-process theories have been used to investigate a wide range of phenomena, including mathematical thinking (Gillard, Van Dooren, Schaeken & Verschaffel, 2009a; Vamvakoussi, Van Dooren & Verschaffel, 2012a, 2012b), social persuasion (Petty & Cacioppo, 1986), fear of death (Pyszczynski, Greenberg & Solomon, 1999), memory (Brainerd & Reyna, 2002; Reyna & Brainerd, 1995), moral judgements (Greene, Sommerville, Nystrom, Darley & Cohen, 2001), and self-esteem and stereotypes (Greenwald & Banaji, 1995).

It was mentioned above that some single-process theories could be considered explanations of one of the process types in dual-process theories. Pragmatic reasoning schemas theory with its context-specific heuristics, for example, could fit into Type 1 processes here, while mental logic theory's proposition of a series of logical steps could be seen as an explanation for how Type 2 processes operate. The important thing to remember here is that the various theories do not have to be seen as competitors. Support for one is not always a contradiction to another.

Turning back to dual-process theories, an important issue is how the two process types interact. On the majority of reasoning tasks, the two process types would come to the same conclusion. Take for example a valid syllogism with a believable conclusion:

All fish can swim.
Tuna are fish.
Therefore, tuna can swim.

In this case, if Type 1 processes were used they may decide that the syllogism is valid because the conclusion is believable and if Type 2 processes were used they may decide it is valid because it is logically sound. So, the output of the systems would be the same, even though the processing by which they come to their outputs is different. However, on some problems the systems may come to different conclusions. The valid syllogism with an unbelievable conclusion in Figure 2.8 can demonstrate this effect.

In this case, Type 1 processes may quickly decide that the syllogism is wrong

³Although many sources refer to them differently, e.g. System 1 and System 2 (Stanovich, 1999; Evans, 2003), TASS and analytic (Stanovich, 2004), heuristic and analytic (Evans, 1984; Evans, 2006), experiential and rational (Pacini & Epstein, 1999), and associative and rule-based (Sloman, 1996).

All things that are smoked are good for the health.
Cigarettes are smoked.
Therefore, cigarettes are good for the health.

Figure 2.8: A valid and unbelievable syllogism.

because the conclusion strikes the reader as unbelievable. Slow and rational Type 2 processes on the other hand may decide that the syllogism is valid because, as above, it is logically sound. Dual-process theories must account for how such potential conflicts are dealt with, and there are several strands that differ in their explanations: parallel-competitive, pre-emptive, and default-interventionist theories.

Parallel-competitive theories

Parallel-competitive theories (e.g. Sloman, 1996) suggest that the two systems run simultaneously from the start. If they come to the same conclusion processing ends and the conclusion is output. If the processes conflict, the Type 2 processes can override Type 1 processes, but because Type 1 processes are so much faster they often win the competition. Sloman (1996) suggested that participants are always aware of both responses when there is a conflict and he referred to this as Criterion S, for ‘Simultaneous Contradictory Belief’. Although the conflicting responses are not both strong enough to be acted upon, they are both compelling to some extent, and this creates conscious conflict in the mind of the reasoner.

From the view of parallel-competitive theories, there are two ways in which reasoning might be improved. One way would be for Type 2 processes to become more efficient so that they more often win the race with Type 1 processes and so more often determine the output. Another way would be for the Type 1 heuristics that sometimes lead to fallacious reasoning to be altered, so that Type 1 processing can more effectively deal with reasoning problems itself, and the race becomes less important.

The problem with the first solution is that the speeds of the systems are so inherently different that it seems unlikely that the slow and resource-demanding Type 2 processes could ever get to the point of being as fast as the automatic and effortless Type 1 processes (Evans, 2003). The most promising mechanism for the TFD to operate via under parallel-competitive theories is a change to some relevant Type 1 heuristics. Although some heuristics may be innate, some are also learned from experience in a gradual manner (Chen & Chaiken, 1999), so it is plausible that the study of mathematics could have some gradual influ-

ence on them.

Pre-emptive theories

Pre-emptive models argue that it is decided at the outset of a task, via its superficial characteristics, which system will be used to solve it. This means that there is never a conflict between system outputs. An example model of this type is the selective scrutiny model of Evans, Newstead and Byrne (1993), which was developed to account for belief bias in syllogistic reasoning.

Belief bias occurs when a participant judges the conclusion of a syllogism based on whether or not it is believable rather than whether or not it is logical (Evans et al., 1983; Sá et al., 1999, also see Section 2.3). Take for example the syllogism given above (Figure 2.8) where it is logically concluded from the premises that cigarettes are good for the health. A person biased by their prior beliefs would answer that the syllogism is invalid, whereas a person ignoring their beliefs and concentrating on the logical steps of the argument would answer that it is valid. The selective scrutiny model suggests that ‘belief comes first’ so that when a conclusion is believable, it is decided at the outset that it is correct and when a conclusion is unbelievable, it is decided that Type 2 processes should evaluate it more thoroughly.

It is difficult to identify a way in which reasoning might improve under the pre-emptive type of theory. If the task characteristics determine which system is used, perhaps the only way to improve reasoning would be to alter the heuristics that evaluate the task so that they become more conservative, and more often judge that a task requires Type 2 processing.

Default-interventionist theories

Default-interventionist models suggest that Type 1 processes are always used as the default method to solve any task, but in some cases Type 2 processes may override them.

An example is the heuristic-systematic theory proposed by Chen and Chaiken (1999). It proposed that reasoners have the aim of expending the minimum effort necessary whilst also striving for accuracy. This means they use Type 1 processes as much as possible, but will engage Type 2 processes when necessary for confidence. The sufficiency principle states that individuals hold a continuum of confidence about their judgements, with one point on the scale relating to their actual confidence in their judgement, and another relating to the desired level of confidence. Reasoners will use Type 1 processes as far as possible, but if the actual confidence gained by this does not reach the desired confidence level, they will use Type 2 processes to close the gap.

Another default-interventionist model is that proposed by Evans (2006). In

Evans's (2006) model (shown in Figure 2.9) Type 1 processes output an answer first, and then some form of evaluation occurs. Often it is decided that the output is plausible and so it is accepted and given as the person's response. However, it is occasionally decided that the Type 1 response is not satisfactory, and the problem is re-processed with Type 2 processes.

This approach assumes that we are cognitive misers as much as is possible, but that we have a desire to respond correctly and that there is a mediating evaluation mechanism which determines when the minimum effort response is not sufficient. The feeling of rightness (FR, Thompson, 2009) is a meta-cognitive process that can explain how the interaction between the systems is moderated in the case of the default-interventionist model.

The output of every cognitive process is proposed to come with an associated FR – an intuition about whether the output is correct or not possibly based on the fluency with which the output was computed (Thompson, 2009). The stronger the FR, the more likely it is that an output will be rationalised with shallow analytic processing rather than re-processed analytically. In the case of

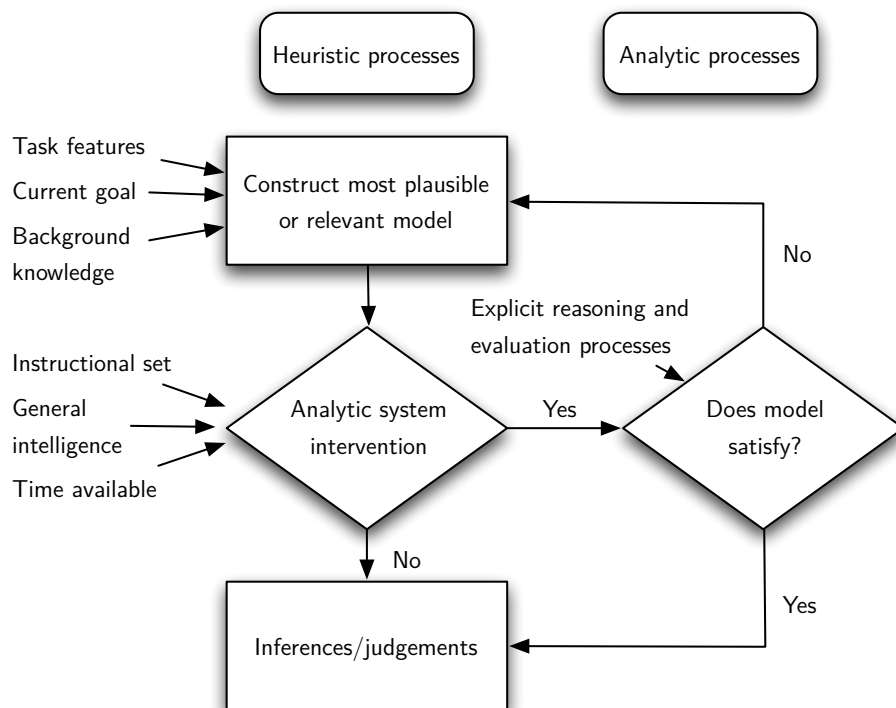


Figure 2.9: Evans's (2006) default-interventionist model. The model uses the terms 'heuristic' and 'analytic' to refer to Type 1 and Type 2 processes respectively.

1. A bat and a ball costs \$1.10 in total. The bat costs \$1.00 more than the ball.
How much does the ball cost?
2. If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?
3. In a lake there is a patch of lily pads. Every day the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake?

Intuitive answers: Q1 = 10 cents, Q2 = 100 minutes, Q3 = 24 days.

Correct answers: Q1 = 5 cents, Q2 = 5 minutes, Q3 = 47 days.

Figure 2.10: Frederick's (2005) three item Cognitive Reflection Test.

the default-interventionist model, it can be supposed that the Type 1 process output has an associated FR, and if this is high, the output is accepted, but if it is lower than some threshold, the analytic system will conduct further processing.

From the view of a default-interventionist model, there are two ways in which mathematics might improve reasoning performance: there might be some change to the threshold for analytic system intervention, so that it is more likely to be engaged; or it might be that the analytic system becomes more efficient, so that when it is engaged it is more likely to compute a correct response. Of course, it could also be a combination of these two. The Cognitive Reflection Test (CRT, see Figure 2.10), developed by Frederick (2005), provides a good way of testing the first possibility.

The three items on the CRT prompt intuitive answers, which quickly spring to mind but are incorrect. These are presumably Type 1 processing default responses, and if a person responds with them, it is assumed that they have not engaged Type 2 processes. If a person does engage Type 2 processes on the task, they will almost certainly give the correct response because once you stop to think about it it is easy to see that the first answer is wrong and easy to calculate the correct answer (Frederick, 2005). Therefore, a person's answer to each question gives a good measure of whether or not they have engaged Type 2 processes.

It may be useful to demonstrate the default-interventionist model shown in Figure 2.9 using the CRT as an example. Upon reading question 1, the answer '10 cents' springs to mind – this is the most plausible model. Because

the answer comes with such a high FR, it is likely that it will be decided that the analytic system need not intervene, and the answer given will be ‘10 cents’. However, if a person has a lower FR, a more conservative threshold for FR, is more intelligent, or is prompted by the instructions that the task is difficult (see Alter, Oppenheimer, Epley & Eyre, 2007), for example, then they may decide that the analytic system should intervene.

When explicit reasoning and evaluation occurs, it is quite likely that the mistake will be spotted. One merely has to work through the simple arithmetic to spot the problem: ‘if the ball costs 10 cents and the bat costs \$1.00 more than the ball then the bat costs \$1.10, and in total they cost \$1.20’. By now the problem has been spotted and the reasoner decides that the model does not satisfy. A new (and usually correct in the case of the CRT, Frederick, 2005) model is then created, and the answer will either be accepted without any further analytic intervention, or the new model will be evaluated analytically and deemed to satisfy. The new answer (usually ‘5 cents’) is then given.

In Frederick’s (2005) data from 3,428 people over 11 different studies, the mean number of CRT items answered correctly was 1.24 out of 3. Over all of the studies, only 17% of participants got all three questions correct, 23% got two correct, 28% got only 1 correct, and 33% did not answer any questions correctly. Even in the highest scoring sample, from Massachusetts Institute of Technology, only 48% of participants answered all three items correctly. The CRT then, is very difficult. This is presumably because the intuitive responses come to mind so easily that they have a high FR and the questions are rarely dealt with by Type 2 processes.

It is possible but rare to give non-intuitive incorrect responses (Frederick, 2005), and in this case it may indicate that Type 2 processes have been engaged, rejected the default response, but failed to compute the correct response. Although this appears to be useful for differentiating engagement of the analytic system and efficiency of the analytic system, it happens with such low frequency (because the arithmetic is so simple) that it is unlikely to be useful (Frederick, 2005).

In sum, the CRT provides a good indication of Type 2 process engagement. Therefore, it can be used to test the possibility that studying mathematics increases the likelihood of Type 2 process intervention. If mathematics does improve reasoning ability but not the likelihood of Type 2 process involvement, then presumably the change has come from more efficient Type 2 processing, according to the default-interventionist perspective.

Comparing the three strands of dual-process theory

Gillard (2009) conducted several experiments to distinguish between the

three strands of dual-process theory using probability tasks. Participants were shown an image of two boxes of black and white marbles and asked which box gave the highest probability of picking a black marble. Participants often display so-called ‘denominator neglect’ and choose the box with the highest frequency rather than preferable ratio.

Gillard varied the ratios and absolute frequencies to construct congruent and incongruent trials (Figure 2.11). On congruent trials, the box with the highest probability of picking black also had the highest frequency of black marbles, e.g. $1/3$ compared to $4/7$. On incongruent trials, the box with the highest probability of choosing black had the lowest frequency of black marbles, e.g. $2/3$ compared to $3/7$. In one experiment time pressure was manipulated, in another working memory was manipulated, and in a third the proportion of congruent and incongruent trials was manipulated, to investigate whether any of these factors would influence which system was used. Reaction times were analysed and compared to predictions stemming from each theory.

Over all three experiments, Gillard (2009) found consistent support for the FR-moderated default-interventionist account. She argued that depending on the strength of the FR, Type 2 processes can range from very minimal activation in the form of simply accepting the heuristic output, to very strong activation that might be a complete reformulation of the problem. FR, therefore, determines the extent of analytic system input, not merely whether it will be activated or not.

Gillard (2009) also proposed that analytic processing can be strengthened as it progresses. If minimal analytic intervention turns up information that conflicts with the heuristic model, the analytic processing will be increased. Therefore, even if the initial FR was high and analytic processing low, it is still possible that the heuristic output could be completely overthrown.

What this means for the current purpose of identifying mechanisms that the

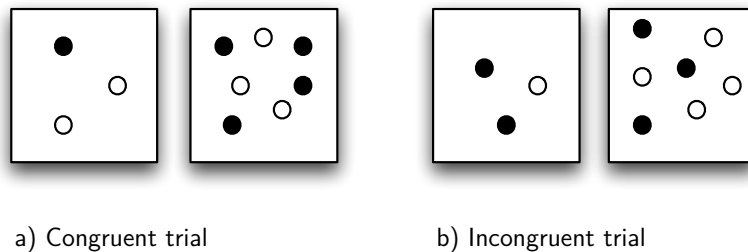


Figure 2.11: An example of a) a congruent and b) an incongruent trial of the denominator neglect task used by Gillard (2009).

TFD may operate by is that as far as dual-process theories of reasoning are concerned, the default-interventionist-with-FR model may be the most promising place to look. As discussed above, the ways for reasoning to improve according to this view are for the analytic system to be used more often or for the analytic system processing to become more efficient when it is used. Reasons for the analytic system to be used more often would be a general lowering of the FR associated with heuristic outputs, or for the threshold for sufficient FR to become more conservative. Either of these changes would result in the analytic system more often being engaged. The CRT can determine whether the analytic system is engaged or not, but not necessarily whether it was because of low FR or conservative threshold. It may even be the case that it is impossible to distinguish these two accounts empirically because each would have the same effect on observable behaviour – in either case more frequent slow analytic processing would be expected.

A tri-process theory

Keith Stanovich has long been an advocate of the dual-process approach to reasoning, but recently he has gone a step further to propose a tripartite model (Stanovich, 1999; Stanovich, 2009a). He suggested that Type 2 processing can actually be divided into two further types – in this case, those of the algorithmic level and those of the reflective level. The algorithmic level can be thought of as the computational element to Type 2 processes, while the reflective level is the dispositional element (Stanovich, 2009a). The algorithmic level is subordinate to the reflective level in that the reflective level determines when the algorithmic level will override Type 1 processes. Type 1 processes in this theory are no different to the dual-process theories discussed above.

The tripartite theory could be considered analogous to the FR-moderated default-interventionist theory discussed above. There it was suggested that Type 1 processes are used by default and the output has an associated FR. When the FR is less than some threshold, Type 2 processes intervene to conduct a more rigorous analysis of the problem. In the case of the tripartite theory, Type 1 processes are again the default processing method, but it is the conscious reflective level of cognition rather than the intuitive FR which determines when to use more rigorous analytic processing. When the reflective level deems it necessary, the algorithmic level takes over.

One of the reasons to propose a tripartite theory of this structure is that ability with critical thinking measures has been shown to be somewhat separate from general intelligence (Stanovich & West, 2008), and both thinking disposition and general intelligence explain unique variance in reasoning ability

(Stanovich & West, 1997; Toplak & Stanovich, 2002). This suggests that general intelligence and thinking dispositions map onto different types of cognition – the algorithmic and reflective levels, respectively. This also means that behaving rationally depends on more than just analytic system capacity – it also depends on the disposition to put effort into using one’s algorithmic capacity, in line with the Meliorist position on human rationality.

Implications for the TFD

Assuming that there are indeed three types of processing in reasoning – heuristic, algorithmic, and reflective – via which type might the TFD operate? The possibilities are that a) studying mathematics changes Type 1 processing heuristics, b) studying mathematics improves algorithmic efficiency or adds new algorithms, and c) studying mathematics alters the reflective level to make the individual more keen to put effort into reasoning tasks.

If System 1 heuristics are the root of differences between mathematicians’ and non-mathematicians’ reasoning behaviour, then the best methods for identifying this would be eye-tracking, reaction times, or speeded tasks. In these ways it is possible to separate initial intuitions from slightly later analytic processing, whereas non-speeded accuracy based tasks such as the standard Conditional Inference and Belief Bias Syllogisms tasks do not allow such a distinction.

With eye-tracking it is possible to identify which aspects of a task draw the participant’s attention first and hold it for longest, and with reaction times it is possible to infer which type of processing determined an answer, based on the speed with which it is given. Similarly, speeded tasks allow us to see how a participant will respond when they do not have time to effectively engage Type 2 processes.

If studying mathematics changes one’s intuitions when faced with a reasoning task, we would expect to see differences between mathematicians and non-mathematicians in conditional inference accuracy even when each item has a short time limit. If mathematicians were to outperform others when they only have, say, 5 seconds to answer each item, we could infer that their advantage lies with Type 1 processing.

If the TFD operates via the algorithmic level, we would need to identify this through constructs such as general intelligence and executive functions. General intelligence is a fundamental aspect of cognition – it contributes to an individual’s ability with any cognitive task. When individuals complete a series of tests measuring different cognitive abilities, there tends to be large correlations between performance on each one, and this is thought to be due to the single underlying factor of general intelligence (g , Spearman, 1927; Jensen, 1998, see Figure 2.12).

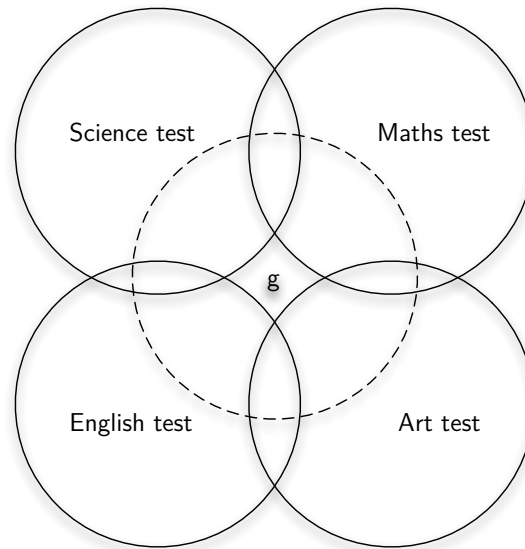


Figure 2.12: An illustration of g as the shared variance between cognitive abilities, such as school subjects.

Intelligence and education are strongly related. Those of higher intelligence tend to do better in school and stay in education longer (Neisser et al., 1996). Looking at the relationship in the other direction, we see that formal schooling develops intellectual abilities and a lack of schooling or schooling at weak institutions can have significant negative effects on intelligence scores (Neisser et al., 1996). As noted in Section 2.3, people of higher intelligence also get better degrees (Farsides & Woodfield, 2003), have higher job performance and success (Deary, 2001, Judge et al., 1999), receive higher incomes (Ashenfelter & Rouse, 1999), and live longer (Deary, 2008).

It seems from this that intelligence may be a factor in the development of reasoning skills. It is already known that the students studying mathematics may have a higher group mean intelligence than those studying English and that intelligence scores are related to material conditional reasoning (Inglis & Simpson, 2009a). If the TFD operates via Stanovich's (2009a) algorithmic level, intelligence seems like a possible mechanism – perhaps studying mathematics increases one's intelligence and that in turn changes one's behaviour with reasoning tasks such as the Conditional Inference Task.

However, despite some malleability in intelligence due to schooling, it is generally agreed in the intelligence literature that g is stable in adulthood both over time (Larsen, Hartmann & Nyborg, 2008; Jensen, 1998; Reeve & Lam, 2005; Rönnlund & Nilsson, 2006) and environmental change (Locurto, 1990),

so although it is possible, it is unlikely that studying mathematics at advanced levels could increase intelligence. Nevertheless, if intelligence were to increase through the study of mathematics, it is quite plausible that this would lead to improvement in general reasoning skills, so g is a worthwhile factor to consider in a study looking for mechanisms of reasoning improvement.

Raven's Progressive Matrices (Raven, Raven & Court, 1998) is a non-verbal intelligence test that is thought to be the best single measure of g (Jensen, 1998). The task is a series of matrices, each of which show a pattern with one piece missing, and participants are asked to select the missing piece from a choice of eight possibilities. An example item is shown in Figure 2.13. If intelligence is a factor in reasoning improvement, Raven's Matrices should be a good measure to reflect that. It should be noted that intelligence tests, including Raven's Matrices, have been shown to be susceptible to repeat testing effects (Ruston & Jensen, 2010; Bors & Vigneau, 2003). In the case of comparing participants from different subjects for improvement, though, it is the difference between groups in degree of improvement that is important, not the absolute extent of improvement. In other words, repeat testing should affect both groups equally (except in the case of a selection-maturation effect, see Chapter 3), and any difference between groups in the extent of improvement may be considered a real change in g .

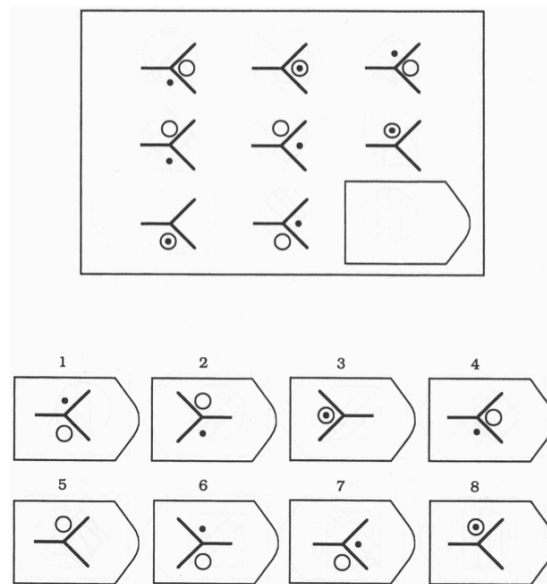


Figure 2.13: An example item from Raven's Advanced Progressive Matrices. Participants are asked to select which of the eight numbered pieces correctly completes the grid.

Another algorithmic level construct is executive function, which is actually a group of cognitive functions: working memory (including capacity for, monitoring of, and updating of information being held in mind), inhibition of irrelevant stimuli or dominant responses, and shifting attention between tasks or processes (Miyake, Friedman, Emerson, Witzki & Howerter, 2000). These abilities are all proposed to be involved in how we regulate our cognitive processes in order to reach a goal. When we encounter a cognitive task, particularly a novel task, these skills guide the way we approach and solve it (Banich, 2009). There is dispute over whether the three functions have the same underlying mechanism (Banich, 2009), and although Miyake et al. (2000) found moderate correlations among performance on each aspect, they concluded that they are actually clearly separable.

Stanovich (2009a) made the point that executive functions are not actually ‘executive’ at all. ‘Executive’ implies the highest level of processing, which according to the tripartite theory is the reflective level. However, executive function tasks actually constrain reflective level processing because the tasks are always fairly simple and come with instructions, and the measure instead is how efficiently one can perform given that they know exactly what to do. A truly executive-tapping task would require the participant to work out how to solve the task themselves and then be able to put in a self-determined level of effort to complete it. As stated above then, executive function tasks measure efficiency of algorithmic level functioning.

The name issue aside, the three main executive functions alongside general intelligence provide some insight into the algorithmic mind and are therefore a useful place to look for differences between mathematicians’ and non-mathematicians’ reasoning behaviour. Moutier, Angeard and Houdé (2002), for example, found some evidence that matching bias may be due to a failure of inhibition rather than a problem with logic. This demonstrates how executive functions may indeed be a promising line of investigation when searching for mechanisms of better reasoning. If studying mathematics improves one’s algorithmic functioning, we may find that reflected in one or more of: higher general intelligence, higher working memory capacity, stronger inhibition, or more efficient shifting ability.

Finally, it may be the case that the TFD operates via the reflective level, which is tapped via measures of thinking disposition. Stanovich (1999) described thinking dispositions as “relatively stable psychological mechanisms and strategies that tend to generate characteristic behavioural tendencies and tactics” (p. 157). They may also be referred to as intellectual styles, cognitive styles, or habits of mind. As discussed before, the CRT provides a behavioural measure of the tendency for effortful thinking. Other measures that assess thinking dis-

positions include the Need for Cognition (NFC) scale (Cacioppo, Petty & Kao, 1984) and the thinking dispositions questionnaire devised by Stanovich and West (1998) that includes the actively open-minded thinking subscale, the counterfactual thinking subscale, the absolutism subscale (adapted from Erwin, 1981), the dogmatism subscale (adapted from Trolldhal & Powell, 1965 and Rokeach, 1960), and finally the paranormal beliefs subscale (adapted from Jones, Russell & Nickel, 1977 and Tobacyk & Milford, 1983).

A person's thinking disposition may influence how long they persevere at a difficult task, whether they seek or avoid effortful thinking or how open or closed their thinking tends to be. As Stanovich's (2009a) tripartite theory proposes, thinking disposition may be just as important as cognitive ability, or intelligence, because it determines the extent to which a person's intellect will actually be used.

The NFC scale has 18 self-report items (Figure 2.14) that provide a measure of enjoyment of effortful thinking. It is considerably shorter than Stanovich and West's (1998) thinking dispositions questionnaire, but is nonetheless related to grade point average in undergraduates (Elias & Loomis, 2002) and complex problem solving behaviour (Nair & Ramnarayan, 2000). The CRT is a behavioural as opposed to self-report measure, and so the two measures used side-by-side may provide a more broad reflection of thinking disposition, making them suitable measures for a study of reasoning skills in students. If it is the case that studying mathematics improves reasoning skills by influencing an individual's reflective mind, then we may find that their Need for Cognition or CRT score increases as well.

Section summary

- Psychological theories of reasoning tend to describe one, two or three process-types.
- Gillard's (2009) work supported the default-interventionist model of reasoning. According to the tripartite model of reasoning, which falls into the default-interventionist category, there are several possible mechanisms via which the TFD may operate.
- The TFD may operate via Type 1 processes, presumably though changing the heuristics that sometimes cause errors in reasoning.
- It may also operate via the algorithmic level of Type 2 processes, i.e. learning new logical rules, increased intelligence or improved executive functions resulting in more effective or more efficient processing.

- Finally, the TFD may operate via the reflective level of Type 2 processes, i.e. a change in thinking disposition that means individuals are more willing to engage in effortful thinking and so use algorithmic level processing more often or more effortfully.
- These possibilities are not mutually exclusive. If the TFD is correct, it may operate through any combination of these mechanisms.

- 1 I would prefer complex to simple problems.
- 2 I like to have the responsibility of handling a situation that requires a lot of thinking.
- 3 Thinking is not my idea of fun.*
- 4 I would rather do something that requires little thought than something that is sure to challenge my thinking abilities.*
- 5 I try to anticipate and avoid situations where there is likely a chance I will have to think in depth about something.*
- 6 I find satisfaction in deliberating hard and for long hours.
- 7 I only think as hard as I have to.*
- 8 I prefer to think about small, daily projects to long-term ones.*
- 9 I like tasks that require little thought once I've learned them.*
- 10 The idea of relying on thought to make my way to the top appeals to me.
- 11 I really enjoy a task that involves coming up with new solutions to problems.
- 12 Learning new ways to think doesn't excite me very much.*
- 13 I prefer my life to be filled with puzzles that I must solve.
- 14 The notion of thinking abstractly is appealing to me.
- 15 I would prefer a task that is intellectual, difficult, and important to one that is somewhat important but does not require much thought.
- 16 I feel relief rather than satisfaction after completing a task that required a lot of mental effort.*
- 17 It's enough for me that something gets the job done; I don't care how or why it works.*
- 18 I usually end up deliberating about issues even when they do not affect me personally.

* Reverse scored

Figure 2.14: The Need for Cognition Scale.

2.7 Current status of the Theory of Formal Discipline

Despite the many changes in psychological opinion and the conflicting evidence about whether thinking skills can be generalised across contexts or not, the TFD view is still held by many academics and policy makers today. Oakley (1949) argued that “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) and Amitsur similarly stated that “Through mathematics we also wish to teach logical thinking – no better tool for that has been found so far” (Sfard, 1998, p. 453).

The TFD has also motivated arguments for mathematics to receive a special status in the UK National Curriculum. In his report on secondary mathematics education in the UK, Professor Adrian Smith stated that as well as being important for its own sake, mathematics is also important because it “disciplines the mind, develops logical and critical reasoning, and develops analytical and problem-solving skills to a high degree” (Smith, 2004, p. 11). Based on this, the report recommended that high priority be given to “encouraging and funding a significant increase in the number of mathematics graduates admitted to the Fast Track Scheme” (Smith, 2004, p. 45), a programme which gives new teachers an additional spine point on the pay scale and a £5,000 bursary.

This was one of many recommendations intended to raise the status of mathematics and increase the numbers of mathematics students and teachers. More recently, the Walport report (2010) identified quantitative and logical thinking as science, technology, engineering and mathematics (STEM) skills, and quoted the Advisory Committee on Mathematics Education who stated that “problem solving abilities, perseverance and logic are [...] highly sought after and are commonly found in those with a high level of competency in mathematics” (p.185).

The arguments cited above are evidence that the TFD belief is still held, both in academic writings and in policy debates. It has even been found that people with an A level in mathematics earn 7-10% more than those with similar ability and qualifications at the age of 33 (Dolton & Vignoles, 2002). Mathematics qualifications are clearly highly valued and it is possible that this is due, at least in part, to the belief that studying mathematics improves one’s reasoning ability.

The TFD is not supported by all, though. The recent book ‘Why Learn Maths?’ (Bramall & White, 2000) challenged the TFD in several chapters. Bramall argued explicitly that mathematics does not deserve to be prioritised in the curriculum and that subjects should be more equally balanced (Bramall,

2000). Peter Huckstep reviewed several theories of how mathematics trains the mind and concluded that the most defensible version was that of Thomas Tate, which applies only to the elementary level – not to the advanced level where recent policy debates have focused (Huckstep, 2000). Tate suggested that mathematics provides a useful introduction to reasoning but that it only trains students in how to deal with mathematical knowledge – he did not believe that mathematics provides reasoning skills for other contexts (Huckstep, 2000).

Despite the fact that the TFD debate is still unresolved, UK education policies are being influenced in a way that assumes the TFD is correct. At best, the lack of conclusive evidence means that the TFD may not be wielding as much clout as it could, and at worst, it leaves open the possibility that it is just not true. The aim of this thesis, therefore, is to provide a direct test of the TFD by investigating whether, and if so how, studying mathematics at A level and undergraduate level improves reasoning skills.

Section summary

- Belief in the Theory of Formal Discipline has been held since the time of Plato and is still held today, despite some disputes from recent psychologists.
- The Theory of Formal Discipline is used in modern day policy reports to support the prioritisation of mathematics in the curriculum.
- This thesis aims to provide a direct test of the TFD which may be able to inform debates about the place of mathematics in the UK's national curriculum.

Chapter 3

Methodology

3.1 Introduction

The aim of this chapter is to outline the necessary considerations for conducting high quality ethical research. The areas covered will be research ethics, experimental and quasi-experimental designs, reliability, and validity. Potential problems in each of these areas will be identified and addressed in order to lay a solid base for the research presented later in the thesis.

3.2 Overview of longitudinal study

In order to put some of the methodological considerations discussed below into context, it would be helpful to include a brief overview of the main studies presented later in the thesis. The main studies, presented in Chapters 5 and 6, were designed to answer the question of whether studying mathematics at advanced levels is associated with improvement in logical reasoning skills. The studies followed longitudinal quasi-experimental designs. In Chapter 5 students studying AS level mathematics were compared to students studying AS level English Literature for improvement in logical reasoning skills over their year of study and in Chapter 6, mathematics and psychology students were compared for development in reasoning skills over a year. This type of design raised several issues, which are discussed in the sections below.

3.3 Research Ethics

The first thing that may spring to mind when the subject of research ethics comes up is the need to look out for participants' welfare – we cannot subject

them to just any conditions for the sake of answering our research question and this is discussed below. First, the issue of the relationship between science and society is considered (Diener & Crandall, 1978) due to its relevance to research on the TFD.

3.3.1 Science and society

The issue of the relationship between science and society concerns the extent to which society should influence which research topics are pursued (Christensen, 2000). Competition for research funding can be intense, and the money tends to go to the projects that are ‘in vogue’. For example, our society is currently concerned about poverty in third world countries, the economic climate of our own country, obesity, and immigration (amongst other things). Correspondingly, the Economic and Social Research Council (ESRC), one of the primary research funding bodies in the UK, states in its priorities for the 2009-2014 period ‘reducing poverty among the poorest countries in the world’ (p. 6), ‘understanding of individual and household responses to the rapidly changing economic climate’ (p. 6), ‘the reduction of obesity’ (p. 8), and understanding ‘the dynamics of migration into and from the UK’ (p. 21, ESRC, 2009).

What this subsequently means is that researchers may alter their research focus for the sake of obtaining funding. While it is not necessarily wrong for researchers to investigate what is important to society, especially when research funding is coming from taxpayers’ money, it does compromise the objectivity of science. Traditionally, science is supposed to be the uncovering and explaining of the nature of our world by objective scientists but this may well be compromised when researchers have to compete for funding by tailoring their research questions to social or political motives. As Christensen (2000) points out, researchers’ interests may also be determined by personal experiences. For example, a dyslexic researcher may want to investigate dyslexia in school children, and again, this may compromise the objectivity of the research.

The issue of society influencing science is relevant to the research reported in this thesis. As discussed in Section 2.7 on the current status of the TFD, there have been many recent arguments for mathematics to be prioritised in the National Curriculum based on the assumption that studying mathematics improves thinking skills. Investigating the accuracy of that assumption is a part of the motivation for conducting this research. It is important, therefore, that the research is conducted as objectively as possible – not biasing the results towards either supporting or refuting the TFD. One way in which this issue is dealt with is by withholding the purpose of the study from the participants (discussed below). Another benefit to objectivity is that I am approaching the

question with a background in psychology, not mathematics, and so do not feel the need to support mathematics by finding the TFD to be true, whereas others with a background in mathematics might.

3.3.2 Welfare of participants

Participants in research studies should always be asked for informed consent, given the right to withdraw, and debriefed after participation. Wherever possible, research should be conducted in a way in which participants do not suffer any negative effects, such as failure, stress or embarrassment, and they should not be deceived (Ethics Committee of the British Psychological Society, 2009). Two of these issues which have particular importance for the research presented in this thesis are informed consent and deception.

Consent

Informed consent means that participants should be told all relevant information about the nature and purpose of the research and that they should only subsequently take part having given their voluntary consent. When the participants are vulnerable adults or children special considerations are required for informed consent. The participants in all studies reported in this thesis, including in the AS level study, were aged 16 or over. According to the BPS (The British Psychological Society, 2010), only those under the age of 16 are considered to be children. Nevertheless, parental consent was obtained for the participants aged under 18 in Chapter 5. The participants in all of the studies reported here were given information about the purpose and nature of the studies they were asked to take part in (except for some minor withholding of information in the case of the AS level study, see below), and all gave consent before their participation began.

Deception and withholding information

In the longitudinal studies reported in this thesis, it was necessary to withhold some of the details of the purpose of the study from participants. If they were aware that I was testing the hypothesis that mathematics students would improve in reasoning to a greater extent than the English/psychology students, the results may have been invalidated by stereotype threat.

Stereotype threat occurs in test situations when a person could potentially reinforce a negative stereotype about a social group they belong to. For example, when men and women were given the same mathematics test and told either that it did or did not tend to show a gender difference in scores, Spencer, Steele and Quinn (1999) found that gender differences in performance followed accordingly:

for the group of participants who were told the test did not produce a gender difference, no gender difference was found, but for the group that were told that it did produce a gender difference, women scored significantly lower than men. In another study, Asian-American women were found to perform better on a mathematics test when their ethnic identity was primed, and worse when their gender was primed, than a control group who had neither identity primed (Shih, Pittinsky & Ambady, 1999). White men’s mathematics performance was also found to suffer when they were reminded that Asian people tend to outperform white people on mathematics tests, compared to a control group that was not reminded of the stereotype (Aronson, Lustina, Good & Keough, 1999). These examples demonstrate what a powerful effect stereotype threat is – a single statement from researchers can create a significant difference in the performance of participant groups.

In the longitudinal studies reported here, stereotype threat could be a problem if participants were told that I was testing the claim that mathematics students would improve in reasoning to a greater extent than the English/psychology students – it could be enough to create the difference between groups that I was testing for. Instead, participants were told that the research was looking at improvement in reasoning over A-levels without emphasising any comparisons between subjects. The participants were not deceived, therefore, rather information was withheld. In this case, the cost to participants was minimal, and the findings would simply not be valid if the information were not withheld. In line with the BPS code of conduct (Ethics Committee of the British Psychological Society, 2009), it was deemed that this withholding of information was appropriate, and the study was approved by Loughborough University’s Ethics Approvals (Human Participants) Sub-Committee.

3.4 The Experimental Method

Most of the studies reported in this thesis follow experimental and quasi-experimental designs, and the current and following sections will elaborate on each of these methods respectively.

The experimental method is used to test hypothesised causal relationships by systematically manipulating one or more variables and measuring the effect on other variables. This can occur in a highly controlled laboratory setting or in a less controlled field setting. The variables that are manipulated are referred to as independent variables, and those that are measured are dependent variables. In an experiment, the independent variable is split into two or more conditions, which could be something like a drug condition and a placebo condition, or stimuli display time conditions of 250ms, 500ms, and 750ms, for example. The

Order	Complete						Latin Square		
	1	2	3	4	5	6	1	2	3
Task 1	A	A	B	B	C	C	A	C	B
Task 2	B	C	A	C	A	B	B	A	C
Task 3	C	B	C	A	B	A	C	B	A

Table 3.1: Demonstration of complete and Latin Square counterbalancing for a set of three tasks. A = administered first, B = administered second, and C = administered third.

experimenter then measures and compares the dependent variable under each condition, e.g. patients' reports of their symptoms following treatment.

One way in which experiments can differ is in whether the conditions are administered between- or within-participants. In a between-participants design, the participants are randomly assigned to different conditions. Random assignment is very important because it means that any differences between groups, other than the experimental manipulation, are due to random variation rather than systematic variation. Any differences between groups found in the dependent measure can therefore be said to be due to the manipulation.

In a within-participants design, all participants experience all conditions, and the order in which they are administered must be counterbalanced. Complete counterbalancing means that all possible orders of tasks are administered (with participants randomly assigned to one order each). Latin square counterbalancing means that tasks are presented in a set but rotating order. Figure 3.1 demonstrates complete and Latin square counterbalancing for a set of three tasks. The reason for counterbalancing is to prevent any order effects. For example, if participants become tired towards the end of a study, counterbalancing ensures that this does not affect performance on one task alone but balances the effect between all tasks.

Christensen (2000) stated that there are three conditions that must be met for a good experimental research design. The first criterion is that the design must allow the research question to be answered. The second criterion is that extraneous variables (variables that are not of interest but that affect the dependent variable) are controlled for (also known as internal validity, see Section 3.6.2). The third is that the findings are generalisable.

The first criterion, that the design must allow the research question to be answered, seems so fundamental as to not require stating. However, it is not impossible for a researcher to get as far as trying to interpret data before realising that this criterion has not been met. For example, take the research question

‘Is intervention X effective in helping dyslexic children improve their reading speed?’. A flawed approach to answering this question would be to take a sample of dyslexic children and measure their reading speed before and after administering intervention X. Suppose the findings showed that the children’s reading speed was faster after the intervention. Does this answer the research question? No, it merely shows that the children’s reading speed became faster over time, whether or not that has anything to do with the intervention is impossible to say in the absence of a control group. Recall the study by Lehmann (1963) which showed that undergraduate students’ critical thinking skills improved throughout their university education. It was mentioned in the literature review in Chapter 2 that this could have been a change that occurs in all college-aged people. Due to the lack of a control group, it was not reasonable to assume that the change had any relation to the participants’ educational experiences. To adequately answer such questions, the design needs an experimental group that receives the intervention and a control group that does not, and the participants need to be randomly assigned to one group or the other to allow any differences to be attributable to the manipulation.

The second criterion, that extraneous variables are controlled for, is necessary to be able to eliminate rival hypotheses. An extraneous variable is something other than the independent variable that influences the dependent variable. It would be no good if you concluded that participants who read a happy story recalled more details than those who read a sad story if the participants in the happy story group were more intelligent, for example. The best way to control for extraneous variables is to include a control group and to randomly assign participants to groups. The control condition should be identical to the experimental condition in all aspects except that it does not receive the experimental manipulation. In this way, the independent variable is isolated as the difference between conditions and a research question about that variable can be answered. By randomly assigning participants to conditions, there should be no known or unknown variables that affect one group more than another, except for the independent variable.

The third and final criterion is generalisability (also known as external validity, see Section 3.6.2) – the extent to which the results can be applied beyond the study itself. Generalisability is restricted by having a non-representative sample or an artificial experimental situation. For example, if your sample is entirely made up of females the findings may not generalise to males. Similarly, if you study factors affecting attractiveness of faces using photographs in a lab setting, your results may not apply to attractiveness of faces as seen in natural public settings. It is likely that participant samples will always be restricted in some dimension, but the important thing is to be aware of the boundaries of

the population to which your sample belongs and thus how far your results can generalise when you draw your conclusions. Artificiality of study environments, however, is a more interesting issue.

A complaint often levelled at lab based and experimental research is that the artificiality of the setting makes the results non-generalisable to other settings (Mook, 1983). This complaint often reflects a fundamental misunderstanding of the aims of the research. On the whole, such research is not conducted with the aim of generalising the results to other situations, it is done with the aim of testing a theory. In psychological research, theories describe and explain real world behaviour and the role of experiments is to test and refine those theories. In testing a theory, a researcher derives a hypothesis that can be tested in a controlled situation. The results of the experiment are only used to accept or reject that hypothesis, thus informing the researcher about the accuracy of the theory from which it was derived. This means that it is irrelevant whether or not the experiment resembles real life, what is important is that the experiment is highly controlled so that the hypothesis is being tested accurately – in the absence of confounding variables. Mook (1983) provided a very detailed discussion of this issue, and it is also discussed further in Section 3.6.2 below on external validity.

3.5 The Quasi-Experimental Method

A quasi-experiment differs from a true experiment in that the independent variable is not randomly assigned. It may be that it is a pre-existing characteristic such as gender, nationality, number of siblings or degree subject, in which case the conditions cannot be assigned or manipulated by the experimenter. It may instead be the case that participants cannot be randomly assigned because of ethical reasons, for example, if you were looking at the effectiveness of an alcohol addiction treatment it might be unethical to deny treatment to an alcoholic by assigning them to a control condition (Christensen, 2000), although of course this regularly happens in medical research, presumably because the knowledge gained is deemed to justify the temporary withholding of treatment.

The design of the longitudinal studies in this thesis is quasi-experimental. The hypothesis is that studying mathematics improves students' reasoning skills to a greater extent than studying other subjects. Students who had already chosen to study mathematics were compared to students who had already chosen to study other subjects. It would be unethical, and practically impossible, to randomly assign people to studying different A levels or degrees and so a quasi-experimental design is the only way to test the hypothesis.

The problem with non-random assignment to conditions is that extraneous

variables, and in particular, *confounding* extraneous variables, are not properly controlled for. An extraneous variable is a factor other than the independent variable that affects the dependent variable. An extraneous variable becomes confounding when it also systematically varies with the independent variable. If a confounding variable is not controlled for by random assignment to conditions, it can become an alternative explanation for any effects found and this creates a problem for determining causation.

A way to deal with this problem in a quasi-experimental design is to measure and statistically control for any factors that are anticipated to be confounding. In the case of the longitudinal study, the mathematics group may have a higher mean intelligence than the non-maths group (Inglis & Simpson, 2009a), and intelligence may affect reasoning ability and development. In this case, the way to deal with the problem is to measure participants' intelligence and statistically control for its influence in the analysis.

In a true experiment with random assignment to conditions, causation can be established because no other variables could be creating the effects observed. In a quasi-experiment, it is only possible to establish plausible causation by ruling all alternative hypotheses implausible. Christensen (2000) gave the example of a person who dies immediately after being hit by a car. It is possible that they actually died from a heart attack, independently of being hit by the car, but that is quite implausible so you can reasonably conclude that the collision was the cause of death.

Beyond the problem of causation, there are several other issues associated with quasi-experimental designs. In each of the examples given below, the results of a study would seem to support the hypothesis that an intervention or treatment works. However, there are alternative explanations of the results that have not been ruled out.

In the longitudinal studies reported in Chapters 5 and 6 participants' reasoning skills are being measured both before and after the conditions are experienced. One potential outcome of this design identified by Christensen (2000) is that the comparison group does not change in reasoning skills and the mathematics group does, as demonstrated in Figure 3.1. This would imply that the hypothesis is correct, but there is a potential problem known as selection-maturation.

Selection-maturation means that one group may already be developing faster in the dependent variable than another group, for example, because they are more intelligent. Perhaps mathematics students are more intelligent than those of other subjects, and perhaps high intelligence individuals are on a faster developmental trajectory for reasoning skills than lower intelligence individuals.

There are several approaches to dealing with this potential issue. One is

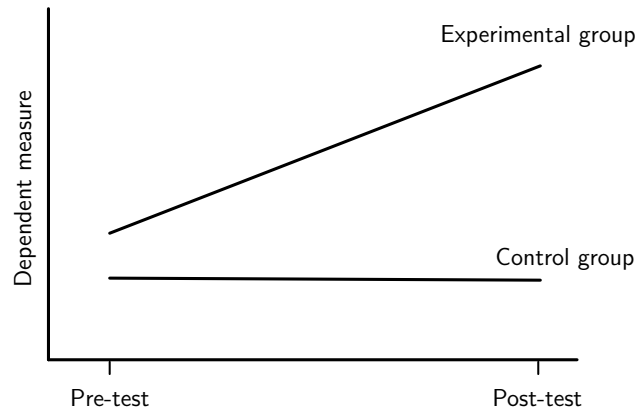


Figure 3.1: Possible quasi-experimental design outcome 1.

to match the groups on the extraneous variable that could be responsible for a selection-maturation effect. In this case, we would need to match the participants in each group on intelligence scores. This would mean taking participants from the lower range of the higher scoring group and from the upper range of the lower scoring group in such a way that the selected groups means are equal on the intelligence measure. There is a problem with this solution though: the participants that are at the extremes of their group range of intelligence scores at pre-test may regress towards the mean of their group by post-test, which could lead us to underestimate the effect of the independent variable.

Another possible solution is to use statistical methods such as analysis of covariance (ANCOVA) that take into account the effect of the confounding variable when determining the results. Van Breukelen (2006) discussed the advantages and disadvantages of using ANCOVA compared to repeated measures analysis of variance (ANOVA) for inferring a treatment effect. The ANCOVA method is to perform an analysis on Time 2 scores with Time 1 scores as a covariate along with the suspected confounding variables, while the repeated-measures ANOVA method involves comparing change-from-baseline in each group with only the confounding variables as covariates. Van Breukelen (2006) argued that in randomised studies both methods are unbiased but ANCOVAs have more power. However, where there is not random assignment to conditions, repeated-measures ANOVAs are less biased because ANCOVAs assume no baseline difference, which cannot be certain in non-randomised designs. Therefore, repeated measures ANOVA with covariates of intelligence and thinking disposition will be used in the longitudinal studies presented in this thesis. A major benefit of this approach as opposed to matching participants is that the sample is not

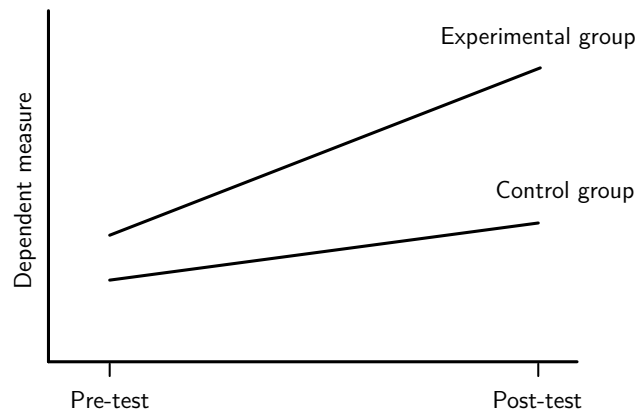


Figure 3.2: Possible quasi-experimental design outcome 2.

reduced.

Besides a selection-maturation effect, another potential threat with such a pattern of results is a local history effect. This is where some event affects one group but not the other. Some possible events that could create this problem in the longitudinal study would be if the mathematics group were also more likely to be taking a logic, critical thinking, or perhaps physics course. This is easy to identify by recording all of the subjects that the participants are studying.

Another possible outcome of the design that was identified by Christensen (2000), shown in Figure 3.2, is that both groups change on the dependent measure over time, but the experimental group changes more than the control group. Again, this could be the result of a selection-maturation effect where the experimental group are on a faster developmental trajectory than the control group.

A third possibility identified by Christensen (2000) is shown in Figure 3.3. In this case the experimental group scores lower than the control group at pre-test, and increases to nearer the level of the control group by post-test. This pattern of results is more likely to occur when the experimental group is a disadvantaged group and the treatment is an intervention designed to help them. For example, an intervention to help dyslexic students improve in their reading speed.

It is not likely that possibility 3 would occur in the case of reasoning ability in mathematics and non-mathematics students, but if it were to occur there would be a danger that the effect was due to a regression towards the mean by the unusually low-scoring experimental group. In this case it would be necessary to also track the deprived group's scores over time in the absence of any intervention. If the scores were consistent over time, it would help to support the conclusion that the improvement in the experimental group was in fact due to the treatment.

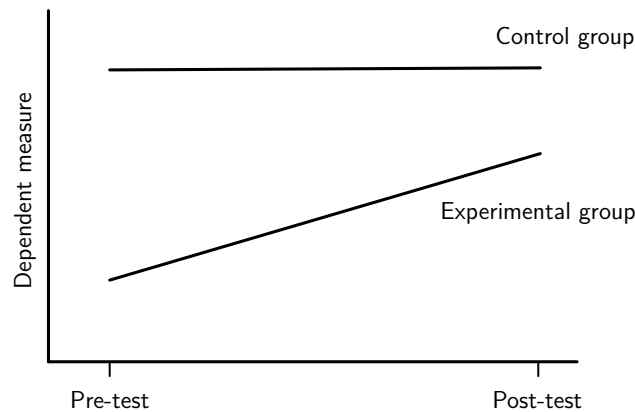


Figure 3.3: Possible quasi-experimental design outcome 3.

The fourth and final possibility identified by Christensen (2000) is that shown in Figure 3.4 where the experimental group's scores start below the control group's and finish higher, while the control group scores do not change. In this case the alternative hypotheses that threaten the other possible results are not an issue. Regression towards the mean is not a plausible possibility, and neither is a selection-maturation effect because it is usually the group that scores highest at pre-test that develops fastest.

To conclude, a quasi-experimental design is not ideal, but because participants cannot ethically or practically be assigned to studying different subjects at A level or degree level, it has to suffice. The main issue is that by not randomly assigning participants to conditions, confounding extraneous variables are not ruled out. This means that there may be alternative explanations for any effects found. Unless these alternatives can be statistically controlled for or deemed implausible, it is not possible to establish causation in a quasi-experimental design. In fact, even if all known confounding variables are ruled out, it is still not safe to conclude a causal relationship because there may be unknown confounding variables that are having an influence. Random assignment to conditions is really the only way to avoid any problems of this kind. This does not mean that the findings of quasi-experimental studies are not important or useful, but it means that they must be interpreted very carefully so as not to overstate any relationships found.

3.6 Reliability and Validity

This section will discuss the concepts of reliability and validity in research, specifically in relation to my own studies. In general, reliability refers to consistency

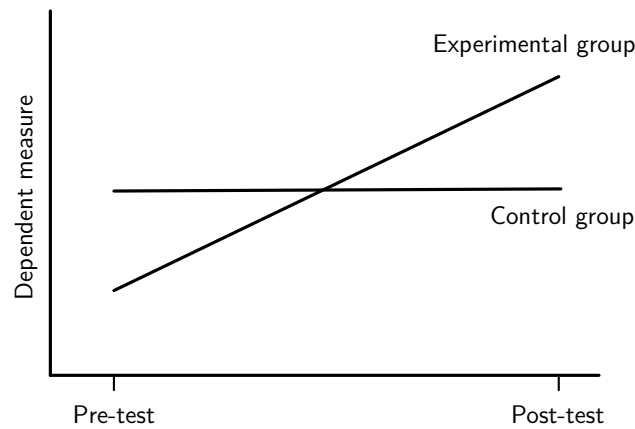


Figure 3.4: Possible quasi-experimental design outcome 4.

of measurement, and validity refers to the extent to which an instrument measures what you intend it to measure. However, these terms can be divided into many sub-types, as will be seen below.

3.6.1 Reliability

Beginning with test-retest reliability, the issue is whether the concept you are measuring is being measured consistently across time (Christensen, 2000). If, for example, you measured your participants' heights twice and got a different measurement each time, then unless they were on some sort of growth-spurt drugs you could conclude that the way in which you are measuring height is not very reliable. Unless the construct you are measuring is theorised to be unstable, then it is desirable that your measure provides the same results on different occasions.

How, then, can you ensure that your measures have high reliability? Changes in measurements can be the result of either systematic error (a changing factor of the situation that biases the results) or random error (Christensen, 2000). As the name suggests, random error happens for no particular reason and so cannot be controlled. The way in which to increase reliability, therefore, is to tightly control the experimental situation, ensuring that it is as similar as possible for all participants every time the study is run.

The main way in which reliability is maximised in the longitudinal study reported in this thesis is by ensuring that all participants complete the tasks in the same exam style conditions at every testing point. This includes working alone, in silence, and having the same amount of time to complete the tasks. This control should reduce the chance of any distractions or time pressure in-

terfering with performance. It would be a problem if, for example, the tasks were completed in silence at the first testing point and in a noisy environment at the second testing point – this could inhibit performance the second time round giving the misleading impression that the participants had become systematically worse at reasoning or it could hide any real improvements. As it is, any change that is found should be the product of an actual change in ability, not an error in measurement.

Internal reliability is the extent to which all items in a measure are related, i.e. consistently measuring the same construct (Heiman, 2002). If you have a measure made up of multiple items, it is a good idea that all items are measuring the same construct to some extent, although it is also desirable that each item brings something slightly unique as well. This ensures that the measure is useful, interpretable and not unnecessarily long. To take an example, suppose you have a 15 item task that is supposed to measure attitudes towards immigration, but three of the items actually measure attitudes towards emmigration. Attitudes towards the two things may be separate and unrelated so scores on the 12 immigration items may not be correlated with scores on the three emmigration items and this would give the task low internal reliability. Note how this is different from test-retest reliability – the immigration task may produce consistent results if re-administered every week, but the items within the task are not producing consistent responses.

Internal reliability can be measured with a split-half analysis or a Cronbach's alpha after the task has been completed by a number of participants (Novick & Lewis, 1967). In a split-half analysis, the items on a task are split into two groups, usually by alternating trials or at the midpoint. Participants' scores on the two halves of items are then subjected to a correlation analysis. In a Cronbach's alpha, a similar correlation is computed but averaging across every possible combination of test halves. The resulting correlation, or reliability coefficient, from either method is considered to be good when it is over .8 (Heiman, 2002). A reliability of below .7 would indicate questionable internal reliability, but these are just rules of thumb (George & Mallery, 2003). Split-half and Cronbach's alpha analyses are reported for the reasoning measures used in later chapters.

3.6.2 Validity

Validity is a slightly broader concept than reliability, encompassing external validity and internal validity. External validity is the extent to which a study generalises to other people and other situations, and it was touched upon in Section 3.4. It includes ecological validity – whether the study closely resembles

a real-life situation – and temporal validity – whether the findings would apply in the past and future as well as the present (Christensen, 2000). Reasoning is a cognitive process, and cognitive psychology assumes that all humans are born with the same cognitive processing systems (Miller, 2011; Neisser, 1967). This means that findings related to cognitive processing are assumed to apply across time and across the species. That said, this assumes that the manipulation is the same. If other people, at other times, studied the current UK A level mathematics syllabus, then we could hope for the results to be the same. This is not to say that the results can generalise to students studying different mathematics syllabi in non-UK education systems, or in the past or future.

In terms of ecological validity, one issue is whether the measures of reasoning ability used in my research are relevant to the TFD claims about the sort of reasoning that is valued by the job market. Are pen and paper based tasks valid measures of the types of reasoning skills that mathematics graduates might demonstrate in their future jobs? On the one hand, it is unlikely that the tasks closely resemble tasks that would be encountered in day-to-day life. On the other hand, it can be argued that an improvement in logical reasoning skills in general (as suggested by the TFD) would be demonstrated on any logical reasoning task because of the universal nature of cognitive processing (Miller, 2011). Since the TFD does not specify the exact types of logical reasoning skills that are improved by studying mathematics, this is the best we can hope for.

As Mook (1983) argued, however, ecological validity is often misunderstood and is not necessary if the research is designed to test a theory as opposed to generalise directly to the real world. In psychological research ecological validity is often compromised for the sake of internal validity (absence of confounding variables). This allows a hypothesis derived from a theory to be tested effectively. It is simply not possible to control for all confounding variables in a real world setting (and it is also difficult to find a real world setting that is the same as every possible real world setting to which you want to generalise, Mook, 1983). By accurately testing a hypothesis derived from a theory in a controlled artificial setting, it is possible to support, refute or refine that theory, and use the theory to explain real world behaviour. In this thesis, I am testing the TFD, and it is that theory, not my data, which generalises to the real world. The TFD argues that studying mathematics improves general reasoning skills. If this is true, then performance on an ‘artificial’ task that requires reasoning should be changed as part of the umbrella development.

As stated previously internal validity is basically the absence of confounding extraneous variables. High internal validity means that a relationship between two factors can be taken to be a relationship between those two factors alone (Heiman, 2002). If a study has low internal validity, there may be extraneous

variables that have not been controlled that are influencing the relationship (making them confounding variables). The best way to ensure that a study has high internal validity is to randomly assign participants to conditions so that any extraneous variables are balanced out and cannot systematically bias the results. Another safeguard is to measure any variables that you know may be confounding factors so that you can control for them in the analysis (Christensen, 2000).

As already stated, the quasi-experimental design of the longitudinal study means that participants are not randomly assigned to conditions. However, it is likely that intelligence is a confounding factor in the study – it may differ between conditions and it is related to reasoning ability. To deal with this issue, intelligence will be measured so that its effects can be accounted for in the analysis.

Internal validity may be compromised by the quasi-experimental design of the longitudinal study. However, despite the fact that external validity is arguably irrelevant because the study is testing the TFD rather than trying to generalise to real-world settings directly (Mook, 1983), it does have fairly high external validity due to the manipulation (mathematical study) being carried out ‘in the field’ rather than in a laboratory setting, and because it is dealing with a cognitive process that is believed to be common across humans (Miller, 2011).

Chapter 4

Measures of reasoning

4.1 Introduction

The psychology of reasoning has been a strong and growing area of research since the 1960s, when Peter Wason first demonstrated that people systematically fail to behave logically on his famous Selection Task (Wason, 1966). Since that time, a huge amount of research has been conducted with the Wason Selection Task and a collection of other tasks. The aim of this chapter is to justify why the Conditional Inference task was chosen as the primary measure of logical reasoning in this thesis, and why the Belief Bias Syllogisms task was chosen as a secondary measure. In order to do this, the most commonly used tasks in the field are described and discussed.

The chapter is split into two broad sections: judgment and decision making tasks and deductive reasoning tasks. The deductive reasoning section is further divided into three parts: disjunctive reasoning tasks, conditional reasoning tasks, and syllogisms tasks. After each of the tasks has been described, I will present an argument for why the Conditional Inference and Belief Bias Syllogisms tasks were chosen to measure reasoning in the studies presented in this thesis.

4.2 Judgment and Decision Making

In a commentary piece, Kahneman (1991) characterised the judgment and decision making field by three features: the role of the normative theory of rational belief and choice, the emphasis on risky choice, and the cognitive, rather than social or emotional, focus. Put simply, the judgment and decision making field aims to explain the cognitive basis of human thinking, and in particular, its

departure from rationality. This latter aspect is known as the heuristics and biases tradition and was pioneered by Kahneman and Tversky in the early 1970s (e.g. Tversky & Kahneman, 1971, 1974). This area of work has attempted to understand human rationality (or irrationality) by examining the biases we are prone to and their basis in heuristic processes.

4.2.1 Heuristics and Biases tasks

Law of large numbers problems

Tversky and Kahneman (1974) gave their participants the following problem:

A certain town is serviced by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50 percent of all babies born are boys. However, the exact percentage varies from day to day. Sometimes it may be higher than 50 percent, sometimes lower.

For a period of one year, each hospital recorded the days on which more than 60 percent of the babies born were boys. Which hospital do you think recorded more such days?

- The larger hospital
- The smaller hospital
- About the same (that is, within 5% of each other)

According to the statistical law of large numbers, the greater the sample size the more closely it represents the population. By this rule, the large hospital should have a birth rate of boys that is closer to 50% than the small hospital. Therefore, the small hospital will have more days where over 60% of the babies born are boys. In contrast to the law of large numbers, most participants (56%) thought that the hospitals were ‘about the same’ in terms of the number of days on which more than 60% of babies were boys. Equal numbers of participants (22% each) chose the smaller hospital or larger hospital. This and many other studies (e.g. Neilens, Handley & Newstead, 2009; Nisbett, Krantz, Jepson & Kunda, 1983; West et al., 2008) have shown that, on the whole, people do not invoke the law of large numbers when they should.

Base rate neglect

A base rate is the probability of an event occurring in the absence of any other information. For example, in a sample of 100 people where 99 are women, the

base rate of women in the sample is 99/100. Base rate neglect is another issue that was raised by Kahneman and Tversky (1972). The issue is nicely described with Tversky and Kahneman's (1982) taxi problem:

A taxi is involved in a hit and run accident at night. In the city, there are two taxi firms, the Green Cab Company and the Blue Cab Company. Of the taxis in the city, 85% are Green and the rest are Blue.

A witness identifies the offending cab as Blue. In tests under similar conditions to those on the night of the accident, this witness correctly identified each of the two colours 80% of the time, and was wrong 20% of the time.

What is the probability that the taxi involved in the accident was in fact blue?

According to Bayes's rule, given that the base rate of Blue cabs is .15 and that the witness said it was blue with .8 accuracy, the probability of the taxi actually being Blue is .41. Conversely, most of Tversky and Kahneman's (1982) participants rated the probability as .8, which is simply the accuracy of the witness. This suggests that participants do not take account of the base rate information at all.

The Linda problem

The famous Linda problem originated with Tversky and Kahneman's (1983) study, and demonstrated a bias towards thinking that a conjunction of two factors could be more probable than either factor alone. The problem goes like this:

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 anti-nuclear demonstrations.

1. Linda is a teacher in elementary school.
2. Linda works in a bookstore and takes Yoga classes.
3. Linda is active in the feminist movement.
4. Linda is a psychiatric social worker.
5. Linda is a member of the League of Women Voters.
6. Linda is a bank teller.

7. Linda is an insurance salesperson.
8. Linda is a bank teller and is active in the feminist movement.

Participants were asked to rank the eight statements associated with Linda in order of their probability. The interesting finding was that participants were inclined to rank the conjunction ‘Linda is a bank teller and is active in the feminist movement’ as more probable than the constituent part ‘Linda is a bank teller’, presumably because the description leads people into think that Linda *must* be a feminist, whether or not she is a bank teller. However, it is simply not possible for a conjunction of two characteristics to be more probable than either one alone.

Framing bias

Framing bias describes the finding that participants may give different responses to two questions which are essentially the same, but which are framed differently. Take this example from Tversky and Kahneman (1981):

1. You are a health service official making plans for dealing with a new disease that is about to break out. It is expected to kill 600 people. Your scientific advisors tell you about the consequences of two possible treatment programmes: Programme A will definitely save 200 lives, whereas Programme B will have a one-third (.33) chance of saving 600. Which programme will you approve?
2. Your colleague has a choice between Programme C, which will certainly result in 400 deaths, and Programme D, which has a two-thirds chance (.67) that 600 people will die. Which should she approve?

In the first scenario participants are more inclined to choose Programme A, whereas in the second scenario Programme D is the preferred choice, although of course A and C are equivalent and B and D are equivalent. This is thought to reflect aversity to risk when it is framed in terms of positive outcomes, but preference for risk when it is framed in terms of negative outcomes. Nevertheless, this pattern of responses demonstrates a departure from logic.

Sunk cost fallacy

The sunk cost fallacy refers to the tendency to allow previous sunk costs (past costs that cannot be recovered) to affect current decision making. For example, say you bought a non-refundable ticket to a concert, but on the day you felt

very ill and did not want to go. If you thought you should go anyway because you had paid for the ticket then you would be committing the sunk cost fallacy.

Toplak, West and Stanovich (2011) gave their participants the film problem from Frisch (1993). First, participants are told to imagine that they are staying in a hotel room, and they have just paid \$6.95 to see a film on TV. Then they are told that 5 minutes in the film seems pretty bad and they are bored. They are asked whether they would continue to watch the film or switch to another channel. Second, participants see the same scenario except that they have not had to pay for the movie. They are again asked whether they would continue to watch the movie or switch to another channel. If participants report that they would change the channel when the film was free but that they would keep watching when they had paid for it, they were presumed to have committed the sunk cost fallacy. This was the case for 35.8% of participants in Toplak et al.'s (2011) study.

Outcome bias

Outcome bias is the tendency to rate the quality of a decision based on the outcome rather than on the situation at the time the decision was made. A problem that is often used to measure outcome bias derives from Baron and Hershey (1988) and has been used in many studies by Stanovich (Stanovich & West, 1998, Stanovich & West, 2000, Stanovich & West, 2008, Toplak et al., 2011). Participants are told about a 55-year-old man who had a heart condition and who was given an operation with an 8% mortality rate. The surgery was successful and participants rated the quality of the decision on a seven point scale. Later participants are told about a patient with a hip condition who was given an operation with a 2% mortality rate. Despite the decision to operate being objectively better, the patient died during the operation. If a participant rates the first decision (with a positive outcome) as better than the second (with a negative outcome), they have displayed outcome bias.

Gambler's fallacy

The Gambler's fallacy refers to people's misunderstanding of chance. Often, people incorrectly believe that what has happened in the past can affect the probability of future events. Toplak et al. (2011) gave their participants two problems designed to tap into the gambler's fallacy. The first problem went as follows:

When playing slot machines, people win something about 1 in every 10 times. Julie, however, has just won on her first three plays. What

are her chances of winning the next time she plays?
_____ out of _____

The correct response to this problem is 1 out of 10, the odds given in the question. The fact that Julie has already won three times has no bearing on the probability that she will win on any subsequent tries. Toplak et al. (2011) found only 69.4% correct responses to this problem in their study with undergraduate and graduate students. The second problem they gave participants was as follows:

Imagine that we are tossing a fair coin (a coin that has a 50/50 chance of coming up heads or tails) and it has just come up heads 5 times in a row. For the 6th toss do you think that:

1. It is more likely that tails will come up than heads.
2. It is more likely that heads will come up than tails.
3. Heads and tails are equally probable on the sixth toss.

The correct answer is 3, because again the past events are irrelevant to future probabilities, and in this case 92.2% of participants in Toplak et al.'s (2011) study answered correctly, suggesting that the bias is shown inconsistently across tasks.

Summary

This section has presented some of the common tasks used to measure pervasive biases in human judgement and decision making. The problems tend to resemble real world scenarios and each measures a small aspect of human reasoning (usually on a binary scale) which may be important in a limited range of scenarios but which are not necessarily more widely relevant.

The next section discusses deductive reasoning, which may be measured with problems resembling the real world, but which often is not. Deductive reasoning tasks require a necessary conclusion to be derived from given premises. Necessity means that the conclusion must be true when the premises are true. As such, deductive reasoning is about assessing logical validity when all the necessary information is available, rather than about making decisions in the face of limited information.

4.3 Deductive Reasoning

Deductive reasoning is the process of drawing necessary conclusions from given premises. It is based on absolute certainty, even though the premises may be

assumed rather than known. For example, the premises ‘All archbishops are believers’ and ‘No believers are Cannibals’ can lead to the necessary conclusion that ‘No archbishops are cannibals’. The premises may or may not be true, but when we assume they are true the conclusion becomes necessarily true. This certainty is what defines a deduction as valid or invalid: a deduction is valid if the conclusion must be true when the premises are true, and is invalid otherwise. In this section three types of deductive reasoning are discussed: disjunctive, conditional and syllogistic reasoning.

4.3.1 Disjunction tasks

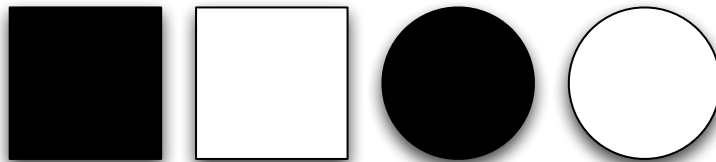
The THOG task

The most famous disjunction problem, the THOG task, was created by Wason and is shown in Figure 4.1 (Wason & Brooks, 1979). Readers unfamiliar with the THOG problem should read it now before moving on.

The answer to the THOG problem is that the white square and black circle cannot be THOGs while the white circle must be a THOG, but only 35% of

In front of you are four designs:

Black Square, White Square, Black Circle, White Circle



You are to assume that I have written down one of the colours (black or white) and one of the shapes (square or circle). Now read the following rule carefully.

If, and only if, any of the designs includes either the colour I have written down, or the shape I have written down, but not both, then it is called a THOG.

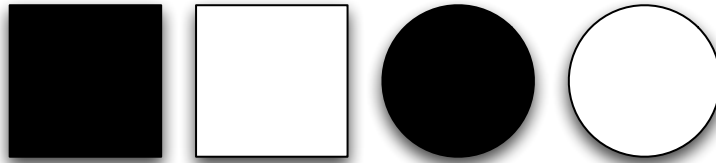
I will tell you that the Black Square is a THOG.

Each of the designs can now be classified into one of the following categories:

- A. Definitely is a THOG
- B. Insufficient information to decide
- C. Definitely is not a THOG

Figure 4.1: Wason’s abstract THOG problem.

I have brought a deck of cards. It contains only these four types of card:
Black Square, White Square, Black Circle, White Circle



I deal one for myself from the deck, and I won't show it to you. Now I'll deal you each a card, and I will pay for a dinner for each person who has a card including either the colour of my card, or the shape of my card, but not both.

(The four cards above are given to Rob, Tim, Paul and John, respectively.)

Without showing you my card, I can tell you that I owe Rob a dinner. Which card do you think I could have? And do you think that I have to pay for a dinner for someone else? If so, for whom?

Figure 4.2: The Pub problem, a contextualised version of the THOG task.

participants in Wason and Brooks's (1979) study gave this response. The problem states that the black square is a THOG, which means that the experimenter must be thinking of a white square or a black circle (a THOG shares one characteristic with the design the experimenter is thinking of). If the experimenter is thinking of a white square then the black circle (shares neither characteristic) and white square (shares both) can be ruled out as THOGs. If the experimenter is thinking of the black circle then the black circle (shares both) and white square (shares neither) can be ruled out as THOGs. Under both alternatives, a white circle shares one characteristic, and is therefore a THOG.

Giroto and Legrenzi (1989) created the pub problem, a reformulation of the THOG problem using realistic content. The problem is about a character called Charles who plays a game with four friends in a pub. The problem stated by Charles appears in Figure 4.2.

This problem is analogous to the abstract problem (in this case the answer is that John is also owed a dinner), yet 89% of people answered correctly according to Giroto and Legrenzi (1989). Furthermore, when Giroto and Legrenzi (1993) simply gave the name SARS to the hypothesised shape in the abstract version of the task, so that a THOG has one feature in common with a SARS, they observed 70% correct performance. The explanation given for the difficulty of the original THOG problem is called confusion theory and argues that people

p	q	$p \text{ or } q$
T	T	F
T	F	T
F	T	T
F	F	F

Table 4.1: Truth Table for the exclusive disjunction ‘ p or q ’ where T = true and F = false.

simply treat the exemplar THOG as if it was the design chosen by the experimenter. They then look for other designs that have one feature in common with the exemplar (Newstead, Girotto & Legrenzi, 1995). It is suggested that when people have to keep several hypotheses in mind at once, as with the exclusive disjunction in the THOG problem, they experience a cognitive overload and resort to more intuitive strategies. In this case, the intuitive strategy is to match the values of the exemplar with the test cases.

Truth Table tasks

Disjunctive reasoning can also be measured with a Truth Table task. A Truth Table is used in logic to demonstrate how the truth or falsity of each variable determines the validity or invalidity of a proposition about those variables. For example, Table 4.1 presents a Truth Table for the exclusive disjunction rule ‘ p or q ’. The fact that the disjunction is exclusive means that either p or q must be true, but not both. Each line represents a different combination of truth and falsity of the values p and q , and the final column denotes whether that combination makes the disjunctive rule true (valid) or false (invalid). In an inclusive disjunction, either p or q must be true, but both can be true as well. The Truth Table for an inclusive disjunction is shown in Table 4.2.

Truth Table tasks given to participants to measure their conceptions of

p	q	$p \text{ or } q$
T	T	T
T	F	T
F	T	T
F	F	F

Table 4.2: Truth Table for the inclusive disjunction ‘ p or q ’ where T = true and F = false.

<i>I order wine</i>	<i>I order water</i>	I'll order wine or water
T	T	T/F?
T	F	T/F?
F	T	T/F?
F	F	T/F?

Table 4.3: Truth Table for the disjunction rule ‘I’ll order wine or water’, where T = true and F = false.

the conditional include the truth and falsity of the variables but leave the rule column blank for the participant to complete, i.e. the participant decides whether each combination of variables makes the rule true or false. This can be given in thematic as well as abstract form, as demonstrated in Table 4.3 for the disjunction ‘I’ll order wine or water’. By asking participants to complete the final column we can infer how logical people are in their assessment of disjunctions and whether they prefer an exclusive or inclusive interpretation of the disjunction.

Evans (1993) reviewed a set of studies that used abstract disjunctive Truth Table tasks. He found that the not- p not- q case was always rated false, as it should be under both exclusive and inclusive readings, but the p not- q and not- p q cases were rated true about 80% of the time, despite both being true under both readings. This suggests that people do not reason entirely logically with disjunctions. As for a preference for exclusive or inclusive readings, the findings were inconsistent. In some studies there was a clear preference for an exclusive reading (where the p q case is rated false), in some there was a clear preference for an inclusive reading (where the p q case is rated true) and in others there was no clear preference (Evans, 1993).

Disjunctive Inference task

In a Disjunctive Inference task, participants are given a disjunctive rule along with a premise about that rule, followed by a conclusion derived from the rule and premise. The participant then assesses whether the conclusion is valid or invalid. For example:

Rule: Either A or B
Premise: not B
Conclusion: A

There are two denial inferences and two affirmation inferences. The denial

p	q	Material Conditional	Biconditional
t	t	t	t
t	f	f	f
f	t	t	f
f	f	t	t

Table 4.4: Truth Table for ‘if p then q ’ where t = true and f = false.

inferences are ‘Either p or q ; not p ; q ’ and ‘Either p or q ; not q ; p ’. Both of these inferences are valid under both the exclusive and inclusive readings. The affirmation inferences are ‘Either p or q ; p ; not q ’ and ‘Either p or q ; q ; not p ’. Under an exclusive reading, both affirmation inferences are valid. Under an inclusive reading, the conclusions may or may not be true, so the inferences are invalid (not *necessarily* true).

As with the Truth Table task, the Disjunctive Inference task can be given with thematic as well as abstract content. For example, ‘My sister keeps tropical fish, which are either angels or neons; they’re not angels; therefore they’re neons’. As with the Truth Table task, this type of task can be used to assess the extent to which participants conform to normative logic and whether they prefer an exclusive or inclusive reading.

4.3.2 Conditional tasks

Conditional reasoning is the process of drawing conclusions from rules based on ‘if’. As Manktelow (1999) argued, the word ‘if’ has probably sparked more interest from psychologists, philosophers and logicians than any other word in the English Language.

Conditional rules come in many forms, including ‘ p only if q ’, ‘ q if p ’, ‘ p if and only if q ’ and the most commonly used form, ‘if p then q ’. According to formal propositional logic, each of these statements should be treated as the material conditional except for ‘ p if and only if q ’, which is biconditional. Under a material conditional reading, the rule is only proved false when p is true and q is false. Under a biconditional reading, p implies q but q also implies p , so the rule is false when p is false and q is true but also when p is true and q is false. The material conditional and biconditional readings are represented in Figure 4.3 as Euler diagrams and in Figure 4.4 as Truth Tables.

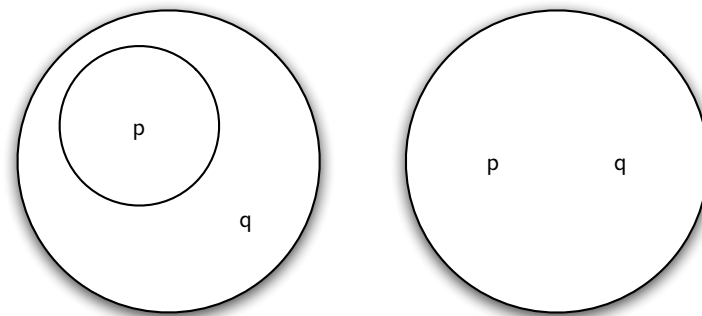


Figure 4.3: Euler diagrams to represent ‘if p then q ’ under the a) material conditional interpretation, where p implies q , and b) biconditional interpretation, where p implies q and q implies p .

Wason Selection Task

The most famous task in the psychology of reasoning is the Wason Selection Task (WST), which was designed to measure conditional reasoning. It was developed in the 1960s and has spawned a great deal of research since, so much so, that one journal stopped publishing any research that used it (Manktelow, 1999). Figure 4.4 displays the task. Participants are shown four cards that each have a letter on one side and a number on the other side. Two cards are letter side up, say A and D, and two are number side up, say 3 and 7. Participants are asked to choose those cards, but only those cards, that they would need to turn over in order to tell whether a rule such as ‘if there is an A on one side of the card then there is a 3 on the other side’ is true or false. In this case, the A and 7 cards would need to be turned over. If the A card had a not-3 number on the other side, or if the 7 had an A on the other side, then the rule would be falsified. However, hundreds of participants across many studies have failed to make this choice. Most chose the A card and sometimes the 3 card too (Wason & Johnson-Laird, 1972; Evans, 1993).

Matching bias (Evans, 1998) is commonly observed on the Wason Selection Task (Wason, 1968). The typical response, A or A and 3, still tends to be given even when the rule is changed from ‘if A then 3’ to ‘if A then not 3’, despite the ‘A, 3’ response then becoming logically correct. This result has been interpreted as indicating that participants simply match the cards to the rule rather than using any systematic reasoning strategy. Wason and Evans (1975) demonstrated that participants showed no awareness of their bias in verbal reports, suggesting that it is an unconscious attentional bias.

The WST has been investigated with thematic content as well as abstract content and in many cases, this has been found to improve performance. Wason

Each of the cards below has a letter on one side and a number on the other side.
Select all those cards, but only those cards, which would have to be turned over
in order to discover whether the rule is true.

Rule: If there is an A on one side of the card, there is a 3 on the other side.

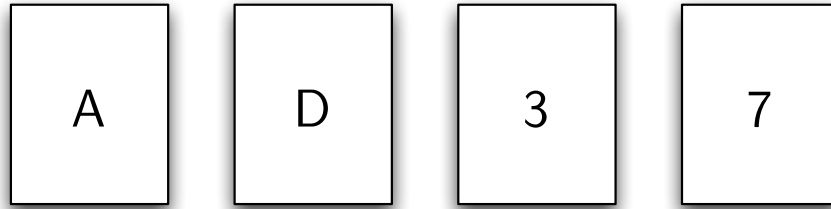


Figure 4.4: A Wason Selection Task example

and Shapiro (1971) gave participants the rule “every time I travel to Manchester I travel by train”, with the cards ‘Manchester’, ‘Leeds’, ‘train’ and ‘car’. While success on the abstract version of the task tends to be lower than 10%, in the thematic case 10 out of 16 participants selected the correct cards: Manchester and car. The facilitative effect of thematic content doesn’t always hold true though, it appears only to help when the rule is deontic rather than descriptive, i.e. conveying a rule, permission or obligation, such as ‘if a person is drinking alcohol then they must be over 18 years of age’ or ‘if a person is on the train then they must have a valid ticket’. It has been argued that this is because familiar rules elicit evolutionarily developed schema, for the purposes of such things as cheater detection (Cosmides, 1989; Cosmides & Tooby, 1992).

Despite its popularity the effectiveness of the WST as a measure of reasoning has been challenged. Sperber et al. (1995) have suggested that the task doesn’t necessarily measure conditional reasoning at all. Instead, performance is highly influenced by relevance-guided mechanisms that pre-empt any reasoning mechanisms. When faced with a reasoning problem, or any other text, we first need to comprehend the information given and this will include inferring the writer’s intended meaning. In the case of inference tasks participants need to infer or evaluate conclusions derived from premises, so although their interpretation of the premises may be influenced by relevance principles, it is explicitly clear that they must go further and engage in reasoning processes as well. In the case of the selection task, participants are not asked to reason from premises to conclusions but are instead asked to judge the relevance of each of the cards to the rule. In this case, the judgments of relevance that come from the comprehension

p	q	if p then q
T	T	T
T	F	F
F	T	T
F	F	T

Table 4.5: Truth Table for ‘if p then q ’ where T = true and F = false.

process provide an intuitive answer to the problem and there is no explicit need to engage in any further reasoning. This may be the source of the pervasive matching bias.

This interpretation of the selection task was supported by six studies across two papers (Sperber et al., 1995; Sperber & Girotto, 2002). Sperber and his colleagues have shown that success rates in the task can be dramatically manipulated by altering the relevance factors of the content. Success in descriptive versions of the task can be increased to over 50%, more in line with the success rates usually found with deontic versions (Sperber et al., 1995). Furthermore, success in deontic versions can be reduced to below 20%, similar to the rates usually found with descriptive and abstract versions (Girotto, Kimmelmair, Sperber & van der Henst, 2001).

Due to this controversy, the WST was ruled out as a measure of reasoning ability for this thesis. Had mathematics students been found to change in WST performance alongside their mathematics study, it would be unclear whether the change had been in conditional reasoning ability or in interpretation processes.

Truth Tables

Truth Table tasks were discussed in Section 4.3.1 as a measure of disjunctive reasoning, but they can also be used for measuring conditional reasoning. Table 4.5 presents a Truth Table for the conditional rule ‘if p then q ’. Again, each line represents a different combination of truth and falsity of the values p and q , and the final column denotes whether that combination makes the conditional rule true or false. Again, Truth Table tasks can be given to participants to complete in thematic as well as abstract form, as demonstrated in Table 4.6 for the conditional ‘If it rains then I take an umbrella’.

By asking participants to decide which lines of the table they consider to be valid, it is possible to determine which reading of the conditional most closely matches their interpretation. As discussed in Section 2.2, there are at least four ways in which people may interpret a conditional statement: material con-

If the letter is H then the number is 5
 The letter is H
 Conclusion: The number is 5
☐ Yes
☐ No

Figure 4.5: Example Modus Ponens item from the Conditional Inference task.

ditional (that endorsed by formal propositional logic), biconditional, defective conditional and conjunctive conditional, and these differ in terms of Truth Tables (see Table 2.2).

Conditional Inference Task

In a Conditional Inference task, participants are given a conditional rule along with a premise about that rule, followed by a conclusion derived from the rule and premise. The participant then deduces whether the conclusion is valid or invalid. Alternatively, participants may generate a conclusion that they consider to be valid, but this is far less common in the literature and so the evaluation version is focused on here.

Figure 4.5 shows a typical conditional inference task item with a valid conclusion. This is an example of a Modus Ponens inference, one of four inference types used in the task. The four inferences, modus ponens (MP), denial of the antecedent (DA), affirmation of the consequent (AC) and modus tollens (MT) are shown in Table 4.7, along with the four rule forms created by rotating the presence of negations, and whether the inferences are considered valid according to the four interpretations.

An abstract (using only letters and numbers) 32-item version of the conditional inference task was used by Inglis and Simpson (2008, 2009a) to compare mathematics and non-mathematics students' reasoning behaviour. The task in-

<i>it rains</i>	<i>I take an umbrella</i>	if it rains then I take an umbrella
T	T	
T	F	
F	T	
F	F	

Table 4.6: Truth Table for the conditional rule 'if it rains then I will take an umbrella', where T = true and F = false.

If the letter is S then the number is 6	If the letter is M then the number is 4
The number is not 6	The letter is not M
Conclusion: The letter is not S	Conclusion: The number is not 4
<input type="radio"/> Yes	<input type="radio"/> Yes
<input type="radio"/> No	<input type="radio"/> No
a) Modus tollens	b) Denial of the antecedent

Figure 4.6: Example items from the Conditional Inference task showing a) a Modus Tollens inference and b) a Denial of the Antecedent inference.

cluded the four inference types, each presented four times with the four different rule forms shown in Table 4.7, which were created by varying the position of negatives. This created 16 items, with explicit negations. In a further 16 items the problems were identical in structure except that the negations were implicit (e.g. ‘not 3’ might be represented as ‘6’). Figure 4.6 shows some example items from the task.

Inglis and Simpson (2009a) found that mathematics students outperformed non-mathematics students (based on the material conditional being the normative reading, see Section 2.3), even when the groups were matched for general intelligence. The mathematics students did not improve in task performance over the course of a year, but the initial difference left open two possibilities: the mathematics students may have improved in conditional reasoning during

Conditional	MP		DA		AC		MT	
	Pr	Con	Pr	Con	Pr	Con	Pr	Con
if p then q	p	q	$\neg p$	$\neg q$	q	p	$\neg q$	$\neg p$
if p then $\neg q$	p	$\neg q$	$\neg p$	q	$\neg q$	p	q	$\neg p$
if $\neg p$ then q	$\neg p$	q	p	$\neg q$	q	$\neg p$	$\neg q$	p
if $\neg p$ then $\neg q$	$\neg p$	$\neg q$	p	q	$\neg q$	$\neg p$	q	p
Minor Premise Type	Affirmative		Denial		Affirmative		Denial	
Material Validity	Valid		Invalid		Invalid		Valid	
Defective Validity	Valid		Invalid		Invalid		Invalid	
Biconditional Validity	Valid		Valid		Valid		Valid	
Conjunctive Validity	Valid		Invalid		Valid		Invalid	

Table 4.7: The four inferences and conditional statement types with and without negated premises (Pr) and conclusions (Con). The symbol \neg should read ‘not’. At the bottom, the validity of each inference under each interpretation is given.

pre-university but post-compulsory study of mathematics, in line with the TFD, or it could be the case that people with more normative reasoning styles are disproportionately filtered into studying post-compulsory mathematics.

In an interview study conducted by Inglis (2012), eight stakeholders in the mathematics community (e.g. members of the education committees of the Institute of Mathematics and its Applications and the London Mathematical Society) were asked to look at the Conditional Inference task and rate their agreement with the statement “This task captures some of the skills that studying advanced mathematics develops”. Of those eight participants, six strongly agreed with the statement (five on a five-point Likert scale) and two agreed (four on a five-point Likert scale). One participant even went as far as to say that “If studying A-level maths doesn’t make you better at that, there is something wrong with the syllabus”.

4.3.3 Syllogisms tasks

Syllogisms represent the oldest form of formal deductive reasoning, dating back to the time of Aristotle over 2,000 years ago (Manktelow, 1999). They are formed from statements of four forms: ‘All A are B’, ‘Some A are B’, ‘No A are B’, and ‘Some A are not B’, known as A, I, E and O respectively. A and I are affirmative while E and O are negative, and A and E are universal while I and O are particular. A syllogism such as ‘All A are B; All B are C; All A are C’ is an example of an AAA structure. In all syllogisms, the conclusion will describe a relationship between the first and last terms of the premises, in this case, A and C. It is possible to construct 512 different syllogisms, but the amount to be considered valid has not been agreed upon: figures range widely, between 14 and 48 (Manktelow, 1999). Human interpretation of syllogisms has been investigated widely, using both abstract and contextual tasks.

Belief Bias Syllogisms Task

A commonly used task is the Belief Bias Syllogisms Task (Sá et al., 1999). In this task, participants see 24 syllogisms, 12 of which are valid and 12 of which are invalid, and they are asked to decide which is which. However, the syllogisms are made to be either believable, unbelievable or belief-neutral (there are 8 of each type, 4 of which are valid and 4 of which are invalid, examples of each type are given in Figure 4.7), and so it is possible to determine the extent to which participants are persuaded by belief and by logic (participants are instructed to put their beliefs to one side and reason logically, but this can be difficult to do). For example, in a valid syllogism with unbelievable content, such as ‘All things that are smoked are good for the health; cigarettes are smoked; therefore

Believable, valid:	Believable, invalid:
Premises: All fish can swim.	Premises: All living things need water.
Tuna are fish.	Roses need water.
Conclusion: Tuna can swim.	Conclusion: Roses are living things.
Unbelievable, valid:	Unbelievable, invalid:
Premises: All things with four legs are dangerous.	Premises: All guns are dangerous.
Poodles are not dangerous.	Rattlesnakes are dangerous.
Conclusion: Poodles do not have four legs.	Conclusion: Rattlesnakes are guns.
Neutral, valid:	Neutral, invalid:
Premises: All ramadions taste delicious.	Premises: All lapitars wear clothes.
Gumthorps are ramadions.	Podips wear clothes.
Conclusion: Gumthorps taste delicious.	Conclusion: Podips are lapitars.

Figure 4.7: Example items from the Belief Bias Syllogisms task.

cigarettes are good for the health’ a person who can overcome their prior belief and reason with logic would be more likely to accept the syllogism as valid than someone who is more greatly swayed by their beliefs than by the instruction to ignore them. What the Belief Bias Syllogisms task provides, therefore, is a measure of participants’ ability to reason independently of their prior beliefs. This is considered a central component of critical thinking (Sá et al., 1999; West et al., 2008).

There are two measures that can be taken from the Belief Bias Syllogisms task: total score, reflecting syllogistic reasoning ability across item-type, and a Belief Bias Index (BBI), which is the total number of consistent items endorsed (valid-believable, invalid-unbelievable) minus the total number of inconsistent items endorsed (valid-unbelievable, invalid-believable) and reflects participants’ ability to reason based on logical validity over believability.

A great deal of research has been conducted with the Belief Bias Syllogisms task, and there are three main findings: valid items are accepted more often than invalid items, believable items are accepted more often than unbelievable items, and believability and validity interact (Manktelow, 1999). Evans et al. (1983) found that valid items were accepted more often than not whether they were believable or unbelievable, but that invalid items were only accepted less often than not when they were unbelievable; believable invalid items were incorrectly accepted 66% of the time. This suggests that when a syllogism is easily believed, people tend to accept it without further thought. Evans et al. (1983) suggested a selective scrutiny model to account for the finding. The idea is that people initially reason with a heuristic that tells them to accept believable items straight

away. Only unbelievable items are subjected to further scrutiny with deliberate and conscious Type 2 processing.

4.4 Summary

It should be clear by now that there are many tasks that have been used to measure various aspects of human reasoning. It is simply not possible to use all of these tasks in a research study and still expect to find willing participants. In particular, the longitudinal studies presented in Chapters 5 and 6 of this thesis were carried out during school and university classes, and so the measures used needed to fit into the length of a standard class (50 minutes). To allow time for covariates to be measured as well as reasoning ability, it was decided that just two of the many reasoning tasks discussed above should be selected.

The aim of this thesis is to investigate changes in reasoning behaviour in mathematics compared to non-mathematics students, and to investigate possible mechanisms for any changes found. It has been suggested that conditional reasoning is central to logical reasoning (Anderson & Belnap, 1975; Braine, 1978; Inglis & Simpson, 2008) and also to mathematics (Houston, 2009). Houston (2009) argued that most mathematical statements are of the form ‘if statement A is true, then statement B is true’, even if they are heavily disguised (p. 63). For this reason, conditional reasoning may be a useful place to begin in the investigation of reasoning skills in mathematics students. However, as seen above, there are three main tasks that have been used to measure conditional reasoning skills: the Wason Selection Task, Truth Table tasks, and the Conditional Inference task.

The findings discussed in Section 4.3.2 suggest that the Conditional Inference task would be the most appropriate measure of conditional reasoning to use in this thesis. Firstly, the task has been widely used and is widely respected in the psychology of reasoning literature (it is free from the type of criticisms that have been levied against the Selection Task, e.g. Sperber et al., 1995). Secondly, it measures an aspect of reasoning on which mathematics and non-mathematics students have already been found to differ (Inglis & Simpson, 2008, 2009a). Thirdly, it would allow me to differentiate between the two hypotheses left open by Inglis and Simpson (2009a): the development hypothesis and the filtering hypothesis. Finally, it was considered by a group of stakeholders in the mathematics education community to be a good measure of the skills that studying mathematics develops (Inglis, 2012).

In addition to the Conditional Inference task, including the Belief Bias Syllogisms task as a measure of reasoning in this thesis allowed me to broaden the scope of the work in two ways: by measuring syllogistic reasoning ability as well

as conditional reasoning ability, and by measuring reasoning with thematic content as well as reasoning with abstract material. The Belief Bias Syllogisms task provides measures of syllogistic reasoning ability and the extent to which people are able to reason with logic over beliefs, and this is an important component of critical thinking (Sá et al., 1999; West et al., 2008).

Recall the quote from the Smith (2004) report that studying mathematics “develops logical and critical reasoning, and develops analytical and problem-solving skills to a high degree” (p. 11.). The Conditional Inference task relates to the logic aspect of this quote, while the Belief Bias Syllogisms task relates to the critical thinking aspect.

Chapter 5

The development of reasoning skills in AS level mathematics students

5.1 Introduction

5.1.1 Testing the Theory of Formal Discipline

The aim of this thesis is to investigate the Theory of Formal Discipline (TFD) claim that studying mathematics improves reasoning skills. The two main research questions are (a) is studying mathematics at advanced levels associated with improvement in reasoning skills and (b) if there is improvement, what might its mechanism be? Some light will be shed upon both of these issues in the current chapter, although that is not to say that both will be conclusively resolved.

The TFD suggests that studying mathematics improves one's logical reasoning and critical thinking skills. This belief is held by philosophers (Locke, 1971/1706; Plato, 2003/375B.C), mathematicians (Amitsur, 1998; Oakley, 1949), and policymakers (Smith, 2004; Walport, 2010) alike. As of yet, however, empirical evidence that supports the TFD is remarkably sparse. There is mixed evidence as to whether thinking skills can be transferred at all (see Chapter 2 for a review of the evidence), and although there is evidence of better reasoning skills from those who have studied mathematics at advanced levels than those who have not (Inglis & Simpson, 2009a), there is no evidence of reasoning skills developing alongside mathematical study. The study presented here investigated whether studying mathematics at AS level (the first year of an A level)

is associated with greater improvement in logical reasoning and critical thinking skills than studying English literature. Improvement in reasoning skills is defined as reasoning that is closer to the relevant normative model, as discussed in Chapter 2 (Stanovich, 1999).

The measures of logical reasoning and critical thinking that I will use are the Conditional Inference and Belief Bias Syllogisms tasks, respectively. The reasons for this were elaborated in Chapter 4, but briefly, conditional reasoning is considered central to logical reasoning (Anderson & Belnap, 1975; Braine, 1978; Inglis & Simpson, 2008), and the ability to decouple one's prior beliefs from logical validity is considered to be a part of critical thinking (Facione, 1990; Sá et al., 1999; West et al., 2008).

If it is found that studying mathematics is indeed associated with greater improvement in reasoning skills than studying English literature, it would beg the question of what the mechanism for such an improvement could be. Stanovich's (2009a) tripartite model was introduced in the literature review as a starting point for identifying potential mechanisms for improvement in reasoning skills. The tripartite model is an extension of dual-process models of reasoning which propose fast and automatic Type 1 processes and slow and deliberate Type 2 processes (Evans, 2003). In the tripartite model Type 2 processing is said to occur at two levels – the algorithmic and reflective levels. The algorithmic level is the computational element to Type 2 processing – the capacity available for effortful processing and the efficiency with which effortful processing can occur. The reflective level is the dispositional element – the processing that regulates when and to what extent the algorithmic level will be used as opposed to Type 1 processing.

It is possible that improvement in reasoning skills could be brought about by changes to any of these three types of processes. Studying mathematics could alter the Type 1 processes that focus our attention on certain aspects of a problem. Alternatively it could improve algorithmic level capacity for or efficiency of effortful reasoning. Finally, it could be that studying mathematics alters the reflective level making the reasoner more willing to put effort into thinking. Alternatively to domain-general factors, a source of improvement in reasoning could be what Stanovich (2009a) termed “mindware”. Mindware consists of domain-specific knowledge, rules and procedures that can be explicitly recalled from memory to aid in solving specific problems, rather than being useful for reasoning more generally. Perhaps knowledge, rules or procedures are taught in mathematics that assist reasoners when solving certain types of tasks.

Type 1 processing is best studied by reaction time or time-limited accuracy measures due to their speed and automaticity (Gillard, 2009; Evans & Curtis-Holmes, 2005). Such methods allow initial responses to be isolated from later

responses that are based on Type 2 processing, but they require computer-based administration and as such were not suitable for the study reported in this chapter (where measures were all administered on paper in schools). Instead, the role of Type 1 processing in the differences between mathematicians' and non-mathematicians' reasoning behaviour is explored in Chapter 8. Here, the algorithmic and reflective levels of Type 2 processing are investigated as potential mechanisms for improvement in reasoning skills.

The algorithmic level can be assessed via measures of intelligence and executive functions, which reflect cognitive capacity and efficiency. The relationship between executive functions and reasoning skills is investigated separately in Chapter 9, while intelligence is considered here. If studying mathematics is associated with improved reasoning skills due to an increase in the capacity for Type 2 processing then this may be reflected in an increased score on intelligence tests. A subset of items from Raven's Advanced Progressive Matrices (RAPM, Raven et al., 1998) were included as a measure of intelligence in the main study presented below to allow for such a mechanism to be identified. As noted in Chapter 2, RAPM is a non-verbal pattern completion task that is thought to be the best single measure of general intelligence (Jensen, 1998). A time-limited subset of items selected for a student population has been shown to have an acceptable split-half reliability of .79 (Stanovich & Cunningham, 1992). This also allowed between-groups differences in intelligence at Time 1 to be controlled for (see the discussion on quasi-experimental methods in Chapter 3).

The reflective level of cognition is assessed by measures of thinking disposition, such as the Actively Open-minded Thinking scale (AOT, Stanovich & West, 1997), the Need for Cognition scale (NFC, Cacioppo et al., 1984) and the Cognitive Reflection Test (CRT, Frederick, 2005). The Cognitive Reflection Test, introduced in Chapter 2, poses three questions which prompt intuitive but incorrect responses. Participants need to inhibit these responses in order to answer correctly. Recently, Toplak et al. (2011) demonstrated that the CRT was a better predictor of normative reasoning than intelligence, executive functions, or the AOT. This suggests two things: firstly, if studying mathematics is associated with more normative reasoning, then changes to the reflective level may be a more likely mechanism than changes to the algorithmic level, and secondly, the reflective level may be better tapped by performance measures (e.g. the CRT) than self-report measures (e.g. the AOT and NFC scales). Here, the reflective level will be measured with two tasks: the CRT, and the NFC scale. The NFC scale was not included in Toplak et al.'s (2011) assessment and it may provide additional predictive power.

As of yet, there is no suggestion of what the nature of any mathematical mindware responsible for improvement might be, so a starting point is to simply

look at the effect of subject studied (mathematics or non-mathematics) and whether this predicts improvement independently of domain-general factors. If domain general factors are not found to be predictors, then mathematical mindware is a remaining possibility.

As a side point, if mathematics develops reasoning skills via domain-general mechanisms then we might expect a wide-spread improvement in reasoning performance. If, on the other hand, mathematics provides some specific mindware, we might expect improvement on only a small set of reasoning tasks for which the mindware is relevant. There will not be enough reasoning measures to test this hypothesis thoroughly here, but perhaps proponents of the TFD would hope for domain-general changes to be responsible for any improvements so that mathematics could be said to have a more useful, widespread influence on thinking skills.

5.1.2 Summary

There are two research questions addressed in this chapter: (a) is studying mathematics at advanced levels associated with improvement in reasoning skills? and (b) if there is such improvement, what might the mechanisms behind it be? These questions are addressed with a longitudinal study, in which mathematics AS level students are compared to English literature AS level students for development in logical reasoning and critical thinking skills. The participants were tested twice, at the beginning and end of their AS year of study.

The Conditional Inference Task was used as a measure of logical reasoning skills and the Belief Bias Syllogisms task was used as a measure of one aspect of critical thinking. Measures of intelligence (RAPM) and thinking disposition (CRT and NFC) were included to indicate whether domain-general factors at the algorithmic or reflective levels of cognition (Stanovich, 2009a) were the mechanisms for any development found. These measures also allowed pre-existing differences between the groups, which are likely to exist due to the quasi-experimental design (see Chapter 3), to be statistically controlled for. Finally, a mathematics test was used as manipulation check (to confirm that the mathematics students did learn more mathematics than the English literature students).

Before describing the longitudinal study in more detail, three pilot studies are presented. The first tested whether the Belief Bias Syllogisms task could be split in half in order to save testing time. The second investigated whether the CRT could be imbedded in mathematical word problems, for the sake of making the ‘trick’ nature of the questions less memorable without altering the way that participants respond to them. The final pilot study assessed the duration and

difficulty of the selected measures to ensure suitability for the study.

5.2 Pilot Studies

5.2.1 Pilot 1: Splitting the syllogisms task

In the Belief Bias Syllogisms task, 24 syllogisms are presented in a contextualised format (Sá et al., 1999; see also Chapter 4). In eight of the syllogisms, prior belief and validity are in accordance (four believable-valid, four unbelievable-invalid), in another eight they are in conflict (four believable-invalid, four unbelievable-valid), and in the final eight the context is neutral (four neutral-valid, four neutral-invalid). This creates four problems for each of six believability-logic combinations (see Figure 5.1 for an example of each type). Two of each of these four make positive (P, Q) statements, and two of the four make negative (not-P, not-Q) statements. Therefore, there are twelve combinations of believability, validity and valence. Two problems for each of these combinations makes the total of 24 items. Because there are two problems of each form, the test can be split in half and still cover all combinations to give a full measure of belief bias.

The test could be split in this way so that half of the problems could be given to participants at Time 1 and the other half at Time 2 (with the order counterbalanced by school), to reduce repeat testing effects and testing time. This method means that any difference between the two halves in terms of believability could cause misleading gains or losses in belief bias between the two time points. Therefore, a pilot study was required to determine whether

Believable, valid:	Believable, invalid:
Premises: All fish can swim.	Premises: All living things need water.
Tuna are fish.	Roses need water.
Conclusion: Tuna can swim.	Conclusion: Roses are living things.
Unbelievable, valid:	Unbelievable, invalid:
Premises: All things with four legs are dangerous.	Premises: All guns are dangerous.
Poodles are not dangerous.	Rattlesnakes are dangerous.
Conclusion: Poodles do not have four legs.	Conclusion: Rattlesnakes are guns.
Neutral, valid:	Neutral, invalid:
Premises: All ramadions taste delicious.	Premises: All lapitars wear clothes.
Gumthorps are ramadions.	Podips wear clothes.
Conclusion: Gumthorps taste delicious.	Conclusion: Podips are lapitars.

Figure 5.1: Example items from the Belief Bias Syllogisms task.

any such differences did exist. In the pilot study the conclusions from each problem, which are where the believability/validity conflicts lie, were rated by participants in terms of how believable they were. The problems of the same format from each half of the test (e.g. the believable problems from Half 1 and Half 2) were then compared for believability ratings.

Methods

Participants Fifty-eight participants (38 male, aged 19-23, $M=20.12$) were recruited by email through a mathematics module tutor and took part unpaid during a larger online study. All participants were undergraduate mathematics and engineering students at Loughborough University.

Procedure Participants took part during an unrelated online study about their degree course study choices (see Inglis, Palipana, Trenholm & Ward, 2011). After completing all sections relevant to their study choices, they were asked to complete a section on the believability of sentences. The instructions read: “Below is a list of sentences. Some of the sentences will be completely believable, some will be completely unbelievable, some will be roughly in the middle, and some will be meaningless. Your task is to decide which is which”.

Below the instructions, on the same page, the 24 conclusions from the Belief Bias Syllogisms task (see Appendix C) were presented in a set order, alternating between Half 1 conclusions and Half 2 conclusions. Next to each sentence was a 5 point scale with the options ‘Very unbelievable’, ‘Moderately unbelievable’, ‘Neither believable nor unbelievable’, ‘Moderately believable’, and ‘Very believable’. Participants rated each statement on the scale before submitting their answers.

Results

The 24 syllogism conclusions fell into six categories for the analysis: Half 1 unbelievable (H1U), half 1 neutral (H1N), half 1 believable (H1B), half 2 unbelievable (H2U), half 2 neutral (H2N), and half 2 believable (H2B). The aim of the analysis was to test for differences between test halves on believability ratings of each item type, i.e. are H1U conclusions rated differently to H2U conclusions?

Participants’ mean responses are shown in Figure 5.2. A 2 (test half: 1 or 2) \times 3 (intended believability: unbelievable, neutral, believable) repeated measures analysis of variance (ANOVA) was conducted with believability ratings as the dependent variable. There was a significant main effect of intended believability on believability ratings, $F(2, 114) = 323.2, p < .001, \eta_p^2 = .85$. Post hoc tests showed that believable items were rated as significantly more

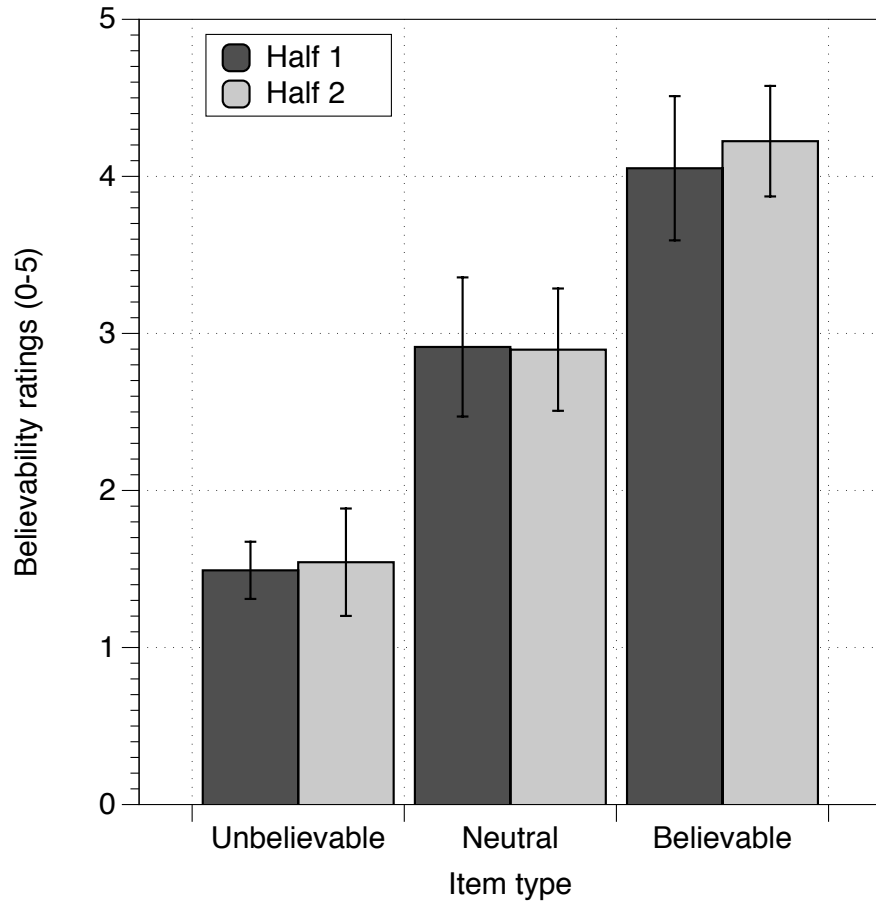


Figure 5.2: Believability ratings for each problem type by test half (error bars represent ± 1 SE of the mean).

believable ($M = 4.14, SD = 0.50$) than neutral items ($M = 2.90, SD = 0.57$), which in turn were rated as significantly more believable than unbelievable items ($M = 1.50, SD = 0.51$). There was no significant main effect of test-half, meaning that there was no evidence of the two halves of the test differing in believability. Importantly, there was also no significant interaction between half and believability ($p = .20, \eta_p^2 = .03$). H1U ($M = 1.49, SD = 0.57$), H1B ($M = 4.05, SD = 0.69$) and H1N ($M = 2.91, SD = 0.59$) conclusions were rated similarly to H2U ($M = 1.54, SD = 0.58$), H2B ($M = 4.22, SD = 0.55$) and H2N ($M = 2.90, SD = 0.58$) conclusions, respectively, so there was no evidence that conclusions in each test half that had the same intended believability status differed in rated believability (see Figure 5.2).

Discussion and Implications

There are three findings from this study: 1) each half of the test can be assumed to be equally believable (at least, there was no evidence of a difference), 2) problems of the same intended believability in each half can be assumed to be equally believable (again, insofar as there was no evidence of a difference), and 3) the intended believability of problems over the test as a whole are in accordance with participants' perceptions of believability. All three of these outcomes are positive for the use of different halves of the test at different time points. The results suggest that a participant would show consistent extents of belief bias in each half of the test, and any difference over time found in the longitudinal study will not be due to a difference in the items used but due to a genuine difference in the participants' susceptibility to belief bias.

5.2.2 Pilot 2: Disguising the Cognitive Reflection Test

To the best of my knowledge, the CRT has not been used in any longitudinal studies before now. A particular concern is that the 'trick' nature of the questions might make them more memorable than other measures at post-test and that this will affect the way in which participants (at least those who inhibited their intuitive response and are thus aware of the 'trick') respond to them.

An attempt to address this issue was to mix the CRT questions in with non-trick mathematical word problems, so that the trick might be less salient at repeated testing points. However, it is possible that this could alter the test properties in some way, perhaps meaning participants no longer process the CRT questions in the way they otherwise would have. The aim of the pilot study reported here was to assess whether this was the case. Participants either saw the three CRT questions followed by the three non-trick problems or they saw the six items in a random order. The three non-trick questions, shown in Figure 5.3, were taken from the Woodcock-Johnson III Applied Problems subtest and were selected for being mathematically simple and of a similar length to the CRT questions. If mixing the questions does not alter the test properties then scores should not significantly vary between groups.

Method

Participants Participants were recruited, without payment, through websites that advertise internet based research studies where they saw a brief description and could open the study webpage. Fifty-four participants completed all six questions and were included in the analysis. The participants were aged 18-59 ($M=29.60$, $SD=10.40$), and 23 were male and 31 were female. Twenty-

1. If a girl saved £1 each week for 1 year, how much money would she have at the end of that year?
2. If a dog can run two and a quarter miles in one hour, how long would it take the dog to run four and a half miles at that same rate?
3. Mileage varies from car to car. Judy's car gets 22 miles to a gallon of gas, and Bob gets 35 miles to a gallon of gas. How many miles can Judy drive on six gallons of gas?

Answers: Q1 = £52, Q2 = 2 hours, Q3 = 132 miles

Figure 5.3: The three items from the Woodcock Johnson III Applied Problems subtest that were used in Pilot Study 2.

seven were randomly allocated to the mixed condition and 27 to the non-mixed condition when they opened the webpage.

Procedure Participants first saw a page providing information about the study. They were told that if they took part they would be asked to answer six arithmetic word problems which would take no more than 10 minutes. They were also told that their data would be kept confidential and used for research purposes only. They were asked to select whether they wanted to seriously participate or just browse the pages before continuing to the study, and only those who wanted to seriously participate were included in the analysis.

Next they were asked to report their sex, age, degree subject (if applicable), and whether their native language was English or non-English. The six questions were each presented on a separate page and required participants to type their answer into a blank response box. In the mixed condition, the six questions were presented in a random order. In the non-mixed condition, the three CRT questions were presented first in a random order, followed by the three non-trick questions in a random order. Finally, participants were thanked for their participation and given my email address in case they wanted to request further information or to withdraw their data (none did).

Results

Two Mann-Whitney U tests were used to compare performance across the two conditions: one analysing number of correct responses to the three CRT questions, and one analysing number of intuitive responses (it is possible, but rare, to give a non-intuitive but incorrect response meaning that these measures are

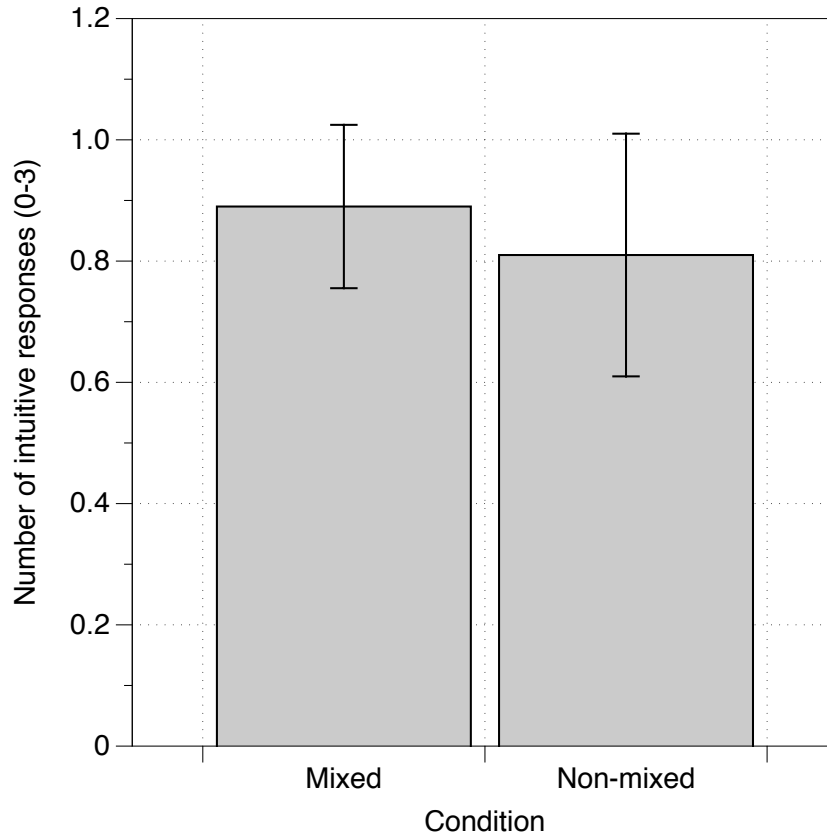


Figure 5.4: Mean number of intuitive responses in the mixed and non-mixed conditions (error bars represent ± 1 SE of the mean).

not exact inverses of each other).

Number of correct responses were not significantly different in the mixed ($M=2.00$, $SD=0.73$) and non-mixed groups ($M=1.93$, $SD=1.07$), $U(54)=358.0$, $z = -.12$, $p=.904$, $r = 0.02$. Similarly, number of intuitive responses were not significantly different in the mixed ($M=0.89$, $SD=0.70$) and non-mixed groups ($M=0.81$, $SD=1.04$), $U(54)=417.00$, $z = -.98$, $p=.325$, $r = 0.13$, see Figure 5.4.

Discussion and Implications

The aim of this pilot study was to assess whether inter-mixing the three CRT questions with three non-trick questions would affect the way in which participants respond to the CRT questions. There was no evidence that this was the case. Neither the number of correct responses nor the number of intuit-

ive responses given to the CRT questions significantly differed when they were inter-mixed with non-trick questions compared to when they were seen first.

One thing this result suggests is that the reflective level determines on an item-by-item basis whether the algorithmic level should be used. It is not the case that answering an item with no intuitive answer which thus requires algorithmic level processing (any of the Woodcock-Johnson questions) sets the reflective level to a ‘use algorithmic processing’ mindset. If this were the case the presence of an item that requires algorithmic processing should mean that the intuitive answer is inhibited for subsequent CRT questions, which was not the case here. This fits with the established finding that participants do not tend to get all CRT questions right or wrong, rather they may get one or two right and give the intuitive responses to the rest (Frederick, 2005).

Returning to the purpose of this pilot study, the results indicate that the three CRT questions and three Woodcock-Johnson questions can be randomly inter-mixed in the longitudinal study without affecting the way participants respond to the CRT questions, which are the real measure (responses to Woodcock-Johnson questions will not be analysed). The Woodcock-Johnson questions appear to be simple enough as to not influence the reflective level in a task-general manner. It is hoped that their inclusion will reduce repeat-testing effects on CRT performance.

5.2.3 Pilot 3: Duration and difficulty of measures

There were several measures selected for the longitudinal study reported below for which the average completion time was unknown. In order to plan an appropriate time slot(s) with schools, a pilot study was conducted to determine the total session duration required. Five undergraduate mathematics students were recruited to complete the entire set of tasks (demographics, the RAPM subset, the Conditional Inference Task, the Belief Bias Syllogisms task, the CRT, the Need for Cognition scale, and a mathematics task). The aim was to record the duration of each individual measure and the overall duration of the test for each participant, so that an average duration could be derived. A further aim was to assess whether the measures being used were of an appropriate difficulty and whether the instructions for each measure were clear to participants unfamiliar with the tasks.

Methods

Participants The participants were five undergraduate mathematics students (one male, four females), aged 19 to 51 ($M=25.80$, $SD=14.10$) who took part in return for £15 each. Recruitment was through an email advertisement to

the students on a differential equations module for first and second year undergraduates. This sample was assumed to be of a higher general ability than the AS level students of the longitudinal study, although not greatly. Presumably it tends to be the most able AS level students who go on to degree level study, so the undergraduate sample here may be similar to the more capable students in the AS level sample. The implication of this is that the scores found here may be slightly higher than in the AS level sample, and so any floor effects found should be of particular concern.

Procedure. Participants were informed that they would be given a test book to complete. They were told that the aim of the study was to determine the length of the test, which would be used in a large scale study in the future. They were then asked to sign a consent form before taking part. Two participants took part simultaneously but working alone, and three took part individually. All testing took place in a quiet seminar room.

Participants were asked to work through the booklet at their own pace, informing the experimenter when they reached the end of a section and began the next section. The sections were presented in a set order for all participants. After they had completed the booklet, participants were asked whether any part of the test was unclear, too easy or too difficult, and whether they had any other comments. They were then thanked, paid and dismissed.

Results

The first section of the results will deal with the length of the measures, the second section will deal with the range of scores obtained, and the final section will discuss the participants' comments on the clarity of the tasks.

Duration. Table 5.1 shows the descriptive statistics for the length of time taken for each measure. The mean total duration was 45.40 minutes with a standard deviation of 11.63 minutes. However, the total duration data is positively skewed (2.13) with four of the data points in the range 39-43 minutes and one data point of 66 minutes. Therefore, four of the five participants completed the test faster than the average time of 45.4 minutes.

The RAPM section of the test has a time limit of 15 minutes. However, it can be seen from the table that some participants finished more quickly than this.

Scores. Mean scores were examined to indicate whether any of the tasks suffered from floor or ceiling effects. Table 5.2 shows the descriptive statistics for the scores obtained on each measure as well as the theoretical minimum and maximum scores. The CRT scores were not examined because the participants had recently been exposed to the task as part of another study. As the table

Measure	Mean	SD	Min	Max
Demographics	2.2	1.64	1	5
Raven's Matrices	14.0	1.73	11	15
Conditional Inference	10.6	3.13	8	16
Belief Bias Syllogisms	3.8	1.3	3	6
Cognitive Reflection Test	1.2	0.45	1	2
Need For Cognition	3.2	1.79	2	6
Mathematics	10.4	3.64	7	16
Total	45.4	11.63	39	66

Table 5.1: Duration information for each measure used in the test book (units are in minutes).

shows, none of the mean task scores were at, or approaching, the theoretical minimum or maximum scores, with the possible exception of the Belief Bias Syllogisms task approaching the theoretical maximum.

Additional comments. All five participants commented that the Conditional Inference Task was not completely clear on first reading of the instructions, although they did find that it became clear as they began to complete it (discussed below).

Discussion and implications

The results of the pilot test have shown that (a) the average duration of the entire test is 45 minutes, with the majority of participants finishing in less time, (b) the measures are at ceiling in some individuals but not on average, and (c) there was one issue noted by the participants; the instructions for the Conditional Inference Task.

The length of time taken means that the whole test can be completed within one school lesson, which are usually 50-60 minutes in length. Requiring only

Measure	Mean(SD)	Min Possible	Observed Min	Max Possible	Observed Max
Raven's Matrices	11.40(3.78)	0	8	18	16
Conditionals	22.40(7.23)	0	15	32	32
Syllogisms	10.60(1.67)	0	8	12	12
NFC	5.47(1.09)	0	4.56	9	7.33
Mathematics	10.20(2.49)	0	7	15	12

Table 5.2: Descriptive statistics for each measure used in the test book (standard deviations in parentheses).

one session will reduce the demand on teachers' and participants' time. On the other hand, it is evident from the one participant who took 66 minutes that some individuals may take longer than one lesson's worth of time, in which case they will either need to leave the test incomplete, stay after the lesson to finish, or continue at another time. This is something that can be decided with the input of the teachers concerned.

With reference to the second part of the results, the scores obtained were only problematic for the Belief Bias Syllogisms task, where there is a slight ceiling effect. It is worth re-emphasising that the participants in this pilot are educationally more advanced than the participants of the main study reported below. The main study will use participants from the beginning to the end of their AS year of study, whereas the participants here were at the end of their first or second year of undergraduate mathematics degrees, so it can be expected that they would perform higher on achievement tests than the majority of the main study participants will. Therefore, all of the measures piloted here are expected to provide enough variation to detect improvements over the course of an AS level.

It was noted by all participants that the Conditional Inference Task instructions were not completely clear. However, none of the participants could suggest how the instructions might be clarified even once they had completed the test and reported that they did understand it. This may reflect the unavoidably complicated nature of the task, since it is not something that is usually encountered in day-to-day life. The instructions used were adapted from Evans, Clibbens & Rood (1995), who did not report any similar issues in their large scale use of the measure. In the interest of consistency with published research, the instructions will be made identical to those used by Evans, Clibbens & Rood for the main study, and it is expected that even if the participants find the instructions complicated in isolation, the task will become clear once they start. An experimenter will always be present when participants complete these tasks, so there will be the opportunity to ask for clarification if necessary.

In sum, the pilot study described here has not raised any problems that require the measures selected to be altered or substituted.

5.2.4 Summary

Three pilot studies have been presented that assessed various aspects of the measures selected for the main study, and each has provided positive results. Next, the main study itself is presented.

5.3 Main study

To recap, there were two research questions for the main study:

1. Is studying mathematics at AS level associated with improvement in reasoning skills?
2. If there is such improvement, what might the mechanisms behind it be?

5.3.1 Method

Design

The study followed a longitudinal quasi-experimental design. Participants were recruited after they had chosen their AS level subjects and were tested at the beginning (during the first term and as close to the start of term as possible) and end (after teaching had finished) of their AS year of study. They completed the same set of tasks at both time points (with the exception of the Belief Bias Syllogisms task as described in Pilot 1).

Participants

One hundred and twenty four participants were recruited from five schools in Leicestershire, Hampshire and Derbyshire, UK. Seventy-seven were studying AS level mathematics or further mathematics amongst any other subjects and forty-seven were studying AS English literature and not mathematics. The English literature students served as a comparison group. Participants and their parents/guardians gave written informed consent.

Of the original sample, 44 of the mathematics students and 38 of the English literature students took part at both time points. There were no differences on any of the Time 1 measures between those who returned and those who did not ($ps > .15$). The mathematics group was composed of 21 females and 23 males and the English group was composed of 23 females and 15 males. Eighty participants reported their first language as English, one reported both Gujarati and English, and one did not report their first language. Four participants in the mathematics group reported having been diagnosed with dyslexia and two in the English group reported that they suspected having dyslexia. None reported having dyscalculia or any other relevant disabilities.

Participants also reported their GCSE (General Certificate of Secondary Education, end of compulsory education exams) grades. From this, a prior attainment score was calculated for each participant to be used as covariate in the analyses reported below. This was the sum of the grades achieved, with an A* being scored 8, an A being scored 7, a B scored 6 and so on. Scores were

summed rather than averaged to take account of the variation in the number of GCSEs taken – a student who achieved 10 GCSEs at grade A could be said to have a higher attainment level than a student who achieved 9 GCSEs at grade A, for example. Of those who studied AS level mathematics, 25 had achieved an A* at GCSE mathematics, 15 achieved an A, 3 achieved a B and 1 did not report their grade. Of those studying English literature, there were 5 A*s in GCSE mathematics, 6 As, 14 Bs, 11 Cs, 1 D and 1 E.

Mathematics Syllabus

There are three different versions of the AS level mathematics course available to students in England and Wales, which all have similar content. The syllabus contains sections on algebra, coordinate geometry, introductory calculus, trigonometry, probability, mathematical modelling, kinematics, Newtons laws of motion, and forces (e.g., Assessment and Qualifications Alliance, 2011), among other topics. The course is considered by some to be fairly basic and not adequate preparation for university-level mathematics study (e.g. Lawson, 1997, 2003), and most importantly, students are not taught any proof-based mathematics, nor the definition of the conditional. This was confirmed with an analysis of every AS-level mathematics examination between 2009 and 2011. Of 929 questions set, only one contained an explicit ‘if then’ sentence, and there were no mentions of the terms ‘modus ponens’, ‘modus tollens’ or ‘conditional’.

Measures

Conditional Inference. Participants completed the standard 32 item Conditional Inference Task (Evans et al., 1995), consisting of eight items each of four inference types: modus ponens (MP), denial of the antecedent (DA), affirmation of the consequent (AC) and modus tollens (MT). Half of the items used explicit negations (e.g. “not 5”) and half used implicit negations (e.g. “not 5” represented as, for example, 6). The inference types used are summarised in Table 5.3 and the full measure is presented in Appendix B.

The lexical content was generated randomly and the items were presented in a random order for each participant. Participants decided whether each item was valid (i.e. the conclusion necessarily followed, assuming that the premises were true) or invalid.

Six measures were taken:

1. A material conditional index (MCI, number of answers out of 32 consistent with the material interpretation), which was calculated as: number of MP inferences endorsed + (8–number of DA inferences endorsed) +

(8–number of AC inferences endorsed) + number of MT inferences endorsed.

2. A defective conditional index (DCI, number of answers out of 32 consistent with the defective interpretation), which was calculated as: number of MP inferences endorsed + (8–number of DA inferences endorsed) + (8–number of AC inferences endorsed) + (8–number of MT inferences endorsed).
3. A biconditional index (BCI, number of answers out of 32 consistent with the biconditional interpretation), which was calculated as: number of MP inferences endorsed + number of DA inferences endorsed + number of AC inferences endorsed + number of MT inferences endorsed.
4. A conjunctive conditional index (CCI, number of answers out of 32 consistent with the conjunctive interpretation), which was calculated as: number of MP inferences endorsed + (8–number of DA inferences endorsed) + number of AC inferences endorsed + (8–number of MT inferences endorsed).
5. A negative conclusion index (NCI), which was calculated as the number of inferences endorsed on arguments with negative conclusions minus the number of inferences endorsed on arguments with affirmative conclusions.
6. An affirmative premise index (API), which was calculated as the number of inferences endorsed on arguments with affirmative premises minus the number of inferences endorsed on arguments with negative premises.

The instructions given were identical to those used by Evans et al. (1995). An example item is shown in Figure 5.5.

Syllogisms. The Belief Bias Syllogisms task (presented in full in Appendix C) was used as a measure of the ability to reason independently of prior beliefs (Evans et al., 1983; Markovits & Nantel, 1989; Sá et al., 1999). The

Conditional	MP		DA		AC		MT	
	Pr	Con	Pr	Con	Pr	Con	Pr	Con
if p then q	p	q	$\neg p$	$\neg q$	q	p	$\neg q$	$\neg p$
if p then $\neg q$	p	$\neg q$	$\neg p$	q	$\neg q$	p	q	$\neg p$
if $\neg p$ then q	$\neg p$	q	p	$\neg q$	q	$\neg p$	$\neg q$	p
if $\neg p$ then $\neg q$	$\neg p$	$\neg q$	p	q	$\neg q$	$\neg p$	q	p

Table 5.3: The four inferences and conditional statement types with and without negated premises (Pr) and conclusions (Con). The symbol \neg should read ‘not’.

If the letter is H then the number is 5

The letter is H

Conclusion: The number is 5

☐ Yes

☐ No

Figure 5.5: Example item from the Conditional Inference Task.

task consisted of 12 contextualised syllogisms, four congruent (believable-valid, unbelievable-invalid), four incongruent (believable-invalid, unbelievable-valid) and four neutral (example items are shown in Figure 5.1). Participants decided whether each syllogism was logically valid or not after being instructed to ignore their prior beliefs. Two measures were taken: a total score out of 12, indicating syllogistic reasoning ability, and a Belief Bias Index. A Belief Bias Index is calculated for each participant by subtracting the number of incongruent items answered correctly from the number of congruent items answered correctly. The resulting score indicates the degree to which a person's answers are swayed by believability or validity. The Belief Bias Index can range from -4 to +4 with positive scores indicating some degree of belief bias.

Raven's Advanced Progressive Matrices (RAPM). An 18 item subset of RAPM (see Appendix A) with a 15 minute time limit (Sá et al., 1999) was used as a measure of general intelligence (or algorithmic level processing, Stanovich, 2009a).

Cognitive Reflection Test (CRT). As suggested by Toplak et al. (2011) the number of intuitive responses given to the three-item CRT (Frederick, 2005, see Figure 2.10) was used as a performance measure of the tendency to use Type 2 processing (at the reflective level). Scores were reversed so that higher scores represented more normative performance, in line with the other measures. The questions were randomly intermixed with three simple mathematical word problems of a similar length from the Woodcock-Johnson III Applied Problems subtest as described in Section 5.2.2.

Need for Cognition (NFC). The NFC scale (Cacioppo, Petty, Feinstein & Jarvis, 1996, see Appendix D) was included as a self-report measure of thinking disposition to compliment the performance based CRT measure. However, there were no between-group differences at Time 1 ($p = .616$) or at Time 2 ($p = .374$) nor a change over time in either group (both $ps > .670$), despite the measure correlating significantly at Time 1 with RAPM scores $r(122) = .19, p = .034$, Syllogisms scores, $r(122) = .31, p = .001$, Belief Bias Index, $r(122) = -.30, p = .001$, correct answers to the CRT, $r(122) = .19, p = .040$, the MCI, $r(120) =$

.21, $p = .022$, and the DCI, $r(122) = .21$, $p = .021$. Due to the lack of differences between groups or changes over time, NFC scores are not discussed any further.

Mathematics Manipulation Check. A 15 item mathematics test was included as a manipulation check. This was to ensure that the students who were studying AS level mathematics were indeed learning more mathematics than those studying AS level English literature and not mathematics. In the case of no improvement in reasoning skills in the mathematics group, it would be useful to be able to rule out the possibility that the reason they did not improve was because they did not actually improve in mathematics.

Twelve items were taken from the Woodcock-Johnson III Calculation subtest. Nine had shown an average accuracy of less than 55% and correlated with performance on the whole test at .86 in a previous dataset with mixed-discipline undergraduate students (Inglis, Attridge, Batchelor & Gilmore, 2011) and the remaining three were taken from the lower range to prevent floor effects in the English literature students.

The final three items were the most difficult questions on the Loughborough University diagnostic test for new mathematics undergraduates based on performance in 2008 and 2009. The diagnostic test is designed to test incoming students' capability with A-level mathematics, and the three items were included to prevent ceiling effects in the mathematics students at the second time point whilst ensuring that the content was not inappropriately advanced. Questions were presented in a set order that was intended to be progressive. The full task is presented in Appendix E.

Procedure

Participants took part in groups (5-34) during the school day under examination style conditions. All tasks were given in a single paper booklet. The RAPM task was always completed first to allow the 15 minute time limit to be enforced, and the order of the subsequent tasks was counterbalanced between-participants following a Latin Square design. Participants were instructed to work at their own pace until they had completed all tasks and the sessions lasted approximately 45 minutes. At four of the five schools, participants were entered into a prize draw to win either a Nintendo DS Lite or a portable DVD player at Time 1. At Time 2 participants were each paid £4 for taking part. In the fifth school, the teacher preferred the students to take part without external incentive.

5.3.2 Results

The results are reported in three sections: (i) preliminary analyses, (ii) development of Conditional Inference and Syllogisms scores, and (iii) mechanisms of

development.

Preliminary analyses

Data inclusion. Forty-four mathematics students and thirty-eight English literature students took part at both time points and were included in the analysis. Those who dropped out had typically changed courses; there were no significant differences in Time 1 scores on any of the measures between those who took part at Time 2 and those who dropped out ($ps > .15$).

Reliability. The reliability of the Conditional Inference Task was assessed with a large data set collated from several studies (including Inglis & Simpson, 2008, 2009a; Inglis, Attridge et al., 2011, and the studies reported in the current chapter and Chapter 6 of this thesis). This resulted in a pool of 656 participants' data from three universities and five schools and colleges. The Cronbach's alpha, .87, was found to be sufficiently high for the measure to be considered internally reliable.

The reliability of the Belief Bias Syllogisms task was assessed using only the data presented here. The Cronbach's alpha was found to be lower, at .65, but not unreasonably low. This may be due to the deliberately disparate nature of two types of items – the belief/validity consistent and inconsistent thirds of the task.

Covariates. The mathematics group scored significantly higher on the RAPM ($M = 9.57, SD = 3.26$) than the English literature group at Time 1 ($M = 7.03, SD = 3.45$), $t(80) = 3.43, p = .001, d = 0.76$. The mathematics group also scored significantly higher on the CRT (reversed number of intuitive answers, $M = 1.77, SD = 1.12$) than the English literature group at Time 1 ($M = 0.89, SD = 0.86$), $U(82) = 466.00, z = -4.29, p < .001, r = .39$. Prior academic attainment scores ranged from 30 to 99 ($M = 64.10, SD = 11.94$) and were marginally higher in the mathematics group ($M = 66.26, SD = 9.75$) than the English literature group ($M = 61.66, SD = 13.45$), $t(79) = 1.75, p = .084, d = 0.39$.

MCI scores at Time 1 were significantly correlated with the RAPM, $r(79) = .41, p < .001$, CRT, $r(79) = .42, p < .001$, and prior attainment scores, $r(78) = .30, p = .007$. Consequently, RAPM, CRT and prior academic attainment scores are used as covariates in subsequent analyses of Conditional Inference scores.

Syllogisms scores at Time 1 were also significantly correlated with RAPM, $r(82) = .43, p < .001$, CRT, $r(82) = .48, p < .001$, and prior academic attainment scores, $r(81) = .38, p < .001$, supporting the use of all three measures as covariates in analyses of Syllogisms scores.

Finally, Belief Bias Index (BBI) scores at Time 1 were also correlated with

RAPM, $r(82) = -.28, p = .011$, CRT, $r(82) = -.26, p = .019$, and prior academic attainment scores, $r(81) = -.28, p = .010$, so the covariates are used in BBI analyses as well.

Although both groups improved their RAPM and CRT scores slightly over the course of the year, neither Group \times Time interaction effect approached significance, $ps > .20$.

Manipulation Check. Change in mathematics test scores was analysed with a 2 (Time: 1 and 2) \times 2 (Group: mathematics, English) Analysis of Variance (ANOVA). There was a significant interaction, $F(1, 80) = 52.91, p < .001, \eta_p^2 = .40$, which suggested that the mathematics group improved to a greater extent (Time 1 $M = 4.82, SD = 1.56$, Time 2 $M = 6.95, SD = 1.94$) than the English group (Time 1 $M = 3.47, SD = 0.95$, Time 2 $M = 3.12, SD = 0.59$, see Figure 5.6). The mathematics group's improvement over time was confirmed by a planned comparison of Time 1 and 2 scores, $t(43) = 7.37, p < .001, d = 1.21$. This suggests that as a group they engaged with and learned from their year of studying mathematics and the quasi-manipulation was successful.

Development of reasoning skills

Conditional Inference analyses.

Endorsement rates. The endorsement rates of each group at Time 1 were analysed with a 2×4 ANOVA with one within-subjects factor: Inference Type (MP, DA, AC, MT), and one between-subjects factor: Group (mathematics, English). There was a significant main effect of Inference Type, $F(3, 231) = 26.29, p < .001$, with MP inferences being most often endorsed ($M = 7.01, SD = 1.27$), followed by MT inferences ($M = 5.94, SD = 1.84$), AC inferences ($M = 5.82, SD = 2.10$), and finally DA inferences ($M = 4.76, SD = 2.54$). There was no interaction between Inference Type and Group, $F(3, 231) < 1$, indicating the two groups responded similarly to the Conditional Inference task at Time 1.

Next, change in endorsement rates of each inference type over time were analysed with a $2 \times 4 \times 2$ ANOVA with two within-subjects factors: Time (start and end of the year) and Inference Type (MP, DA, AC, MT), and one between-subjects factor: Group (mathematics and English literature). This revealed a significant three-way interaction, $F(3, 228) = 7.48, p < .001, \eta_p^2 = .09$, which remained significant after controlling for Time 1 RAPM, Time 1 CRT and prior academic attainment, $F(3, 216) = 5.10, p = .002, \eta_p^2 = .07$ (see Figure 5.7). The means and standard deviations for this interaction are displayed in Table 5.4. At Time 2 the mathematics students endorsed more MP inferences, $t(42) = 2.42, p = .020, d = 0.41$, and fewer DA, $t(42) = 3.98, p < .001, d = -0.67$,

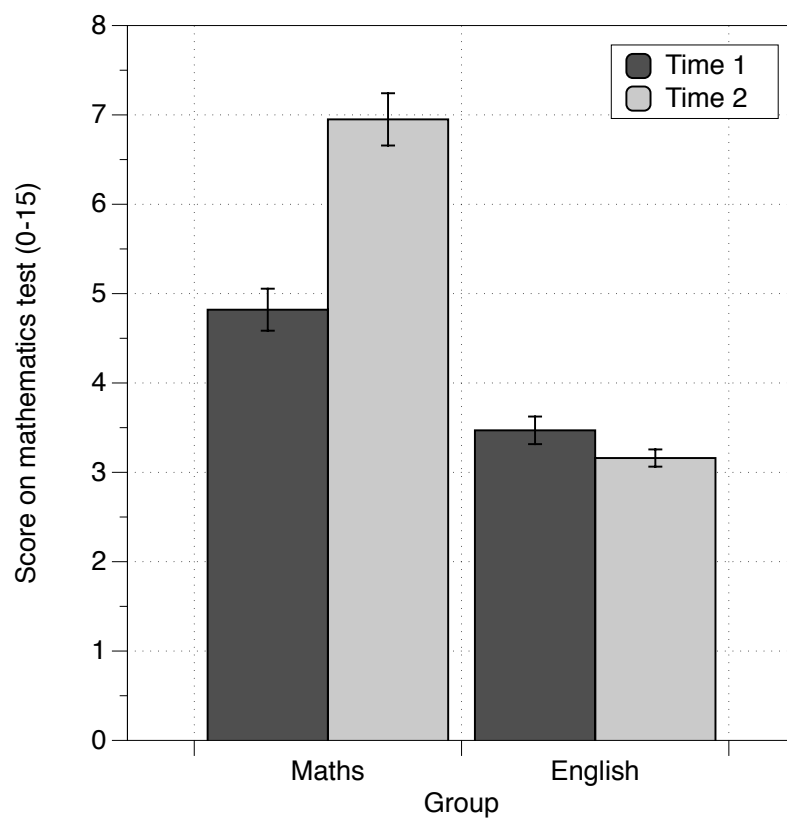


Figure 5.6: Interaction between Group and Time on mathematics test scores (error bars show ± 1 standard error of the mean).

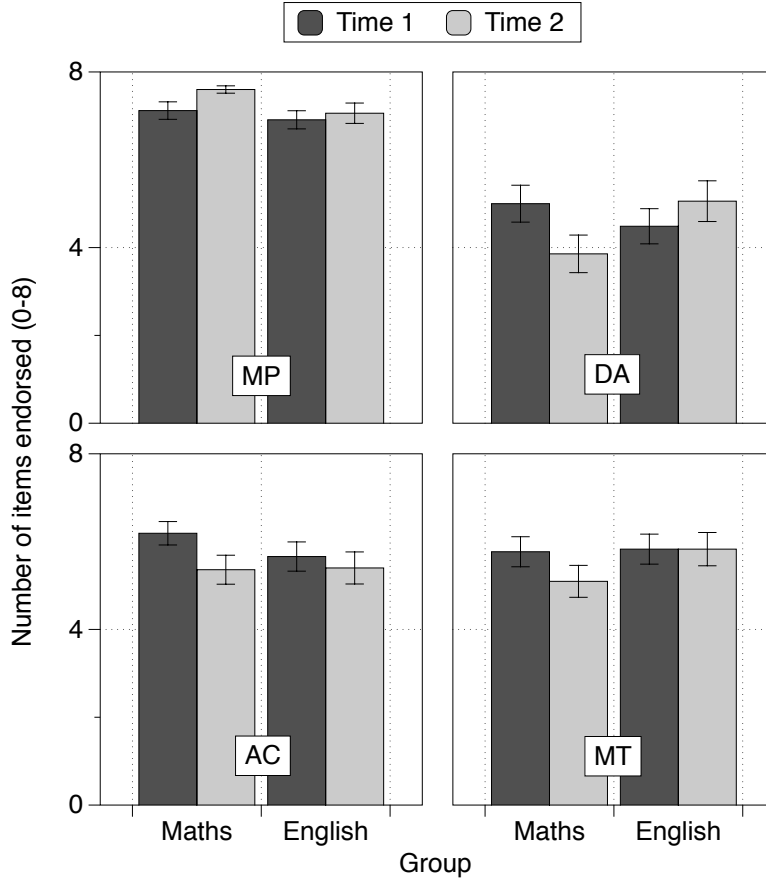


Figure 5.7: Mean endorsement rates for each of the four inferences in each group at Time 1 and Time 2 (error bars show ± 1 standard error of the mean).

AC, $t(42) = 3.06, p = .004, d = -0.47$, and MT inferences, $t(42) = 2.88, p = .006, d = -0.45$ compared to Time 1. In contrast, the English literature group showed no significant differences between Time 1 and Time 2 scores for any inference, although there was a marginally significant increase in the number of DA inferences endorsed, $t(34) = 1.80, p = .082, d = 0.31$.

To summarise, the mathematics group showed an increase in MP endorsement along with a decrease in DA, AC and MT endorsement, which is consistent with a more defective interpretation of the conditional. To investigate this formally, each interpretation index was analysed with a 2×2 ANOVA with one within-subjects factor: Time (start and end of the year) and one between-subjects factor: Group (mathematics, English). The mean index scores for each group at each time point are shown in Figure 5.8.

For the material conditional index (MCI), there was a significant interaction

between Time and Group, $F(1, 76) = 11.86, p = .001, \eta_p^2 = .14$, which remained significant when Time 1 RAPM, Time 1 CRT, and prior academic attainment scores were included as covariates, $p = .007$. The mathematics group became more material, $t(42) = 3.17, p = .003, d = 0.49$, whereas the English literature group did not change, $p = .092, d = -0.17$.

Time and Group also interacted for the biconditional index (BCI), $F(1, 76) = 7.97, p = .006, \eta_p^2 = .10$, although this was only marginally significant when covariates were included, $F(1, 72) = 3.70, p = .058, \eta_p^2 = .05$. The mathematics group became less biconditional, $t(42) = 3.32, p = .002, d = -0.51$, whereas the English literature group did not change, $p = .500, d = 0.07$.

For the defective conditional index (DCI), Time and Group again interacted, $F(1, 76) = 17.65, p < .001, \eta_p^2 = .19$, and this remained significant with covariates, $p = .002$. The mathematics group became more defective, $t(42) = 5.76, p < .001, d = 0.88$, whereas the English literature group did not change, $p = .767, d = -0.03$.

Finally, for the conjunctive conditional index (CCI), Time and Group also interacted, $F(1, 76) = 8.53, p = .005, \eta_p^2 = .10$, which remained significant with covariates, $p = .014$. The mathematics group became more conjunctive, $t(42) = 3.53, p = .001, d = 0.55$, whereas the English literature group did not change, $p = .69, d = -0.06$.

Comparing the effect sizes of these analyses confirms that the change in the mathematics group is best understood as an increased tendency to adopt the defective interpretation of the conditional ($d = 0.88$ compared to $ds < 0.55$ for the other interpretations). Over time the mathematics group became more likely to endorse the MP inference, but less likely to endorse the DA, AC and MT inferences. The English literature group, on the other hand, did not change on

Inference	Group	Time 1	Time 2
MP	Mathematics	7.12(1.29)	7.60(0.54)
	English	6.91(1.22)	7.06(1.37)
DA	Mathematics	5.00(2.72)	3.86(2.77)
	English	4.49(2.37)	5.06(2.74)
AC	Mathematics	5.76(2.22)	5.09(2.36)
	English	5.83(2.02)	5.83(2.24)
MT	Mathematics	6.19(1.73)	5.36(2.14)
	English	5.66(1.97)	5.40(2.16)

Table 5.4: Mean number items endorsed by Inference type, Group and Time point with standard deviations in parentheses.

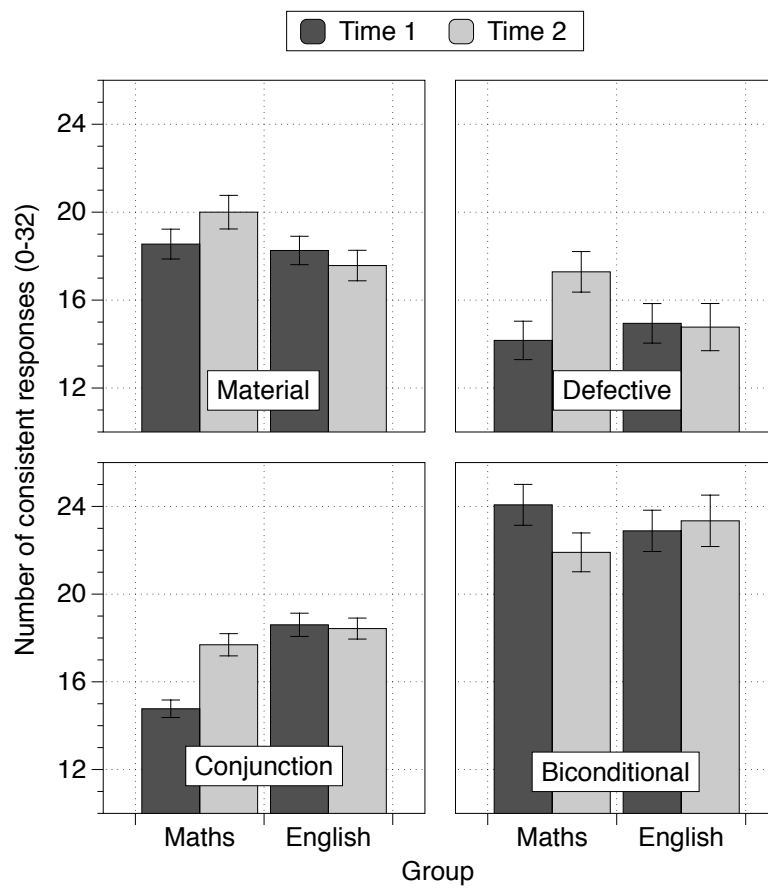


Figure 5.8: Mean interpretation index scores for each group at Time 1 and Time 2 (error bars show ± 1 standard error of the mean).

any of the interpretation indices.

NCI and API scores. NCI scores were subjected to a 2×2 ANOVA with one within-subjects factor: Time (start and end of the year) and one between-subjects factor: Group (mathematics, English literature), and three covariates: prior academic attainment, Time 1 RAPM scores and Time 1 CRT scores. There was a marginally significant interaction between Time and Group, $F(1, 72) = 3.60, p = .062, \eta_p^2 = .05$. Intriguingly, the mathematics students displayed a marginally higher NCI at Time 2 ($M = 2.86, SD = 2.48$) than at Time 1 ($M = 2.05, SD = 2.69$), $t(43) = 1.81, p = .078, d = 0.31$, whilst the English students' scores did not change (Time 1 $M = 2.09, SD = 3.08$, Time 2 $M = 1.91, SD = 3.21$), $t(35) = .33, p = .74, d = -0.06$. This is the opposite of what one would expect given the findings for overall scores – it appears that over the course of the year the mathematics students became *more* biased towards endorsing inferences with negative conclusions.

API scores were also subjected to a 2×2 ANOVA with one within-subjects factor: Time (start and end of the year) and one between-subjects factor: Group (mathematics, English literature), and three covariates: prior academic attainment, Time 1 RAPM scores and Time 1 CRT scores. This did not show a significant interaction, $F < 1$.

Reasoning competency and biases. In the analyses above studying mathematics was found to be associated with an increased defective and material interpretation of the conditional and a marginally significant increase in NCI, compared to studying English literature. This is somewhat counterintuitive. A material or defective interpretation can be said to be an improvement on a biconditional interpretation (see the discussion in Section 2.3), and one would expect that as overall reasoning competence increases there should be less susceptibility to biases. However, it could be the case that a certain level of understanding of the task is required before one is *able* to show systematic biases. If a reasoner has little understanding of the logic of the conditional they may be responding unsystematically to the task (effectively guessing) in which case no systematic bias could be occurring and overall score will be not far above chance level. Once a more sophisticated but not totally consistent understanding of the conditional is reached, the reasoner may be systematic enough to show a higher overall score and be able to show biases, but not competent enough to overcome the biases every time. Of course, when a very high level of understanding is achieved, both overall score and ability to avoid biases should increase. This hypothetical relationship is demonstrated in Figure 5.9.

In order to test the proposed relationship, the data of 656 participants collated from five separate studies (discussed above) was analysed. These data were subjected to a curve fitting estimation, which confirmed that a quadratic



Figure 5.9: Proposed relationship between overall score and a bias index on a reasoning task based on the finding that over time mathematics students showed a higher total score and marginally higher negative conclusion index on the Conditional Inference Task.

curve provided a better fit to the relationship between MCI and NCI on the Conditional Inference task, $R^2 = .09$, $F(2, 653) = 33.61$, $p < .001$, than did a linear relationship, $R^2 = .003$, $F(1, 654) = 2.18$, $p = .14$. The relationship between MCI and NCI is shown in Figure 5.10. This was also the case for the relationship between the DCI and NCI: a quadratic curve provided a better fit to the data, $R^2 = .19$, $F(2, 653) = 74.58$, $p < .001$, than did a linear relationship, $R^2 = .03$, $F(1, 654) = 18.78$, $p < .001$. The relationship between DCI and NCI is shown in Figure 5.11.

However, it is necessarily the case that a high NCI cannot exist when an index score is close to the maximum possible – if the majority of inferences are being categorised consistently then there is not much room for patterns to be found in the inferences categorised inconsistently. Considering this limitation, another way to test the proposed relationship is to look only at data where the index score is below 75% and there is room for biases to occur. Within this range, the hypothesis is that biases will increase alongside the index score as participants become competent enough to think systematically about their answers, whether they be right or wrong. To test this hypothesis, MCI scores were correlated with NCIs for those participants' whose MCI score was less than or equal to 24 (out of 32). This revealed a significant positive relationship, $r(554) = .25$, $p < .001$, as predicted. There was also a significant positive

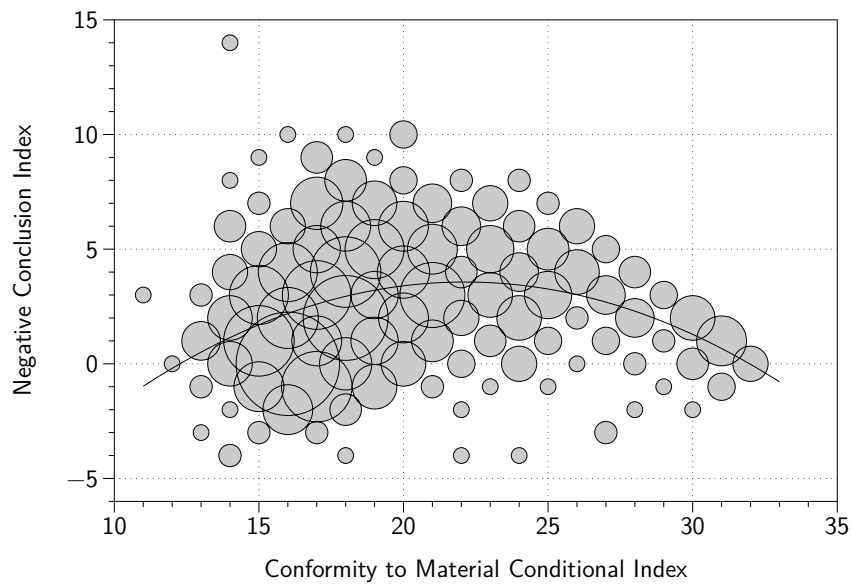


Figure 5.10: Relationship between the Material Conditional Index and Negative Conclusion Index on the Conditional Inference Task when collating data from five studies with a line of best fit. The area of each circle increases with the number of people in that cell.

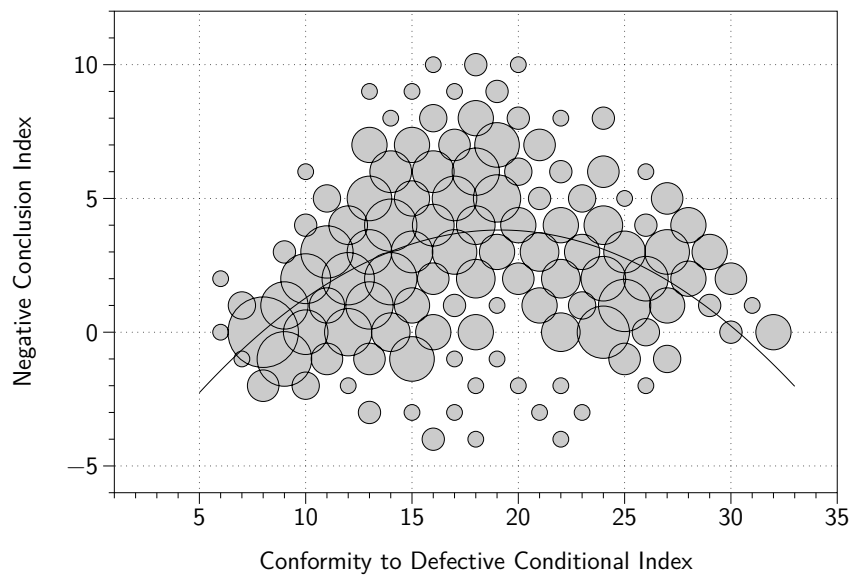


Figure 5.11: Relationship between the Defective Conditional Index and Negative Conclusion Index on the Conditional Inference Task when collating data from five studies with a line of best fit. The area of each circle increases with the number of people in that cell.

relationship between DCI and NCI for participants whose DCI was less than or equal to 24, $r(574) = .32, p < .001$.

The curve fitting estimation and correlation analyses are consistent with the hypothesis that as participants become more successful at thinking about deductions from conditional statements overall, they also become more prone to showing systematic biases in their errors. That is, until they become near-experts. This may be a side-effect of a move from effectively random responding that is absent of any reasoning to a more systematic style of reasoning.

Syllogisms analyses.

Total syllogisms scores. Total syllogisms scores were subjected to a 2×2 ANOVA with one within-subjects factor: Time (start and end of the year) and one between-subjects factor: Group (mathematics, English literature), and three covariates: prior academic attainment, Time 1 RAPM scores and Time 1 CRT scores. This revealed a significant interaction, $F(1, 74) = 4.08, p = .047, \eta_p^2 = .05$. However, the mathematics group's scores did not change between Time 1 ($M = 8.98, SD = 2.26$) and Time 2 ($M = 9.32, SD = 2.25$), $t(44) = 1.06, p = .295, d = 0.15$, and the English literature students' scores marginally decreased between Time 1 ($M = 8.75, SD = 2.13$) and Time 2 ($M = 8.14, SD = 2.21$), $t(36) = 1.90, p = .066, d = -0.28$, see Figure 5.12. The interaction was therefore not in line with the TFD.

Belief Bias Index. Next, the same analysis was conducted with BBI scores. There was no significant interaction between Time and Group, $F(1, 74) < 1$ (see Figure 5.13).

Mechanisms of development

It has been demonstrated that the mathematics group developed a more defective interpretation of the conditional over time. However, they did not significantly improve in syllogistic reasoning ability or avoidance of belief bias (although the English students did decrease in syllogistic reasoning ability, creating an interaction). In this section, Stanovich's (2009a) algorithmic and reflective levels of cognition are evaluated as potential mechanisms of the change in interpretation. RAPM scores were used as a measure of general intelligence (the algorithmic level) and reversed intuitive scores on the CRT were used as a measure of tendency to use Type 2 processing (the reflective level).

Below I present a regression model predicting Time 2 DCI scores from the following blocks of variables:

1. Prior academic attainment, Time 1 DCI scores, Time 1 RAPM scores and Time 1 CRT scores (all covariates);

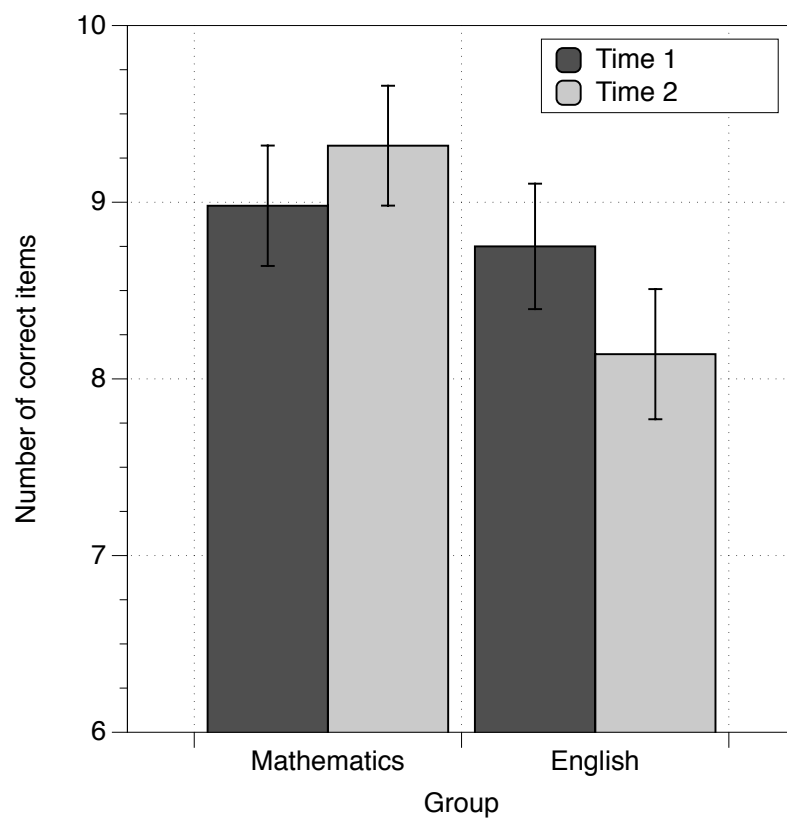


Figure 5.12: Mean syllogisms scores for each group and Time 1 and Time 2 (error bars show ± 1 standard error of the mean).

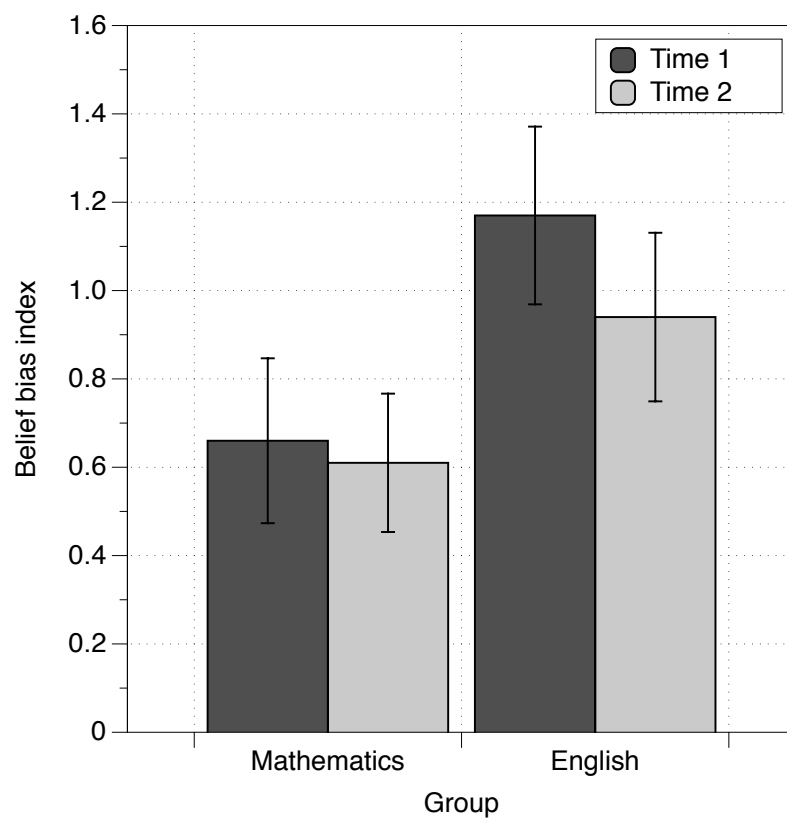


Figure 5.13: Mean belief bias index scores for each group and Time 1 and Time 2 (error bars show ± 1 standard error of the mean).

2. Change-in-RAPM score and Change-in-CRT score, to investigate whether domain-general changes are responsible for Time 2 DCI scores over the whole sample;
3. Group (mathematics or English), to evaluate whether the subject studied predicted Time 2 performance over and above any domain-general changes across the sample;
4. Two interaction terms – Change in RAPM \times Group and Change in CRT \times Group – to investigate whether domain-general changes in the mathematics group specifically are responsible for Time 2 DCI scores.

If it were the case that studying mathematics increases DCI scores by improving the domain-general processing skills, then the interaction terms in the final block should explain a significant amount of variance – one would expect the CRT and RAPM scores to change more in the mathematics group than the English literature group, and for this to predict DCI development. If domain-general changes were responsible for changes in DCI scores independently of subject studied, then the Time 1 or Change scores on RAPM or CRT should be significant predictors. If it were the case that studying mathematics increases DCI scores by a mechanism other than intelligence or thinking disposition, then the Group factor alone should explain a significant amount of variance.

The hierarchical regression model is presented in Table 5.5. The control variables in the first block of predictors accounted for 65.6% of the variance in Time 2 DCI scores, $p < .001$. Change-in-RAPM and Change-in-CRT scores were added in the second block and accounted for an additional significant 3.2% of the variance, $p = .036$. Group (mathematics = 1, English = 0) was entered in the third block and accounted for an additional 2.1% of the variance, $p = .032$. Adding the two interaction terms in the final block accounted for less than 0.5% of additional variance, $p = .581$. In the final model, the only significant predictors were Time 1 DCI ($p < .001$) and Group ($p = .024$).

The regression model does not support the hypothesis that studying mathematics increases conformity to a defective interpretation of the conditional via the domain-general factors of intelligence (at the algorithmic level) or thinking disposition (at the reflective level). Instead, the mechanism of improvement may be domain-specific ‘mindware’, heuristic level factors, or factors at the algorithmic or reflective level of cognition that were not included here, such as executive functions. The control variables included in Block 1 were, as expected, significant predictors. However, it was not the case that mathematics influenced intelligence or thinking disposition in a way that predicted Time 2 DCI: the interaction terms in the final block explained almost zero variance.

Model	R^2	ΔR^2	Predictors	β
1	.66	.66**	Time 1 DCI	.71**
			Time 1 Ravens	.03
			Time 1 CRT	.22**
			Prior academic attainment	.01
2	.69	.03*	Time 1 DCI	.72**
			Time 1 Ravens	.13
			Time 1 CRT	.18
			Prior academic attainment	.01
			Change-in-RAPM	.20*
			Change-in-CRT	.06
3	.71	.02*	Time 1 DCI	.76**
			Time 1 Ravens	.06
			Time 1 CRT	.13
			Prior academic attainment	-.01
			Change-in-RAPM	.15 [†]
			Change-in-CRT	.04
			Group (mathematics = 1, English = 0)	.17*
4	.71	.01	Time 1 DCI	.75**
			Time 1 Ravens	.07
			Time 1 CRT	.12
			Prior academic attainment	-.01
			Change-in-RAPM	.14
			Change-in-CRT	.09
			Group (mathematics = 1, English = 0)	.20*
			RAPM Change \times Group	.02
			CRT Change \times Group	-.09

Table 5.5: A hierarchical regression analysis predicting Time 2 Defective Conditional Index scores. [†] $p < .1$, * $p < .05$, ** $p < .001$.

Influence of other science subjects

At AS level, students tend to study four subjects, and those studying mathematics are often studying science subjects as well. In England in 2009, for example, 63% of students studying post-compulsory mathematics also studied one or more of chemistry, biology and physics (Royal Society, 2011). To investigate whether mathematics was unique as a formal discipline that predicted development in Conditional Inference, the number of core science subjects each participant was studying (physics, chemistry, biology) was correlated with change in DCI score within the mathematics group. This revealed a non-significant

relationship, Spearman's $r(43) = .16, p = .316$, which remained non-significant after controlling for participants' Time 1 RAPM, CRT scores, and prior academic attainment scores, $pr(37) = .22, p = .188$. In the current data set there is no evidence that studying non-mathematics science subjects is responsible for the relationship between mathematics and change in DCI.

5.4 Discussion

This chapter aimed to shed light upon two questions that are central to this thesis and the TFD: (a) is studying mathematics at advanced levels associated with improvement in reasoning skills? and (b) if there is such improvement, what might its mechanisms be? These questions were addressed with a longitudinal study that followed the development of conditional and syllogistic reasoning ability in AS level mathematics and English literature students. The results have shown that (a) the mathematics students changed in conditional reasoning behaviour to a greater extent than the English literature students, (b) the change was best characterised as the mathematics students becoming more defective in their interpretation of the conditional, and (c) that the mechanism for the development did not appear to be domain-general.

5.4.1 Development of reasoning skills

It was found here that mathematics students' conditional reasoning behaviour became more in line with the defective conditional interpretation over time, whereas the English literature students' reasoning behaviour did not change. Inglis and Simpson (2009a) found that, compared to intelligence-matched comparison undergraduates, incoming mathematics undergraduates reasoned more normatively on the Conditional Inference Task but that they did not change over a year of mathematics study. The authors suggested that the initial difference may have been due to one of three possibilities: post-compulsory but pre-university study of mathematics developing reasoning skills; filtering of more material reasoners into the study of mathematics; or between-group differences unrelated to intelligence, such as in thinking disposition. The findings are consistent with the first possibility, that post-compulsory but pre-university study of mathematics develops conditional reasoning skills. At the start of post-compulsory education, the students studying mathematics did not differ from non-mathematics students on the Conditional Inference Task, but they did after a year of study. This change was not due to between-groups differences in initial or changed thinking disposition or intelligence scores.

The mathematics students' reasoning did not become entirely more norm-

ative as we might expect from an all-round improved understanding of logic. There was increased rejection of DA, AC and MT inferences and increased acceptance of MP inferences, reflecting a move towards the so-called ‘defective’ interpretation of the conditional. This is not entirely surprising given the nature of mathematics: Houston (2009) argued that most mathematical statements are of the form ‘if statement A is true, then statement B is true’, even if they are heavily disguised (p. 63). He also argued that in mathematics A is assumed to be true, even if it is clearly not true, and the truth or falsity of B is then deduced. Since the AS level curriculum does not include any explicit reference to conditional logic or the material conditional, it is plausible that exposure to implicit ‘if then’ statements, where the antecedent is assumed to be true, could induce a defective interpretation of the conditional, where false antecedent cases are considered irrelevant. In line with this hypothesis, Hoyles and Kuchemann (2002) argued that the defective interpretation is actually more appropriate for mathematics classrooms than the material interpretation for the same reasons proposed by Houston. Furthermore, Inglis and Simpson (2009b) found that a group of undergraduate mathematics students, who had of course been very successful at A level mathematics, tended to have a more defective than material interpretation of conditionals.

It seems somewhat surprising that studying mathematics was associated with improved conditional reasoning skills given that Cheng et al. (1986) found no improvement in conditional reasoning even after their participants studied a course in formal logic. One possible account for this discrepancy is that the measure used by Cheng et al. (1986) was not suitable for detecting improvement. They used four Wason Selection Tasks, three of which were contextualised, to measure conditional reasoning ability. Since then it has been suggested that Wason Selection Tasks, in particular contextualised ones, may not measure conditional reasoning at all (Sperber et al., 1995, 2002) and so it is possible that Cheng et al. (1986) were simply not measuring improvement that did actually occur.

It is also possible that if Cheng et al. (1986) had looked specifically at a defective conditional index they may have found a change – in the current study there was a shift towards a material interpretation, but a smaller one than the shift towards a defective interpretation. If Cheng et al.’s participants did become more defective over time it would not have been reflected by performance on a Selection Task – it would encourage participants to select only the true antecedent card and no others, which would have been considered wrong in Cheng et al.’s analysis, along with a biconditional interpretation. Incidentally, this was the pattern of responding found in Inglis and Simpson’s (2004) study of undergraduate mathematics students. Compared to a control group, Inglis

and Simpson's (2004) mathematics students gave more true antecedent (p) only answers and fewer true antecedent and true consequent (p and q) answers.

5.4.2 Competency and biases in the Conditional Inference Task

An additional analysis using a large set of data collated from several studies tested the hypothesis that the negative conclusion index increases as the material and defective interpretations increase, up to the point of near-complete consistency within inference type. In support of this, the relationships between MCI and NCI and between DCI and NCI were better fit by quadratic curves than by linear functions. This relationship suggests that as people become more systematic in their reasoning, they also become more, not less, susceptible to common biases. When people are responding unsystematically, either because they do not understand the conditional or because they are not engaged with the task, it is not possible for them to show negative conclusion bias because it requires some understanding and consistency in responses to similar deductions. Those people who do understand the conditional and respond with some consistency to similar deductions are at least *able* to show negative conclusion bias, and may or may not be able to overcome it. Finally, those who have a consistent understanding of the conditional are able to overcome negative conclusion bias and answer all or most items in line with their primary interpretation. In sum, this pattern may reflect a developmental trajectory from unsystematic responding to more systematic but biased responding to completely systematic reasoning.

5.4.3 Limitations

The students who were studying mathematics in the sample also tended to be studying other core science subjects (physics, biology, chemistry). It was not possible to separate out the potential relationships between change in conditional reasoning behaviour and the study of mathematics, physics, chemistry and biology in my sample, due to the small numbers of students studying only one of the subjects. However, there was no evidence that the more science subjects a person was studying, the more their reasoning behaviour changed, so although this confound is not ideal it does not appear to be a severely limiting factor. To investigate the potential confound issue, a much larger sample size would be required to ensure that enough participants were studying one science subject alone. However, the study reported in Chapter 6 investigated the role of studying mathematics without science subjects in undergraduate students and suggested that the relationship found here still holds.

A substantial drawback to the design of this study is that it was a quasi-experiment as opposed to a true experiment. Participants chose their AS level subjects before the study began – it would not have been practical or ethical to randomly assign students to studying mathematics or English literature. Although there were no differences between groups in reasoning behaviour at Time 1, the quasi-experimental design means that it is not possible to establish causal relationships between subject studied and improvement in reasoning skills. As discussed in Chapter 3, non-random assignment to conditions means that not all confounding variables can be prevented from influencing the relationships studied. Intelligence and thinking disposition, two likely confounds in the relationship between subject studied and reasoning skills, were measured and statistically controlled for, but statistical control is not as effective as random assignment to conditions (Christensen, 2000) and, more importantly, there may have been other confounding variables that were not considered. The most effective way to study causation in the TFD would be to randomly assign participants to courses in mathematics or non-mathematical subjects and track their development in reasoning over an extended period of time. However, this is obviously neither practical nor ethical in high stakes assessment.

In terms of relevance to educational policy, the quasi-experimental design of this study also means that the results may not apply to curricula where it is compulsory to study mathematics until the age of 18. The participants in this study had chosen to study mathematics and thus they likely enjoyed it and were engaged with the course. Where students are required to study mathematics until the age of 18 this may not be the case and it might be that only those who do enjoy and engage with the course see the benefits to reasoning that were found. A useful direction for future research would be to compare the reasoning development of students studying mathematics in curriculums where it is and is not compulsory.

5.4.4 Revised status of the Theory of Formal Discipline

The Theory of Formal Discipline suggests that studying mathematics improves one's ability to reason. Although the TFD makes big claims about the relationship between mathematics and reasoning, it is silent on the issue of what the mechanism for the relationship might be. To date, the TFD has been assumed to be true with minimal scrutiny. Here, some evidence was presented that is partly consistent with the TFD, and furthermore, some potential mechanisms for the improvement in reasoning skills were investigated.

In the study presented here it was found that mathematics students' conditional reasoning behaviour changed to a greater extent than that of non-

mathematics students’, but that mathematics students did not show any change in their syllogistic reasoning nor susceptibility to belief bias. I also found no evidence that the mechanism for change in conditional reasoning skills was a change to the algorithmic or reflective level of cognition, as specified in Stanovich’s (2009a) model. Here, RAPM was the only measure of algorithmic processing, but executive function is a separable aspect of the algorithmic level. Executive function refers to the efficiency of carrying out conscious information processing – e.g. updating information in working memory, switching between tasks, and inhibiting irrelevant information or responses, and it has been shown to be separate from fluid intelligence (Ardila, Pineda & Rosselli, 2000; Arffa, 2007; Friedman et al., 2006). The relationship between executive functions and conditional reasoning behaviour is investigated in Chapter 9.

Another possibility is that mathematics students’ defective responding stems from the heuristic level of cognition. This is investigated in Chapter 8, where mathematics and non-mathematics students’ conditional reasoning behaviour is measured under a strict time limit as well as under no time limit so see how behaviour changes when processing is restricted to the heuristic level.

Chapter 7 presents evidence that mathematics students’ ability with conditional reasoning is dependent on the linguistic form of the conditional statement, supporting the hypothesis that exposure to ‘if then’ statements is responsible for the change as opposed to a general understanding of conditional logic. Next, however, a longitudinal study investigating changes in conditional reasoning behaviour in undergraduate students is presented.

5.4.5 Summary of novel findings

1. Mathematics students’ conditional reasoning behaviour conformed more to the defective and material interpretations and less to the biconditional interpretation of conditional statements after a year of AS level study, compared to English literature students.
2. There was no evidence that mathematics students’ syllogistic reasoning behaviour nor susceptibility to belief bias changed over time.
3. The mechanism of the change in mathematics students’ conditional reasoning behaviour did not appear to be baseline or changed scores on intelligence or thinking dispositions measures.
4. The largest change in mathematics students’ conditional reasoning behaviour was in the form of greater adoption of the defective conditional - this may reflect practice with implicit ‘if then’ statements in mathematics where students are expected to assume p and reason about q .

Chapter 6

The development of reasoning skills in undergraduate mathematics students

6.1 Introduction

6.1.1 Testing the Theory of Formal Discipline

Chapter 5 presented a study that investigated the development of reasoning skills in AS level mathematics and English students. It was found that the mathematics students became increasingly defective in their reasoning behaviour, i.e. they became more likely to reject DA, AC and MT inferences, and more likely to accept MP inferences. It was also found that mathematics students did not improve in reasoning with thematic syllogisms. The discrepancy between the abstract conditional reasoning improvement and lack of thematic syllogisms improvement could either be due to the context/abstract aspect or to the conditionals/syllogisms aspect. This will be clarified in the study reported below.

The aim of the current chapter is to investigate the development of reasoning skills in undergraduate mathematics students. Because the study and its motivations are very similar to the AS study, much of the relevant background and description of materials has already appeared in Chapter 5. To avoid repetition only a brief overview of the relevant background and justifications of the study are provided below.

The little amount of research that has previously investigated the development of reasoning skills in association with the study of mathematics has all focused on undergraduate students. Lehman and Nisbett (1990) tested US undergraduates on various types of reasoning at the beginning and end of their four years of study and found that the natural science students (who took the most mathematics modules) reasoned more in line with the material conditional at the end of their degrees. Furthermore, the number of mathematics modules taken was correlated with the extent of change. In the UK, Inglis and Simpson (2008) found that on entry to university, mathematics undergraduates reasoned more normatively than a comparison group on a Conditional Inference task. However, in a follow up study Inglis and Simpson (2009a) found that although mathematics students again outperformed comparison undergraduates on entry to university, their reasoning did not change over a year of study. The authors suggested that studying mathematics at A level may have led to the initial difference between groups, and this was supported by the study presented in Chapter 5 of this thesis. However, the authors did not investigate differences or changes in a defective interpretation of the conditional, and so it is possible that there were changes that went undetected. Nevertheless, there is a discrepancy between the findings of Lehman and Nisbett (1990) and Inglis and Simpson (2009a) for the material interpretation of the conditional, and the current chapter will add evidence that could clarify the discrepancy.

Aside from contributing to the limited base of evidence relating to the development of reasoning skills in undergraduate mathematics students, the study presented below will address two issues that arose in the AS study presented in Chapter 5.

Firstly, in the AS sample, science and mathematics were confounded so it was not possible to isolate the effect of studying mathematics on reasoning. In the UK, students tend to study only one subject at degree level, and so the study presented here allowed the effect of mathematics to be investigated in isolation. This issue may also be a source of the discrepancy described above – Lehman and Nisbett (1990) found their evidence for a relationship between mathematics and conditional reasoning in students at a US university, where it is common for students to study different subjects as their major and minor degree components. There was no evidence of a relationship between studying science subjects and changes in the DCI in the AS level students, but the mix of subjects studied by Lehman and Nisbett’s (1990) participants could potentially be a reason for their finding a change where Inglis and Simpson (2009a) did not.

Secondly, in the AS study mathematics students only changed in abstract conditional reasoning, not in thematic syllogistic reasoning. The lack of improvement in the syllogisms could be due to the use of context – perhaps studying

mathematics only provides an advantage for thinking about abstract problems – or it could be due to the syllogisms – perhaps studying mathematics only provides an advantage for thinking about conditional statements (the increase in the defective interpretation of the conditional is supportive of the latter explanation). This issue will be clarified in the study presented here. Instead of completing a thematic syllogisms task, participants will complete a thematic conditional inference task which is very similar in form to the abstract conditional reasoning task. If mathematics undergraduates change on the abstract version but not the thematic version it would suggest that context interferes with their reasoning. If they change on both the abstract and thematic versions (in particular, if their reasoning becomes more defective in both), it could indicate that studying mathematics only changes interpretations of ‘if’ and that that’s why there was no change on the Syllogisms task in the AS level students. However, an alternative possibility in the latter case would be that reasoning in context is a skill that comes later in mathematical study than abstract reasoning – at a point in between AS level and the first year of an undergraduate degree.

A further aim of this study, as with the AS study, is to identify potential mechanisms for any improvement that occurs. As before, measures of intelligence at the algorithmic level and reflective thinking disposition at the reflective level of cognition in Stanovich’s (2009a) model will be included for this purpose.

6.1.2 Summary

In sum, there are three research questions that will be addressed in this chapter: (a) is studying mathematics (and not science) at undergraduate level associated with changes in abstract conditional reasoning skills, (b) is studying mathematics (and not science) at undergraduate level associated with changes in thematic conditional reasoning skills and (c) can any improvement found be attributed to changes in intelligence or reflective thinking disposition? Based on the findings from the AS level study, it can be hypothesised that:

1. Studying mathematics at undergraduate level will be associated with more defective, and to a lesser extent, more material abstract conditional reasoning,
2. The changes in abstract conditional reasoning will be predicted by the Group factor over and above intelligence and/or thinking disposition.

It is unclear whether the mathematics students will improve in thematic conditional reasoning, but if they do, the same hypotheses would likely apply to the improvement.

6.2 Method

6.2.1 Design

Undergraduate mathematics and psychology students took part at the beginning and end of their first year of study. Participants had already self-selected into degree courses and so the study took a quasi-experimental design. The same set of tasks was administered at both time points to allow a longitudinal investigation of development.

6.2.2 Participants

Eighty-three mathematics students and 64 psychology students took part at Time 1. The mathematics group consisted only of students who were studying the three year single honours mathematics course ($N = 66$) or the four year mathematics undergraduate masters course (MMath, $N = 17$). All participants were first year students at Loughborough University and took part on a voluntary unpaid basis.

6.2.3 Mathematics Syllabus

In the first year of the single honours mathematics and MMath degrees students take the following compulsory modules: Calculus, Linear algebra, Geometry, Vectors and complex numbers, Mathematical thinking, Introduction to applied mathematics, Computer applications in mathematics, Sequences and series, Differential equations and Introductory probability and statistics. Contrary to the A level syllabus discussed in Chapter 5, the undergraduate module ‘Mathematical thinking’ covers various aspects of logic, including conditional statements and truth tables.

6.2.4 Measures

Abstract Conditional Inference. Participants completed the same Conditional Inference task (Evans et al., 1995) used in Chapter 5, consisting of 32 abstract items of four inference types: modus ponens (MP), denial of the antecedent (DA), affirmation of the consequent (AC) and modus tollens (MT). Half of the items used explicit negations (e.g. “not 5”) and half used implicit negations (e.g. “not 5” represented as, for example, 6). Four interpretation indices were taken: MCI, DCI, BCI and CCI. The instructions given were identical to those used by Evans et al. (1995). An example item is shown in Figure 6.1.

Thematic Conditional Inference. A thematic version of the conditional inference task (see Appendix F) was created based on Evans, Handley, Neilens

If the letter is S then the number is 6	If the letter is M then the number is 4
The number is not 6	The letter is not M
Conclusion: The letter is not S	Conclusion: The number is not 4
<input type="radio"/> Yes	<input type="radio"/> Yes
<input type="radio"/> No	<input type="radio"/> No
a) Modus tollens	b) Denial of the antecedent

Figure 6.1: Example item from the Abstract Conditional Inference task.

and Over's (2010) task. Participants decided whether a conclusion necessarily followed from a rule and a premise in 16 items. There were four MP, four MT, four DA and four AC items. Two of each inference type were presented in believable context and two in unbelievable context. An example unbelievable MT item is shown in Figure 6.2.

All negations were represented explicitly: the lack of implicit negations is the reason for there being 16 rather than 32 items. The measures taken were MCI, DCI, BCI, CCI, and a belief bias index (BBI), as with the syllogisms task in Chapter 5. The BBI was calculated as number of consistent items accepted (believable/valid, unbelievable/invalid) minus number of inconsistent items accepted (believable/invalid, unbelievable/valid). The possible range for BBI scores was -8 to +8, with positive scores indicating a degree of belief bias (being more persuaded by belief than validity in the inconsistent items).

Raven's Advanced Progressive Matrices. An 18 item subset of Raven's Advanced Progressive Matrices (RAPM) with a 15 minute time limit (Sá et al., 1999) was used as a measure of general intelligence (at the algorithmic level of processing, Stanovich, 2009a, see Appendix A).

Assume the following is true:
 If third world debt is cancelled then poverty will worsen.
Given that the following premise is also true:
 Third world poverty does not worsen.
Is it necessary that:
 Third world debt is not cancelled.
☐ Yes
☐ No

Figure 6.2: Example item from the Thematic Conditional Inference task.

Cognitive Reflection Test. Number of intuitive responses given to the three-item CRT (Frederick, 2005) was used as a performance measure of the tendency to use Type 2 processing (at the reflective level, Toplak et al., 2011; Stanovich, 2009a). Scores were reversed so that higher scores represented more normative performance, in line with the other measures. The questions were randomly intermixed with three simple mathematical word problems of a similar length from the Woodcock-Johnson III Applied Problems subtest as described in Chapter 5.

Mathematics Manipulation Check. To confirm that the mathematics group learnt mathematics during the year and that the psychology students did not, a mathematics test was included. This consisted of 11 questions, seven of which were taken from the Woodcock-Johnson III Calculation subtest, two of which were the most difficult questions on the Loughborough University diagnostic test for new mathematics undergraduates based on performance in 2008 and 2009, and a final two of which were based on the first year mathematics degree syllabus. The full task is presented in Appendix G.

6.2.5 Procedure

Participants took part during lectures: the mathematics students in one group and the psychology students in another. RAPM was always completed first with a 15 minute time limit. The rest of the tasks followed in one of four Latin square counterbalanced orders to which participants were randomly assigned:

1. Mathematics test, Abstract Conditional Inference task, Thematic Conditional Inference task, CRT
2. CRT, Mathematics test, Abstract Conditional Inference task, Thematic Conditional Inference task
3. Thematic Conditional Inference task, CRT, Mathematics test, Abstract Conditional Inference task
4. Abstract Conditional Inference task, Thematic Conditional Inference task, CRT, Mathematics test

6.3 Results

Of the 147 participants who took part at Time 1, 59 mathematics and 30 psychology students took part again at Time 2 and were included in the analysis. The high drop out rate may have been due to the second testing sessions taking place shortly before the exam period, particularly so in the psychology group.

Task	Mathematics	Psychology
RAPM	59	30
CRT	47	29
Abstract Conditional Inference	53	27
Thematic Conditional Inference	49	28
Mathematics test	51	30

Table 6.1: Number of participants in each group who completed each task at Time 2.

Furthermore, not all participants completed all tasks. Table 6.1 shows the number of participants in each group who completed each task at Time 2. In all analyses presented below, missing data is excluded pairwise to maximise statistical power.

The participants who returned at Time 2 had significantly higher scores on Time 1 RAPM, $t(145) = 2.34, p = .021$, Time 1 mathematics test, $t(138) = 2.49, p = .014$, Time 1 Abstract MCI scores, $t(114) = 2.78, p = .006$, and Time 1 Thematic MCI scores, $t(116) = 2.19, p = .031$ than those who did not return. The sample was therefore biased towards the more able students, but there will always be some degree of bias when the sample is self-selected. There was no difference between those who returned and those who did not in Time 1 CRT scores, $t(145) = .17, p = .862$. Importantly, there were no significant interactions between Group and Return for RAPM, mathematics test or Thematic MCI scores (all $ps > .250$), indicating the bias was not significantly greater in one group or the other. However, there was a significant interaction between Group and Return for Time 1 Abstract MCI scores, $F(1, 112) = 4.13, p = .044, \eta_p^2 = .04$. In the mathematics group, those who returned had scored marginally higher than those who did not return, $t(65) = 1.89, p = .063$, whereas there was no difference between returnees and non-returnees in the psychology group, $t(47) = .99, p = .327$. The mean scores on each task at Time 1 for those who returned at Time 2 and those who did not are displayed in Table 6.2.

The results are reported in two sections: (i) preliminary analyses and (ii) development of Abstract and Thematic conditional interpretation scores.

6.3.1 Preliminary analyses

Covariates. The mathematics group scored significantly higher on the RAPM at Time 1 ($M = 11.1, SD = 2.97$) than the psychology group ($M = 8.63, SD = 3.40$), $t(87) = 3.56, p = .001, d = 0.77$. The mathematics group also scored

significantly higher on the CRT at Time 1 (reversed number of intuitive answers, $M = 2.12, SD = 0.86$) than the psychology group ($M = 1.30, SD = 1.11$), $U(72) = 327.50, z = -2.98, p = .003, r = .35$.

Abstract MCI scores at Time 1 were significantly correlated with Time 1 RAPM scores, $r(63) = .33, p = .009$, and Time 1 CRT scores, $r(63) = .26, p = .038$. Consequently, both Time 1 RAPM and CRT scores are used as covariates in subsequent analyses of Abstract Conditional Interpretation indices.

Time 1 Thematic MCI scores were significantly correlated with both Time 1 RAPM, $r(66) = .36, p = .003$, and Time 1 CRT scores, $r(66) = .43, p < .001$, supporting the use of Time 1 RAPM and CRT scores as covariates in analyses of Thematic Conditional Interpretation scores.

Finally, BBI scores at Time 1 were significantly correlated with Time 1 RAPM scores, $r(66) = -.26, p = .038$, and marginally significantly correlated with Time 1 CRT scores, $r(66) = -.24, p = .058$. The BBI analysis is reported below with Time 1 RAPM scores as a covariate and both with and without Time 1 CRT scores as a covariate.

Although both groups improved their RAPM scores slightly over the course of the year, the Group \times Time interaction effect did not approach significance, $F(1, 80) < 1$. Again, both groups improved their CRT scores slightly over the year but the Group \times Time interaction effect did not approach significance, $F(1, 62) = 1.95, p = .167, \eta_p^2 = .031$ (shown in Figure 6.3).

Manipulation Check. Change in mathematics test scores were analysed with a 2×2 ANOVA with one within-subjects factor: Time (start and end of the year) and one between-subjects factor: Group (mathematics, psychology). There was a significant interaction, $F(1, 76) = 8.24, p = .005$, which suggested that the mathematics group improved to a greater extent than the psychology group, shown in Figure 6.4. The mathematics group's improvement over time was confirmed by a planned comparison of Time 1 and 2 scores, $t(48) = 3.70, p = .001, d = .70$. This suggests that as a group they engaged with and learned from

Task (maximum)	Returned at Time 2	Absent at Time 2
RAPM (18)	10.27 (3.29)	8.91 (3.64)
CRT (3)	2.06 (1.02)	2.07 (1.03)
Abstract MCI (32)	20.58 (4.51)	18.43 (3.43)
Thematic MCI (16)	11.16 (2.91)	9.93 (3.02)
Mathematics test (11)	6.23 (2.94)	4.98 (2.82)

Table 6.2: Mean scores on each task at Time 1 for those who did and did not return to take part at Time 2 with standard deviations in parentheses.

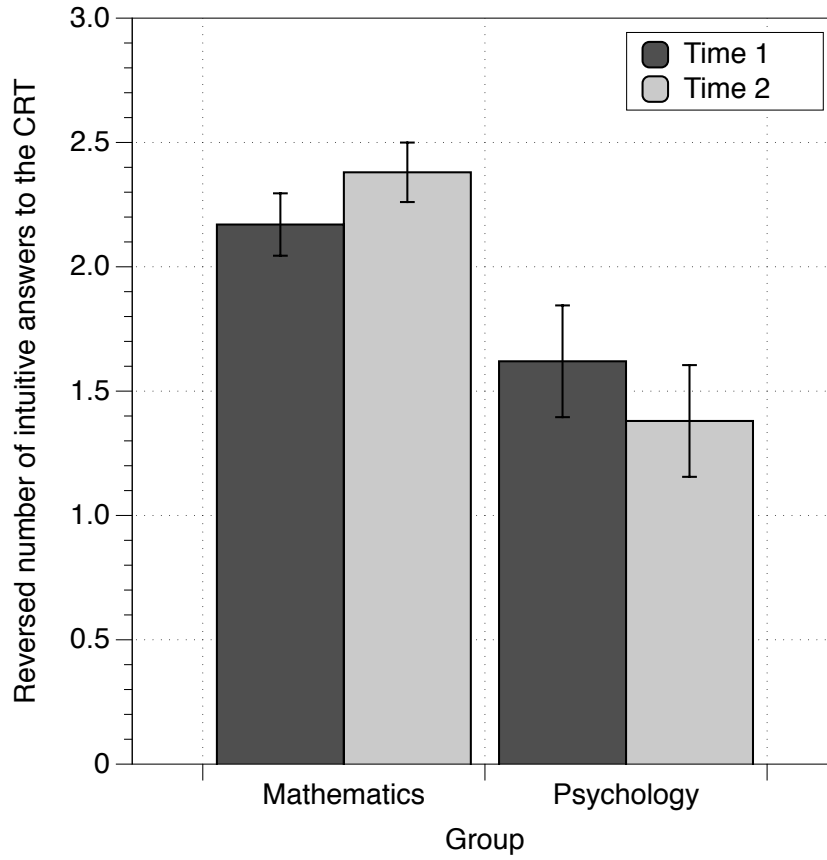


Figure 6.3: Interaction between Group and Time on the (reversed) number of intuitive answers given to the CRT (where higher scores represent better performance, error bars show ± 1 standard error of the mean).

their year of studying mathematics and the quasi-manipulation was successful.

6.3.2 Development of reasoning skills

Abstract Conditional Inference.

Endorsement rates. Endorsement rates of each inference type were analysed with a $2 \times 4 \times 2$ ANOVA with two within-subjects factors: Time (start and end of the year) and Inference Type (MP, DA, AC, MT), one between-subjects factor: Group (mathematics and psychology), and two covariates: Time 1 RAPM and Time 1 CRT. This revealed a marginally significant three-way interaction, $F(3, 177) = 2.58, p = .055, \eta_p^2 = .04$, (see Figure 6.5). The means and standard deviations for this interaction are displayed in Table 6.3. At Time 2 the mathematics students endorsed more MP inferences, $t(43) = 2.33, p =$

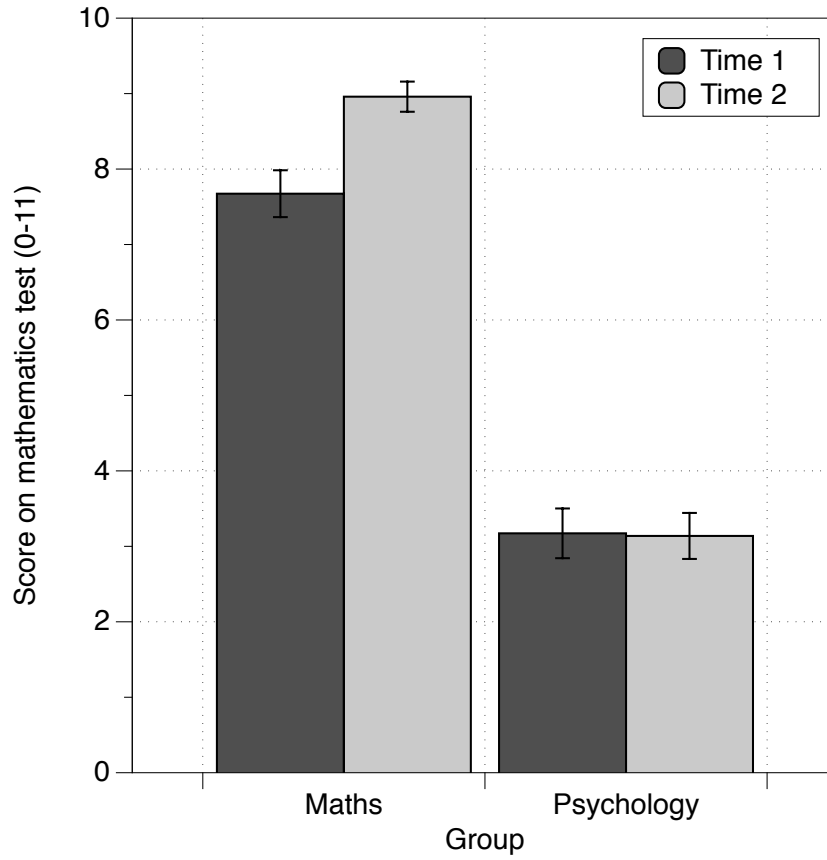


Figure 6.4: Interaction between Group and Time on mathematics test scores (error bars show ± 1 standard error of the mean).

.025, $d = 0.50$, and fewer DA, $t(43) = 3.37, p = .002, d = -0.40$ and AC inferences, $t(43) = 3.30, p = .002, d = -0.42$ compared to Time 1. Mathematics students did not change in their endorsement rate of MT inferences, $t(43) = 1.19, p = .243, d = -0.16$. In contrast, the psychology group showed no significant differences between Time 1 and Time 2 scores for any inference (all $ps > .160$), although there was a marginally significant decrease in the number of MT inferences endorsed, $t(18) = 1.97, p = .065, d = -0.54$.

An increase in MP endorsement along with a decrease in DA and AC endorsement is consistent with a more material or defective interpretation of the conditional. To investigate this formally, each interpretation index was analysed with a 2×2 ANOVA with one within-subjects factor: Time (start and end of the year) and one between-subjects factor: Group (mathematics, psychology). The mean index scores for each group at each time point are shown in Figure 6.6.

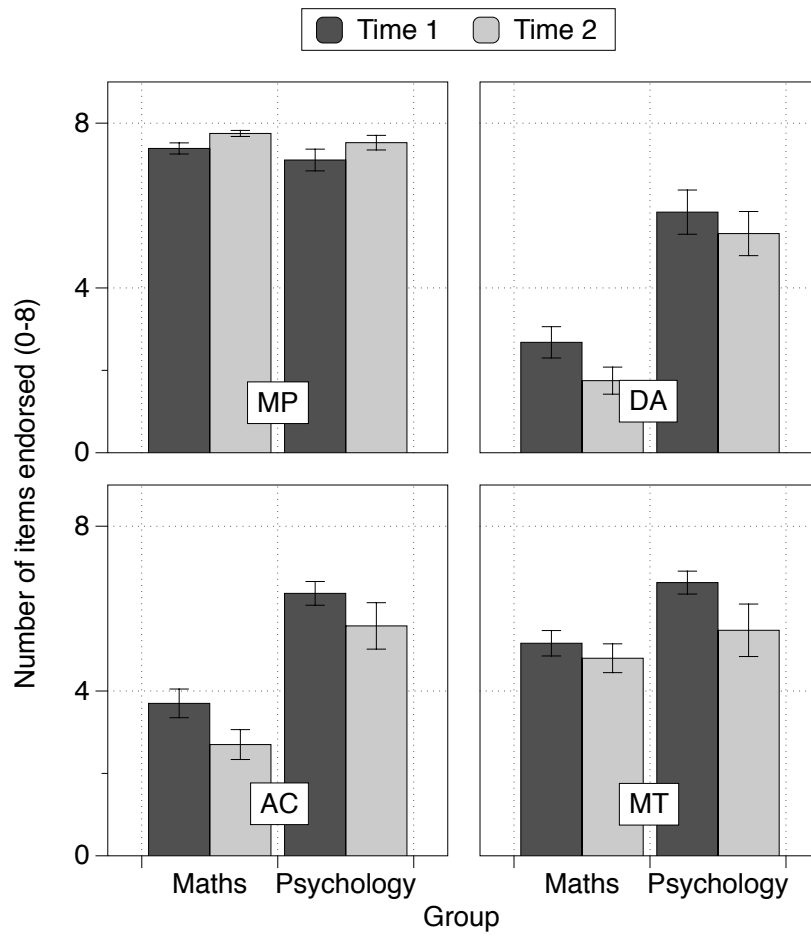


Figure 6.5: Mean endorsement rates for each of the four inferences in each group at Time 1 and Time 2 on the Abstract Conditional Inference Task (error bars show ± 1 standard error of the mean).

Inference	Group	Time 1	Time 2
MP	Mathematics	7.39(0.89)	7.75(0.49)
	Psychology	7.11(1.15)	7.53(0.77)
DA	Mathematics	2.68(2.52)	1.75(2.18)
	Psychology	5.84(2.34)	5.32(2.33)
AC	Mathematics	3.70(2.31)	2.70(2.41)
	Psychology	6.37(1.26)	5.58(2.46)
MT	Mathematics	5.16(2.05)	4.80(2.32)
	Psychology	6.63(1.21)	5.47(2.78)

Table 6.3: Mean number items endorsed on the Abstract Conditional Inference Task by Inference type, Group and Time point with standard deviations in parentheses.

Interpretations. To evaluate changes in each group's inclination towards the four interpretations of the abstract conditional, four 2×2 ANOVAs were carried out, one for each interpretation, each with one within-subjects factor: Time (start and end of the year), one between-subjects factor: Group (mathematics, psychology) and two covariates: RAPM and CRT. The means for each group's interpretation indices at each time point are presented in Table 6.4 and Figure 6.6.

For the MCI, there was no significant interaction, $F(1, 59) = 2.20, p = .143, \eta_p^2 = .04$, but there was a trend in the direction predicted by the TFD. Paired samples t-tests demonstrated that the mathematics group's MCI was significantly higher at Time 2 ($M=24.09, SD=4.10$) than at Time 1 ($M=22.16, SD=4.24$), $t(43) = 3.68, p = .001, d = 0.55$, whereas the psychology group's MCI did not change over time, $t(18) = .70, p = .495, d = 0.18$ (Time 1: $M=17.53, SD=3.61$, Time 2: $M=18.11, SD=2.69$). Furthermore, the mathematics group's MCI was significantly higher than the psychology group's at Time

Group	Time	Material	Defective	Biconditional	Conjunctive
Maths	Time 1	22.16 (4.24)	19.84 (5.58)	18.93 (5.67)	19.25 (2.92)
	Time 2	24.09 (4.10)	22.50 (5.98)	17.00 (5.61)	19.91 (2.76)
Control	Time 1	17.53 (3.61)	12.26 (4.07)	25.95 (3.88)	17.00 (2.87)
	Time 2	18.11 (2.69)	15.16 (7.34)	23.89 (7.20)	18.32 (3.16)

Table 6.4: Mean index scores for each interpretation of the abstract conditional statement at Time 1 and Time 2 in each group (standard deviations in parentheses).

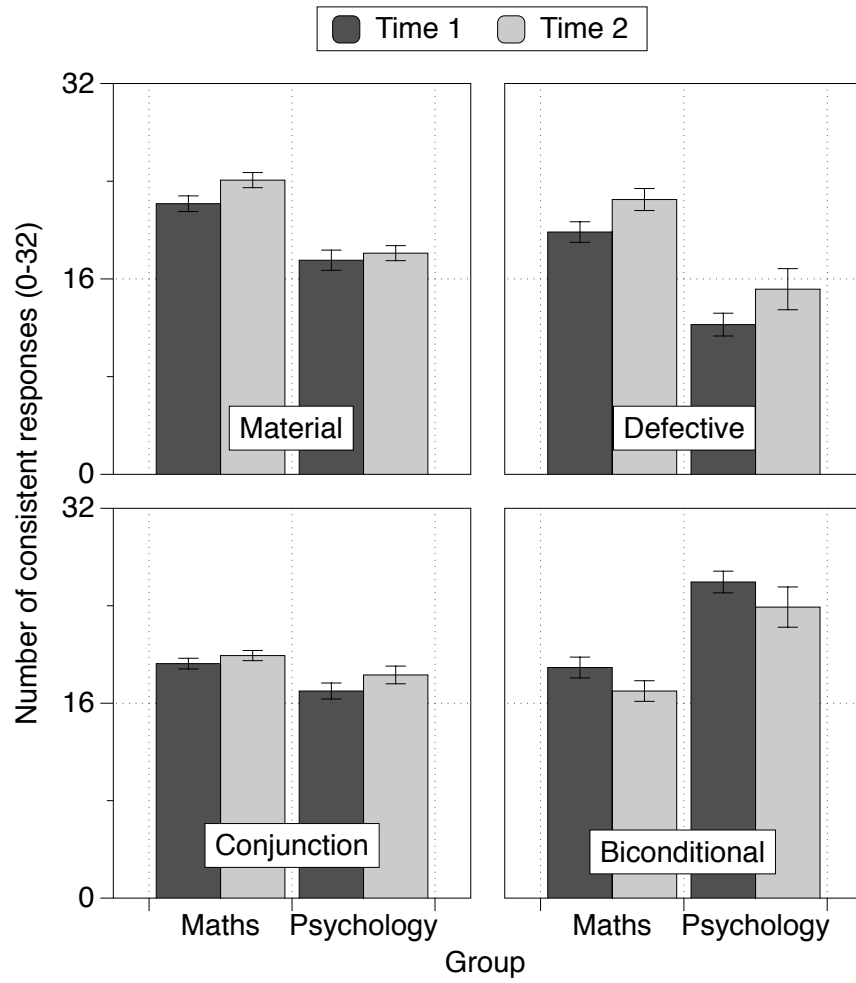


Figure 6.6: Interaction between Group and Time on Abstract Conditional Inference indices (error bars show ± 1 standard error of the mean).

1, $t(67) = 4.23, p < .001, d = 1.17$, and at Time 2, $t(78) = 6.71, p < .001, d = 1.72$.

A similar pattern of results emerged for the DCI. Although there was no significant interaction, $F < 1$, paired samples t-tests demonstrated that the mathematics group's DCI was significantly higher at Time 2 ($M=22.50, SD=5.98$) than at Time 1 ($M=19.84, SD=5.58$), $t(43) = 4.35, p < .001, d = 0.66$, whereas the psychology group's DCI was only marginally higher at Time 2 ($M=15.16, SD=7.34$), than at Time 1 ($M=12.26, SD=4.07$), $t(18) = 2.05, p = .056, d = 0.49$. The mathematics group's DCI was significantly higher than the psychology group's both at Time 1, $t(67) = 5.28, p < .001, d = 1.55$, and Time 2, $t(78) = 4.98, p < .001, d = 1.09$.

For the CCI, there was no significant interaction between Time and Group, $F(1, 59) = 1.50, p = .225, \eta_p^2 = .025$. The mathematics group's scores did not change over time, $t(43) = 1.45, p = .155, d = .23$, while the psychology group's CCI increased marginally over time, $t(18) = 1.99, p = .062, d = .44$, (Time 1: $M=17.00, SD=2.87$, Time 2: $M=18.32, SD=3.16$).

The analysis on BCI scores also showed no significant interaction, $F(1, 59) < 1$, but paired samples t-tests showed that the mathematics group's BCI decreased over time, $t(43) = 3.07, p = .004, d = -0.46$, (Time 1: $M=18.93, SD=5.67$, Time 2: $M=17.00, SD=5.61$), while the psychology group's BCI did not change, $t(18) = 1.46, p = .161, d = -0.36$.

The power for these analyses was low due to the unexpectedly high drop out rate in the psychology group (resulting in only 19 psychology participants who completed the abstract Conditional Inference task and covariates at both time points⁴). The t-tests presented above suggested that there were trends in the direction predicted by the AS level results in the mathematics group. However, the means presented in Table 6.4 suggest that the psychology group showed a similar pattern of changes, although not significantly and with smaller effect sizes.

Thematic Conditional Inference.

Endorsement rates. Endorsement rates of each inference type were analysed with a $2 \times 4 \times 2$ ANOVA with two within-subjects factors: Time (start and end of the year) and Inference Type (MP, DA, AC, MT), one between-subjects factor: Group (mathematics and psychology), and two covariates: Time 1 RAPM and Time 1 CRT. There was no significant three-way interaction, $F < 1$, (see Figure 6.7). The means and standard deviations for this interaction are displayed in Table 6.5. Neither group changed in their endorsement rates of any of the

⁴A second cohort of students are currently taking part in the study and a more powerful analysis will be carried out after May 2013.

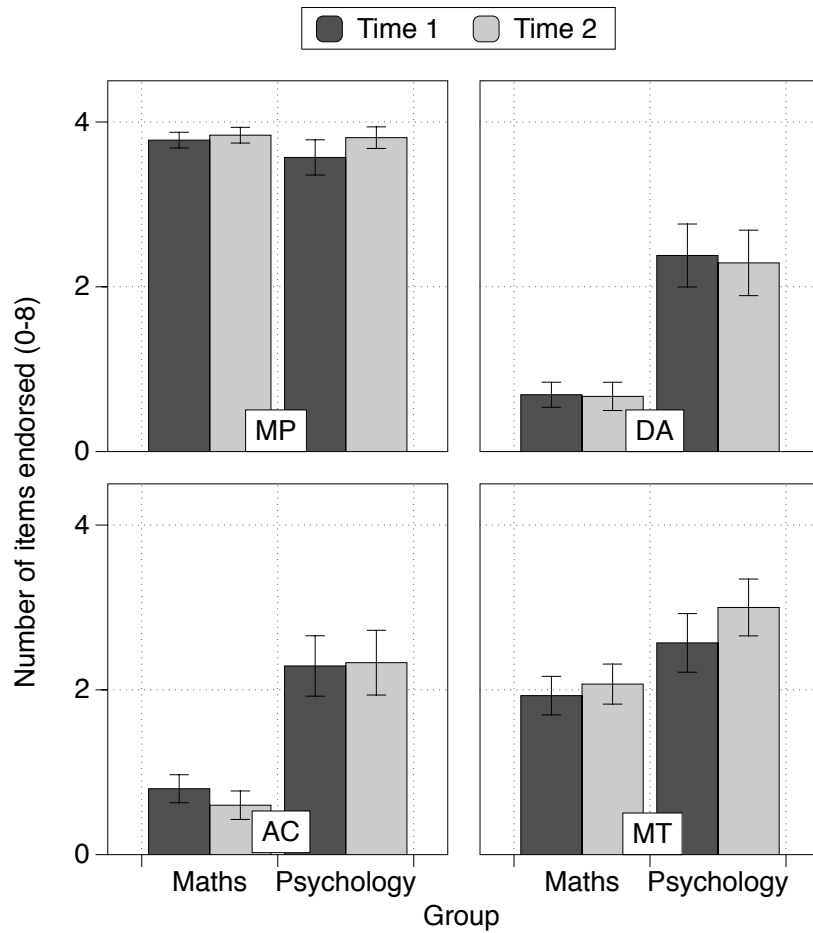


Figure 6.7: Mean endorsement rates for each of the four inferences in each group at Time 1 and Time 2 on the Thematic Conditional Inference Task (error bars show ± 1 standard error of the mean).

inference types over time (all $ps > .25$).

Interpretations. Next, each group's inclination towards the four interpretations of the thematic conditional were investigated.

Four 2×2 ANOVAs were carried out, each with one within-subjects factor: Time (start and end of the year), one between-subjects factor: Group (mathematics, psychology) and two covariates: RAPM and CRT. The means for each groups' interpretation indices at each time point are presented in Table 6.6 and Figure 6.8. There were no significant interactions between Time and Group for any of the interpretation indices, all $ps > .330$. Furthermore, unlike on the Abstract Conditional Inference task, the mathematics students' four index scores did not change over time on the thematic version, all $ps > .235$, and neither did

Inference	Group	Time 1	Time 2
MP	Mathematics	3.78(0.64)	3.84(0.64)
	Psychology	3.57(0.98)	3.81(0.60)
DA	Mathematics	0.69(1.02)	0.67(1.15)
	Psychology	2.38(1.75)	2.29(1.82)
AC	Mathematics	0.80(1.14)	0.60(1.16)
	Psychology	2.29(1.68)	2.33(1.80)
MT	Mathematics	1.93(1.57)	2.07(1.63)
	Psychology	2.57(1.63)	3.00(1.58)

Table 6.5: Mean number of items endorsed on the Thematic Conditional Inference Task by Inference type, Group and Time point with standard deviations in parentheses.

the psychology students' scores, all $ps > .175$.

As with the Abstract Conditional Inference task, the power for the analyses of the Thematic Conditional Inference task was low due to the high drop out rate in the psychology group. However, in this case not even the paired t-tests comparing the mathematics group's interpretation indices across the two time points were significant.

Finally, the BBI scores for each group at each time point were analysed with a 2×2 ANOVA with one within-subjects factor: Time (start and end of the year) and one between-subjects factor: Group (mathematics, psychology), and two covariates, RAPM and CRT. This showed no main effect of Group, $F(1, 62) < 1$, no main effect of Time $F(1, 62) < 1$, and no interaction, $F(1, 62) < 1$ (which remained non-significant without CRT as a covariate, $p = .482$, see Table 6.7 for means and standard deviations). Paired t-tests showed that neither group's BBI changed significantly over time (both $ps > .290$).

Because there were no significant interactions between Group and Time for

Group	Time	Material	Defective	Biconditional	Conjunctive
Maths	Time 1	12.22 (2.79)	12.36 (2.66)	7.20 (2.64)	9.56 (1.59)
	Time 2	12.64 (2.78)	12.51 (2.76)	7.18 (2.38)	9.71 (1.80)
Control	Time 1	9.48 (2.06)	8.33 (4.61)	10.81 (5.16)	8.90 (1.97)
	Time 2	10.19 (3.19)	8.19 (4.69)	11.43 (4.65)	8.86 (1.68)

Table 6.6: Mean index scores for each interpretation of the thematic conditional statement at Time 1 and Time 2 in each group (standard deviations in parentheses).

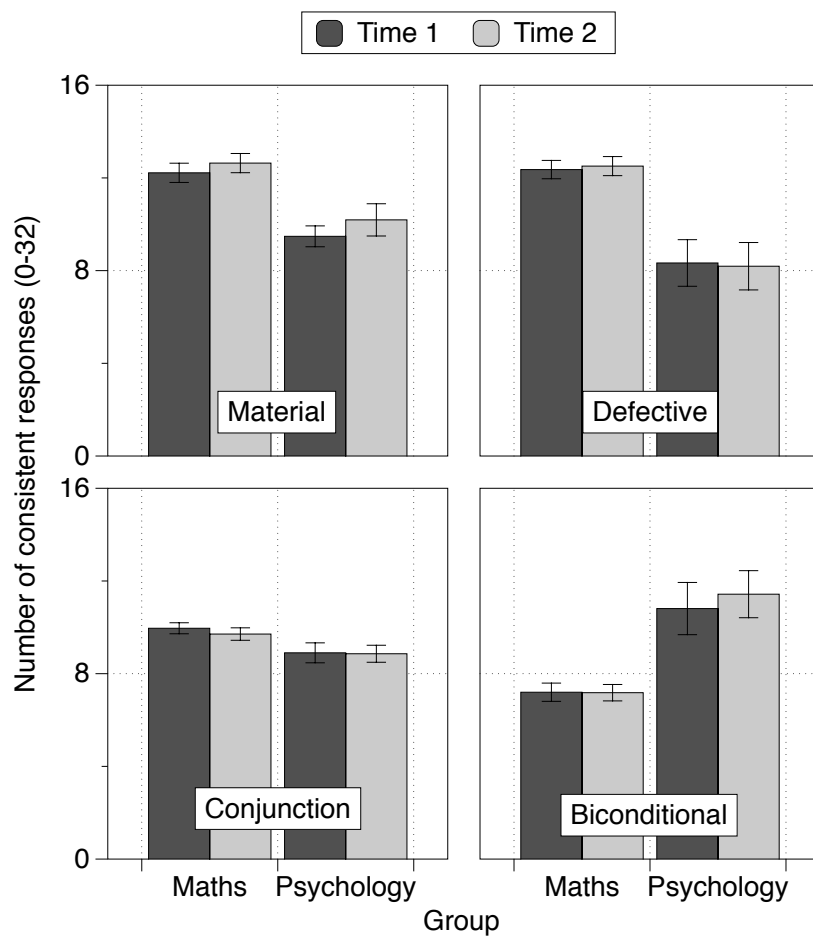


Figure 6.8: Interaction between Group and Time on Thematic Conditional Inference interpretation indices (error bars show ± 1 standard error of the mean).

	Mathematics	Psychology
Time 1	.31 (1.47)	.62 (1.24)
Time 2	-.02 (1.20)	.38 (1.02)

Table 6.7: Mean belief bias index scores for each group at Time 1 and Time 2 (standard deviations in parentheses).

any of the outcome measures, there was no need for regression analyses predicting change from RAPM, CRT and Group.

6.4 Discussion

This chapter aimed to investigate the development of reasoning skills in undergraduate mathematics students. Mathematics students were compared to psychology students in a longitudinal investigation of abstract and thematic conditional reasoning behaviour.

Hypothesis 1 predicted that mathematics students would become more defective, and to a lesser extent more material, in their interpretation of abstract conditional statements. This hypothesis was not supported by a series of ANOVA analyses, but paired t-tests showed that the mathematics students' abstract conditional reasoning became more defective ($d = 0.66$, which is similar to the change in the AS level mathematics students, $d = 0.88$), more material ($d = 0.55$, compared to $d = 0.49$ in the AS level mathematics students), and less biconditional over time ($d = -0.46$, compared to $d = -0.51$ in the AS level mathematics students), whereas the psychology students' behaviour did not change. This is consistent with Hypothesis 1 and with the results from AS level students in Chapter 5, but the lack of statistical significance in the interaction analyses prevents any firm conclusions from being drawn. The analyses suffered from low statistical power due to the small sample size in the psychology group, and this may have prevented the ANOVA analyses from reaching significance.

There were no clear predictions for the Thematic Conditional Inference task. Chapter 5 showed no change in a Thematic Syllogisms task in AS mathematics students, and it was unclear whether the discrepancy with abstract conditional inference change in that study was due to the thematic aspect of the task or the syllogistic aspect. In the present study, the mathematics students' reasoning with abstract conditionals showed a trend in the predicted direction, but did not show any trend in a Thematic Conditional Inference task, despite the items having the same structure. This suggests that context may be the issue. It may be the case that studying mathematics only influences one's ability to think about abstract conditional statements, which is plausible given that much of mathematics itself is abstract in nature. If this is the case there are important implications for the Theory of Formal Discipline. Statements such as that made by Oakley, that "the study of mathematics cannot be replaced by any other activity that will train and develop mans purely logical faculties to the same level of rationality" (1949, p.19), and Plato, that "we must endeavor to persuade those who are to be the principal men of our state to go and learn arithmetic"

(375B.C/2003, p. 256), seem to imply a more general influence of studying mathematics on reasoning skills, an influence that would extend to rational day-to-day reasoning. If mathematics actually only influences abstract conditional reasoning behaviour, then the impact on everyday reasoning is surely quite limited.

Another issue that was raised in Chapter 5 was the confound between studying mathematics and science subjects at A level. It was not possible to investigate the influence of studying mathematics independently of science subjects, but there was no correlation between number of science subjects studied and changes in DCI within the mathematics group. The undergraduate participants in this study were all taking single-honours degrees, meaning that the mathematics students were studying *only* mathematics, and they still showed a trend towards becoming more material and defective in their reasoning. This suggests that mathematics alone is associated with changes in conditional reasoning, although the effect may still be exaggerated by studying other science subjects in conjunction.

A final issue that this study aimed to resolve was the discrepancy in findings between Lehman and Nisbett (1990) and Inglis and Simpson (2009a). Lehman and Nisbett (1990) found that the number of mathematics modules taken by undergraduates at a US university was correlated with the extent of their improvement on a conditional reasoning task. Inglis and Simpson (2009a), however, failed to find any improvement on an abstract conditional inference task in mathematics undergraduates at a UK university. The data presented here adds (a small amount of) weight to the findings of Lehman and Nisbett (1990), suggesting that there is a relationship between undergraduate level mathematics and development of conditional reasoning skills. The reason for no change or trends being found in Inglis and Simpson's (2009a) study is unclear – their dependent measure and participants' level of study was the same as in this study. Perhaps the modules studied by our participants and theirs were different and are differentially related to reasoning skills. Another possibility is that there were changes in the defective interpretation, which Inglis and Simpson (2009a) did not investigate. Further research which compares the influence of studying different mathematics curriculums, both at undergraduate and other levels of education, on development in conditional reasoning skills would be useful for clarifying this issue.

Concerning the potential mechanisms of the mathematics students' improvement in abstract conditional reasoning skills, the data presented in this chapter is unfortunately of little help. The trends reported are consistent with the hypothesis laid out in Chapter 5, that studying mathematics teaches students to assume p and reason about q , thereby making their reasoning from conditional

statements more defective. However, the lack of significant interactions between Time and Group for the DCI or any of the other indices means that the results are only suggestive, and it was not possible to examine the influence of changes to RAPM and CRT on changes to reasoning behaviour.

6.4.1 Summary of novel findings

The study presented here has made three tentative contributions to our understanding of the Theory of Formal Discipline:

1. Studying mathematics at undergraduate level is associated with a trend towards more defective, more material, and less biconditional reasoning.
2. Studying mathematics in isolation from science subjects is associated with similar changes to those found in the AS level students who were studying mathematics in conjunction with science subjects.
3. The changes may only apply to abstract conditional reasoning problems, not thematic ones.

Chapter 7

Linguistic factors in mathematics students' conditional reasoning: 'if then' versus 'only if'

7.1 Introduction

The aim of this chapter is to establish whether or not mathematics students hold a deep understanding of conditional logic, i.e. do they reason based on the logical structure of the conditional, or are they influenced by surface features? Chapters 5 and 6 suggested that mathematics students' development in reasoning was limited to abstract conditional inference, not extending to thematic conditional inference or thematic syllogisms. The question raised here is the extent to which mathematics students' reasoning is generalised *within* abstract conditional inference. The conditional inference task used in the other studies of this thesis presented participants with conditional statements of the form 'if p then q '. A logically equivalent way to phrase this statement is ' p only if q ' (Evans, 1977), as demonstrated in truth table form in Table 7.1. If mathematics students hold a deep understanding of conditional logic, then they should treat 'if then' (IT) and 'only if' (OI) statements in the same way. This was not the case with non-mathematics students in a study by Evans (1977).

Evans (1977) presented undergraduate students, who were not specifically from mathematics courses, with a 16 item version of the conditional inference task. Half of his participants saw the items phrased 'if p then q ' and the other

p	q	p only if q	if p then q
T	T	T	T
T	F	F	F
F	T	T	T
F	F	T	T

Table 7.1: Truth table for the conditional statements ‘ p only if q ’ and ‘if p then q ’.

half saw them as ‘ p only if q ’. Evans hypothesised an interpretational difference between IT and OI rules, based on the principles of necessity and sufficiency. In material implication, the antecedent is sufficient for the consequent and the consequent is necessary for the antecedent. However, these two principles are differentially emphasised in IT and OI rules. Using Evans’s example, the IT rule ‘If he is a policeman then he is over 5ft 9in in height’ seems to emphasise the sufficiency of the antecedent. On the other hand, the equivalent OI rule ‘He is a policeman only if he is over 5ft 9in in height’, seems to emphasise the necessity of the consequent (Evans, 1977, p. 300). Based on this difference, Evans hypothesised two things: firstly, that more MP inferences would be made on IT rules than OI rules, because the minor premise in MP affirms the antecedent, whose sufficiency is emphasised in IT rules, and secondly, that more MT inferences would be made on OI rules than IT rules, because the minor premise in MT negates the consequent, whose necessity is emphasised in OI rules (when the necessary consequent is negated it is more obvious from OI rules that the antecedent must also be negated).

These hypotheses were supported by the data. MP inferences were endorsed 100% of the time in the IT condition, compared to 76% in the OI condition. MT inferences, on the other hand, were endorsed 42% of the time in the IT condition compared to 59% of the time in the OI condition. An unpredicted difference found by Evans (1977) was in AC inferences. Evans did not predict a difference between IT and OI interpretations of AC inferences, but found a higher endorsement rate in the OI condition, of 84%, than in the IT condition, of 67%. Evans conjectured that this could be due to participants making a conversion, where they take ‘ p only if q ’ to mean ‘if q then p ’; an IT reading with the antecedent and consequent reversed. In this case, an OI AC inference would be equivalent to an IT MP inference, which could explain the high endorsement rate of the invalid AC inference in the OI condition. An OI MP would become an IT AC, an OI DA an IT MT, and an OI MT would become an IT DA. When reclassified in this way, the OI endorsement rates became more similar to the IT endorsement rates on all but OI MT/IT DA (see Table 7.2), so the conversion

IT inference	IT endorsements	OI endorsements	OI inference
MP	100	84	AC
DA	38	59	MT
AC	67	76	MP
MT	42	38	DA

Table 7.2: Percentage endorsement rates for each IT inference and the equivalent converted OI inference in Evans’s (1977) study.

hypothesis may go some of the way to explaining the interpretational difference.

In the study presented below, the aim was to see whether advanced mathematics students would show the same interpretational difference found by Evans (1977), or whether they would respond simply to the logic of the conditional and interpret IT and OI statements in the same way, given that they are logically equivalent. The TFD would predict the second scenario; it suggests that studying mathematics “develop[s] mans purely logical faculties” (Oakley, 1949, p. 19), which is taken here to mean that they gain a deep understanding of logic free from the influence of surface-level interference.

However, if mathematics students do simply become familiar with forward inferences, such as ‘if p then q ’, and learn to assume p , they will respond differently to the seemingly ‘backwards’ inference ‘ p only if q ’. In this case, the invitation is not to assume p but rather to question it. It is not clear what the preferred interpretation of an OI statement would be in this case, but the defective interpretation would presumably be significantly lower in the OI condition than in the IT condition. It might be expected that AC inferences would be endorsed significantly more in the OI condition than in the IT condition; an OI conditional emphasises the uncertainty of q , but the necessity of p when q is true. Therefore an AC inference, in which q is confirmed, would presumably lead to high endorsement rates. This was the pattern found by Evans (1977) with non-mathematics students (explained as a conversion from OI AC to IT MP). If mathematics students do reason defectively because they have become familiar with ‘forward’ IT inferences, then in the case of ‘backward’ inferences (OI) we would expect them to reason in the same way as non-mathematics students: with higher endorsement of AC.

7.2 Method

7.2.1 Participants

Participants were 61 third year mathematics undergraduate students at Loughborough University. They took part on a voluntary unpaid basis during a lecture course on Applied Statistics and later analysed the anonymised data as part of an SPSS lab session.

7.2.2 Design

The study followed a between-subjects experimental design with two conditions: ‘if then’ and ‘only if’ phrasing of the conditional inference task. Thirty participants were assigned to the ‘if then’ condition and 31 to the ‘only if’ condition.

7.2.3 Measures

Participants completed the 32 item conditional inference task with the conditional statement phrased as either ‘if p then q ’ (IT condition) or as ‘ p only if q ’ (OI condition). The letters and numbers used in the problem were identical across conditions, as were the task instructions. The order of items was randomised between participants.

7.2.4 Procedure

Participants took part during a lecture. They were informed that they would be given a reasoning task and that they would analyse the data in a subsequent SPSS lab session. Participants were asked to work alone and in silence and were not informed that there were two versions of the task. Task booklets were handed out to the participants with the conditions in an alternating order so that the condition a participant received depended only on the order in which they sat in the lecture hall.

7.3 Results

The results are presented in two parts. First, an analysis of endorsement rates for each inference is presented for comparison to Evans’s (1977) study. Second, an analysis of the interpretation indices in each condition is presented.

	If then	Only if
Modus Ponens	7.20 (1.00)	6.10 (1.90)
Denial of the antecedent	2.90 (2.43)	3.97 (2.17)
Affirmation of the consequent	3.43 (2.52)	6.65 (1.96)
Modus Tollens	4.67 (2.26)	5.74 (1.73)

Table 7.3: Mean number of items endorsed (out of 8) for each of the four inferences by condition. Standard deviations in parentheses.

7.3.1 Endorsement rates

Mean endorsement rates were subjected to a 2 (condition: IT, OI) \times 4 (inference type: MP, DA, AC, MT) analysis of variance (ANOVA). This revealed a significant main effect of inference type, $F(3, 177) = 35.81, p < .001, \eta_p^2 = .38$, with MP items being endorsed most often ($M = 6.64, SD = 1.61$), followed by MT items ($M = 5.21, SD = 2.07$), followed by AC items ($M = 5.07, SD = 2.76$), with DA items being endorsed least often ($M = 3.44, SD = 2.34$). There was also a significant main effect of condition, $F(1, 59) = 8.74, p = .004, \eta_p^2 = .13$, with more items being endorsed in the OI condition ($M = 22.45, SD = 5.27$) than the IT condition ($M = 18.20, SD = 5.95$). Finally, there was a significant interaction between condition and inference type, $F(3, 177) = 16.07, p < .001, \eta_p^2 = .21$. The means for this interaction are shown in Table 7.3. Significantly more MP inferences were endorsed in the IT condition than the OI condition, $t(59) = 2.82, p = .007, d = 0.72$. Conversely, significantly fewer AC inferences, $t(59) = 5.58, p < .001, d = -1.43$, and MT inferences $t(59) = 2.09, p = .041, d = -0.53$, were endorsed in the IT condition. There were also marginally fewer DA inferences endorsed in the IT condition than the OI condition, $t(59) = 1.81, p = .075, d = -0.46$. These data are displayed in Figure 7.1.

To investigate the plausibility of Evans's (1977) conversion hypothesis for explaining the different endorsement rates across conditions, the IT endorsement rates are compared to the converted OI endorsement rates in the same manner (summarised in Table 7.4). This gives 6.09 OI MP inferences endorsed compared to 3.43 IT AC inferences, 3.97 OI DA inferences endorsed compared to 4.67 IT MT inferences, 6.65 OI AC inferences endorsed compared to 7.20 IT MP inferences, and 5.74 OI MT inferences endorsed compared to 2.90 DA inferences. The match provided here is not as close as in Evans's data.

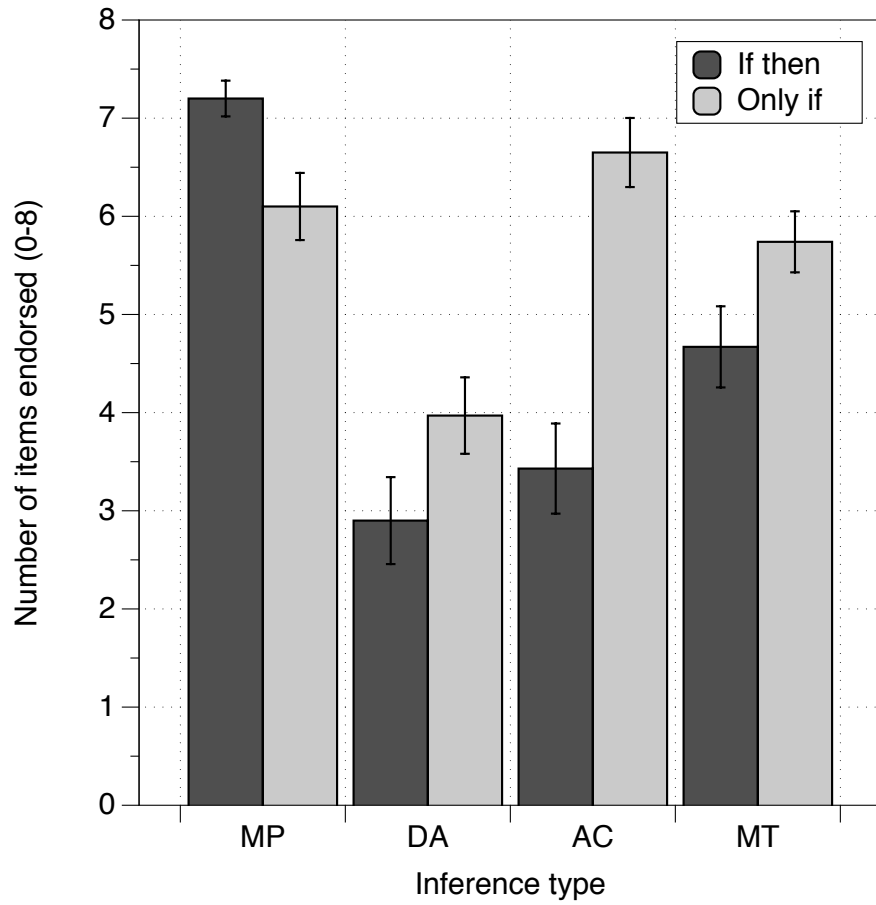


Figure 7.1: Mean endorsement rates for each inference type in the IT and OI conditions (error bars represent ± 1 SE of the mean).

IT inference	IT endorsements	OI endorsements	OI inference
MP	7.20	6.65	AC
DA	2.90	5.74	MT
AC	3.43	6.09	MP
MT	4.67	3.97	DA

Table 7.4: Mean endorsement rates for each IT inference and the equivalent converted OI inference.

7.3.2 Interpretations of the conditional

Next, the four interpretation indices were compared across conditions. A 2×4 ANOVA with one between-subjects factor: Conditional Type (IT, OI) and one within-subjects factor: Interpretation (MCI, DCI, BCI, CCI) revealed a significant main effect of Interpretation, $F(3, 177) = 5.067, p = .002, \eta_p^2 = .08$, a significant main effect of Conditional Type, $F(1, 59) = 14.29, p < .001, \eta_p^2 = .20$, and a significant interaction, $F(3, 177) = 14.27, p < .001, \eta_p^2 = .20$ (see Figure 7.2). The interaction was followed up with an independent samples t-test for each of the four interpretations comparing the means in the IT and OI conditions. The MCI was significantly higher in the IT condition ($M=21.53, SD=4.19$) than in the OI condition ($M=17.23, SD=3.62$), $t(59) = 4.30, p < .001, d = 1.10$. The DCI was also higher in the IT condition ($M=20.20, SD=5.87$) than in the OI condition ($M=13.74, SD=4.40$), $t(59) = 4.87, p < .001, d = 1.25$. Conversely, the BCI was significantly lower in the IT condition ($M=18.20, SD=5.95$) than in the OI condition ($M=22.45, SD=5.27$), $t(59) = 2.96, p = .004, d = -0.76$. There was no difference between the CCI in the IT ($M=19.07, SD=3.27$) and OI conditions ($M=19.03, SD=3.07$), $t(59) = .04, p = .966, d = 0.01$.

7.4 Discussion

In this thesis, mathematics students have been shown to become more defective in abstract conditional reasoning (see Chapters 5 and 6). Here, the nature of mathematics students' conditional reasoning behaviour was investigated; do they respond to the underlying logical structure of a conditional inference problem, as predicted by the TFD, or are they swayed by the linguistic phrasing, as was the case for Evans's (1977) non-mathematics participants?

A group of third year undergraduate mathematics students completed the 32 item conditional inference task in one of two conditions: with IT phrasing or with OI phrasing. The results were largely in line with those of Evans (1977) who studied non-mathematics students. Participants endorsed more MP inferences in the IT condition, and more AC and MT inferences in the OI condition. This suggests two things: firstly that mathematics students do not treat conditional statements based on the underlying logic, and secondly that Evans's unpredicted finding of higher endorsement rates of AC inferences in the OI condition than the IT condition was replicated.

Mathematics students did not appear to respond to conditional inference problems based on the underlying logical structure. Rather, they interpreted the conditional statements differently depending on the phrasing. There were main effects of condition for both endorsement rates and interpretation indices,

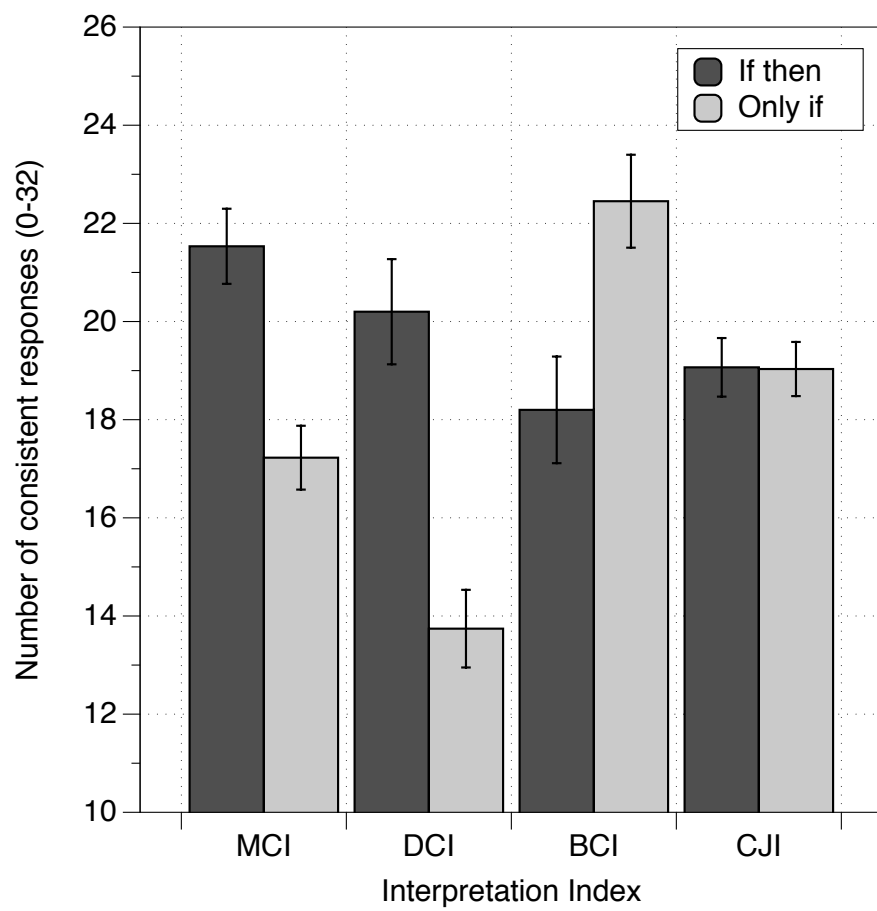


Figure 7.2: Mean interpretation indices in the IT and OI conditions (error bars represent ± 1 SE of the mean).

with the students reasoning more in line with the material and defective interpretations in the IT condition than in the OI condition and more in line with the biconditional interpretation in the OI condition than in the IT condition.

This is contrary to the TFD, which posits that “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” (Oakley, 1949, p. 19), that “through mathematics we also wish to teach logical thinking” (Amitsur in Sfard, 1998, p. 453), and that mathematics “disciplines the mind, develops logical and critical reasoning, and develops analytical and problem-solving skills to a high degree (Smith, 2004, p. 11). It seems clear that these quotes support the hypothesis of mathematics students holding an abstract understanding of logic and being able to avoid the influence of phrasing or context with their ‘disciplined minds’. Instead, they were very much swayed by the phrasing of the problems, with the IT and OI groups responding significantly differently to three of the four inference types (MP, AC, MT) and marginally differently to the fourth (DA).

This is inconsistent with the TFD but consistent with the hypothesis that mathematics students learn to assume p and reason about q , i.e. to become more defective in their conditional reasoning. While it is possible that students do become more defective with OI statements over time, the present study suggests that they are not as defective with OI statements as they are with IT statements by the third year of an undergraduate mathematics degree. The explanation proposed for this finding, as in Chapters 5 and 6 is that mathematics exposes students to statements of the IT form more often than the OI form, and that this familiarity with, practice with, and feedback on problems including IT statements fosters students’ competency with them. A lack of exposure to OI statements could mean that mathematics students treat them in same way as non-mathematicians do. Indeed, Houston (2009) argued that most mathematical statements are of the form ‘if statement A is true, then statement B is true’, even if they are heavily disguised (p. 63). Furthermore, he argued that in mathematics A is assumed to be true, even if it is clearly not true, and the truth or falsity of B is then deduced. This is a forward moving deduction, whereas ‘ A only if B ’ can be seen a backwards deduction where A is in question. As Houston points out, it can be difficult to see this as being equivalent to ‘if A then B ’. We return to this idea below. First, Evans’s (1977) conversion hypothesis is considered, which we shall see points to the same conclusion.

The data supported Evans’s (1977) unpredicted finding of more AC inferences being endorsed in the OI condition than in the IT condition. Evans suggested that this was due to participants making an illicit conversion from ‘ p only if q ’ to ‘if q then p ’. In this case, an OI AC inference is equivalent to

	Maths IT	Evans's IT	Maths OI	Evans's OI
Modus Ponens	90.0%	100%	76.2%	87.5%
Denial of the antecedent	36.3%	68.8%	49.6%	50.0%
Affirmation of the consequent	42.9%	75.0%	83.1%	81.3%
Modus Tollens	58.3%	75.0%	71.8%	81.3%

Table 7.5: Percentage of items endorsed for each inference type across participants in four groups: the current maths IT group, the current maths OI group, Evans's (1977) IT non-maths group and Evans's OI non-maths group.

an IT MP inference, which would explain the high endorsement rate. However, although the conversion hypothesis seemed to provide a good fit to Evans's data in most cases, it did not seem so well suited to this data. There are fairly large differences between both MP and MT endorsement rates from IT problems and the equivalent converted OI endorsement rates.

This may be because, as suggested above, mathematics students perform similarly to non-mathematicians on OI statements but in a more material and defective way than non-mathematicians on IT statements, so performance in the two conditions cannot be compared as if they were a product of the same process. In other words, the mathematicians' IT performance is qualitatively different to their OI performance and to non-mathematicians' performance on both phrasings. An inspection of the endorsement rates in Table 7.5 shows that the mathematics students in the IT group endorsed considerably fewer DA, AC and MT inferences than the non-mathematics IT group in Evans's study. The endorsement rates for the mathematics OI group, however, do not differ drastically from Evans's non-mathematics OI group.

Taken together, the differential performance in the IT group compared to the OI group, and the informal similarity of mathematicians and non-mathematicians OI performance, suggests that mathematics students' advantage in conditional reasoning is limited to IT statements, while their performance on OI statements is similar to that of non-mathematicians. An explanation for this could be that mathematics students become very used to assessing IT statements but don't have so much exposure to OI statements (Houston, 2009). Contrary to the TFD, this leads them to behave more defectively (and materially) on IT conditional statements but it doesn't change their interpretation of conditionals per se.

7.4.1 Summary of novel findings

1. It appears as though the mathematics advantage in conditional reasoning that has been found previously is not due to better understanding of

the underlying logic. Instead, mathematics students only appear to be successful with conditional statements phrased ‘if p then q ’ and not with those phrased ‘ p only if q ’.

2. In the case of OI statements, mathematicians may well behave in much the same way as non-mathematicians.
3. What this means for the TFD is that grand claims about mathematics improving the ‘purely logical faculties’ are in fact overstated. The relationship between mathematics and logical reasoning skills may be far narrower than previously thought.

Chapter 8

The role of the heuristic level in mathematics students' conditional reasoning

8.1 Introduction

According to Stanovich's (2009a) tri-process model of cognition, reasoning occurs through three levels: the heuristic level, the algorithmic level and the reflective level. Heuristic, or Type 1 processing, is fast, automatic, and undemanding of working memory resources. Type 2 processing is deliberate and demanding of working memory, being implemented at the algorithmic level and prompted by the reflective level (see Section 2.6). It is possible that the mathematics students' increased defective reasoning found in Chapter 5 comes about via one or more of these three levels. The aim of this chapter is to investigate the potential of the heuristic level to be the source of the change. Is it the case that mathematics students' increased tendency for defective reasoning stems from the heuristic level of cognition?

The heuristic level of cognition is characterised by fast and automatic processing that is largely obscured from conscious reflection. The output of heuristic processing and an associated feeling of the rightness of the output are available to introspection (Thompson, 2010; Thompson, Prowse Turner & Pennycook, 2011), but can quickly be influenced by Type 2 reflection. This means that the best way to measure a heuristic-level response is with speeded accuracy (Evans

& Curtis-Holmes, 2005; Gillard, Van Dooren, Schaeken & Verschaffel, 2009b; Heit & Rotello, 2010) and reaction time methods (Gillard, 2009). On the assumption that heuristic processing is fast and automatic and Type 2 processing is slower, it can be assumed that responses given under short time limits have been produced with little or no Type 2 input.

Several studies have previously used speeded accuracy methods to investigate heuristic level processing in reasoning tasks. Evans and Curtis-Holmes (2005) gave participants 10 seconds to respond to each item on a belief bias syllogisms task and found that the number of logical responses decreased while belief bias increased, relative to a free-time condition. This suggests that the influence of prior beliefs on reasoning behaviour is a heuristic process and that Type 2 processing is required to override it in order to give normatively logical answers. Similarly, when Gillard et al. (2009b) limited their participants to 17 seconds for responding to problems that induced a proportionality heuristic, the number of proportional responses increased and the number of correct answers decreased. This suggests that proportional reasoning is also heuristic based and that Type 2 processing is required to override it in cases where it leads to incorrect responses.

I am aware of only one study that has investigated conditional reasoning under time restraints. Evans et al. (2009) gave their participants 48 conditional reasoning problems which differed in believability and validity, and had them respond either within 5 seconds of the conclusion being shown or without time pressure. The problems consisted of the four usual inferences, modus ponens (MP), denial of the antecedent (DA), affirmation of the consequent (AC) and modus tollens (MT), and participants responded with a 'yes' or 'no' to indicate whether they thought the conclusion necessarily followed from the premises. Under time pressure participants were less inclined to accept inferences overall, and although an analysis of inference type was not reported, an inspection of means suggests that the decline in acceptance did not differ by inference type.

The mathematics students investigated in Chapter 5 became less inclined to accept MT, DA and AC inferences over time, thereby becoming more defective in their reasoning. It seems possible from Evans et al.'s (2009) findings that this could be due to a heuristic process: if greater reliance on heuristic processing leads to higher rejection rates of conditional inferences, it could be the case that mathematics students become more inclined to rely on their intuitions over time and that this leads to their increased rejection rates. Alternatively, it could be the case that exposure to 'if then' statements leads to a change in the heuristic level whereby practice in assuming p to be true leads to the assumption becoming automatic. The study reported below will investigate these possibilities. Mathematics undergraduate students and a sample of non-mathematicians completed the standard Conditional Inference task under two

conditions: one in which they were forced to respond quickly and one in which they could spend as long as they liked thinking. Several hypotheses can be derived based on previous research:

1. The non-mathematicians will accept fewer inferences overall in the fast condition than in the slow condition, in line with Evans et al.'s (2009) findings from a subject-general sample.
2. Mathematics students will respond more defectively in the slow condition than will the non-mathematicians, in line with the findings reported in Chapter 5.
3. If the heuristic level is responsible for the mathematics students' defective reasoning, then the mathematics students will remain more defective than the non-mathematics students in the fast condition, and will be no less defective in the fast condition than in the slow condition.
4. If the heuristic level is not responsible for mathematics students' defective reasoning, then they will respond no more defectively than the non-mathematicians in the fast condition, and less defectively in the fast condition than in the slow condition.

If the heuristic level is responsible for mathematics students' defective reasoning, the two explanations proposed above will need to be differentiated. The explanations were that mathematics students either become more reliant on heuristic level outputs, or that their heuristic level changes in a way that makes them focus more on assuming p . These hypotheses can be differentiated by comparing the groups' reaction times (RTs) in the slow condition: if mathematicians rely more on their heuristic level processing, then even in the slow condition they should respond faster than non-mathematicians.

8.2 Method

8.2.1 Design

Mathematics and non-mathematics students completed Evans et al.'s (1995) Conditional Inference task twice: first under speeded conditions and second with as much time as they wanted. The fast version was always completed first: in a within-subjects design there was a risk that completing the slow condition first could allow participants to memorise some items or remember which types of inference they considered valid and invalid, and that this could subsequently influence their performance in the fast condition. Participants also completed a

subset of Raven’s Advanced Progressive Matrices (RAPM) after the Conditional Inference Task to control for between-groups differences in intelligence.

8.2.2 Participants

Participants were 16 undergraduate and postgraduate mathematics students and 16 undergraduate and postgraduate non-mathematics students and staff from Loughborough University. There were 16 females and 16 males and the ages ranged from 18 to 51 ($M=23.90$, $SD=8.09$). Each was paid £5 for their time.

8.2.3 Procedure

The experiment was administered on a computer using E-prime 2.0 (Schneider, Eschman & Zuccolotto, 2002). First, participants completed an unrelated study involving the Cognitive Reflection Test. For the present study the first task they completed was the fast version of the Conditional Inference task. They saw instructions, 4 practice items, and 32 real items. In each trial, the conditional premise was presented alone for 1.5 seconds before the minor premise and conclusion were added simultaneously for an additional 2.5 seconds, in which time the participant was required to respond. The time limit was based on pilot testing which suggested that these timings prevented participants from being able to consciously reflect on the questions while preserving a heuristic level response (matching bias allows MP items to be readily accepted without conscious processing, see more below). In the slow version, participants again saw the conditional premise alone for 1.5 seconds, but when the minor premise and conclusion were added participants were not allowed to respond for the first 5 seconds, after which they could take as long as they wanted to respond. This restriction prevented the fast version from influencing the participants into responding quickly in the slow version.

Finally, participants completed a subset of items from RAPM (Stanovich & West, 1998), composed of 18 items with a 10 minute time limit. Participants were then thanked, paid and dismissed.

8.3 Results

8.3.1 Manipulation check and covariate assessment

As a manipulation check, endorsement rates on Modus Ponens items in the fast condition were compared to chance level with a one sample t-test. Because MP can be easily endorsed through matching bias, a heuristic response, we would

expect to find high endorsement rates even when participants are restricted to heuristic level processing, as long as the time limit wasn't too short. The time limit used here (which was based on pilot testing) was somewhat shorter than had been used in previous studies. Nevertheless, participants endorsed MP inferences at above chance levels ($M=5.78$, $SD=1.58$, maximum possible = 8), $t(31) = 6.37, p < .001$, suggesting that the time limit was sufficient for the heuristic level to generate a response.

Responses to the Conditional Inference task were coded into four variables to reflect consistency with each of the four interpretations of the conditional statement: Material Conditional Index, Defective Conditional Index, Biconditional Index and Conjunctive Index. Each variable was a score out of 32, with higher scores indicating that responding was more consistent with that interpretation of the conditional statement.

Scores on the RAPM were significantly higher in the mathematics group ($M=10.31$, $SD=2.21$) than the non-mathematics group ($M=6.19$, $SD=2.10$), $t(30) = 5.40, p < .001, d = 1.91$, and were significantly positively correlated with the Material Conditional Index in the slow condition, $r(32) = .63, p < .001$, the Defective Conditional Index in the slow condition, $r(32) = .59, p < .001$, and significantly negatively correlated with the Biconditional Index in the slow condition, $r(32) = -.53, p = .002$. RAPM and Conjunctive responding in the slow condition did not correlate significantly ($p=.223$).

Responses to the Conditional Inference task in the fast condition were not expected to correlate with the RAPM, since the fast condition was designed to reduce algorithmic level (measured by the RAPM) responding. This was indeed the case for the Defective Conditional, Biconditional and Conjunctive Indexes (all $ps > .130$). The significant correlation between RAPM and the slow Defective Conditional Index and the non-significant correlation between the RAPM and fast Defective Conditional Index were significantly different, $t(29) = 2.17, p = .019, r = .48$. This was also the case for the significant correlation between RAPM and the slow Biconditional Index and the non-significant correlation between RAPM and the fast Biconditional Index, $t(29) = 2.04, p = .025, r = .46$. This serves as a second manipulation check, suggesting that the time limit in the fast condition was not so long as to allow the algorithmic level to interfere.

However, the Material Conditional Index in the fast condition was positively correlated with the RAPM, $r(32) = .39, p = .029$. This may be due to the fact that the mathematics group had higher intelligence scores, and as shown in previous chapters, are more likely to respond in line with the material conditional than non-mathematicians. This explanation was supported by correlations performed on each group separately, which revealed no significant correlations between RAPM scores and the Material Conditional Index in the

fast condition within either group (both $ps > .400$).

8.3.2 Main analyses

Hypothesis 1: Endorsement rates in the non-mathematics group

Evans et al. (2009) found that a non-mathematics sample of participants accepted fewer inferences overall when forced to respond quickly to a Conditional Inference task. To investigate whether this was also the case here, the non-mathematics group's endorsement rates of each of the four inferences was subjected to a 2×2 ANOVA with two within-subjects factors: Time (fast, slow) and Inference (MP, DA, AC, MT). The mean endorsement rates under each condition are presented in Figure 8.1. The mathematics students were not included in this analysis because of the assumption that they reason in a qualitatively different manner to non-mathematicians (as elaborated on in Chapter 7).

Contrary to Evans et al.'s (2009) findings, there was no main effect of Time, $F(1, 45) = 1.41, p = .254, \eta_p^2 = .09$, but there was a significant interaction between Time and Inference, $F(1, 45) = 8.20, p < .001, \eta_p^2 = .35$. In the fast condition participants accepted fewer MP inferences, $t(15) = 4.65, p < .001, d = 1.46$, fewer MT inferences, $t(15) = 2.44, p = .028, d = -0.64$, and marginally more DA inferences, $t(15) = 2.03, p = .061, d = 0.58$, than in the slow condition. Whereas Evans et al. (2009) found an overall decline in endorsement rates when participants were forced to respond quickly, here it appears that only the valid inferences were significantly less likely to be endorsed.

This was further investigated with a 2×2 ANOVA with two within-subjects factors: Time (fast, slow) and Validity (valid, invalid), which revealed a significant interaction, $F(1, 15) = 24.94, p < .001, \eta_p^2 = .62$. T-tests with Bonferroni corrections revealed that participants did indeed accept fewer valid inferences in the fast condition ($M=8.25, SD=2.27$) compared to the slow condition ($M=11.38, SD=2.75$), $t(15) = 5.17, p < .001, d = -1.24$, whereas there was no significant difference between the number of invalid inferences endorsed in the fast ($M=8.25, SD=4.37$) and slow ($M=9.81, SD=3.33$) conditions, $t(15) = 1.61, p = .128, d = -0.40$, see Figure 8.2. Hypothesis 1, that non-mathematicians would endorse fewer inferences in the fast condition than in the slow condition, is therefore partially supported.

Hypothesis 2: Interpretations of the conditional without time pressure

Hypothesis 2 stated that mathematics students would respond more defectively than the non-mathematics group in the slow condition, in line with the results of

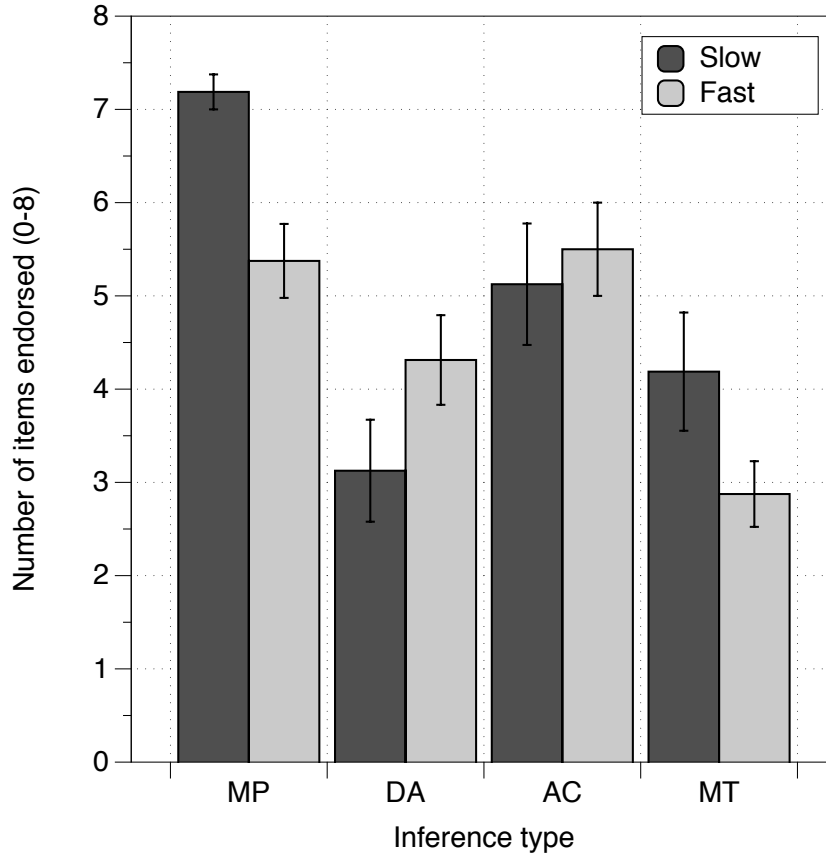


Figure 8.1: Mean endorsement rates for each inference type (out of a maximum of 8 items) in the fast and slow conditions for the non-mathematics group (error bars represent ± 1 standard error of the mean.).

Chapter 5. The mean interpretation indices in the slow condition are presented in Figure 8.3.

The interpretation indices in the slow condition were subjected to a 2×4 ANOVA with one within-subjects factor: Interpretation (Material, Defective, Biconditional, Conjunctive) and one between-subjects factor: Group (mathematics, non-mathematics). Because the indices are derived from the same set of responses and are therefore not independent, it was expected that this analysis would show a main effect of interpretation. There was in fact a marginally significant main effect of Interpretation, $F(3, 90) = 2.68, p = .052, \eta_p^2 = .082$, where the Material Conditional Index had the highest mean ($M=22.09, SD=5.06$), followed by the Defective Conditional Index ($M=20.66, SD=5.76$), then the Conjunctive Index ($M=19.91, SD=3.48$), and finally the Biconditional Index ($M=18.34, SD=5.59$).

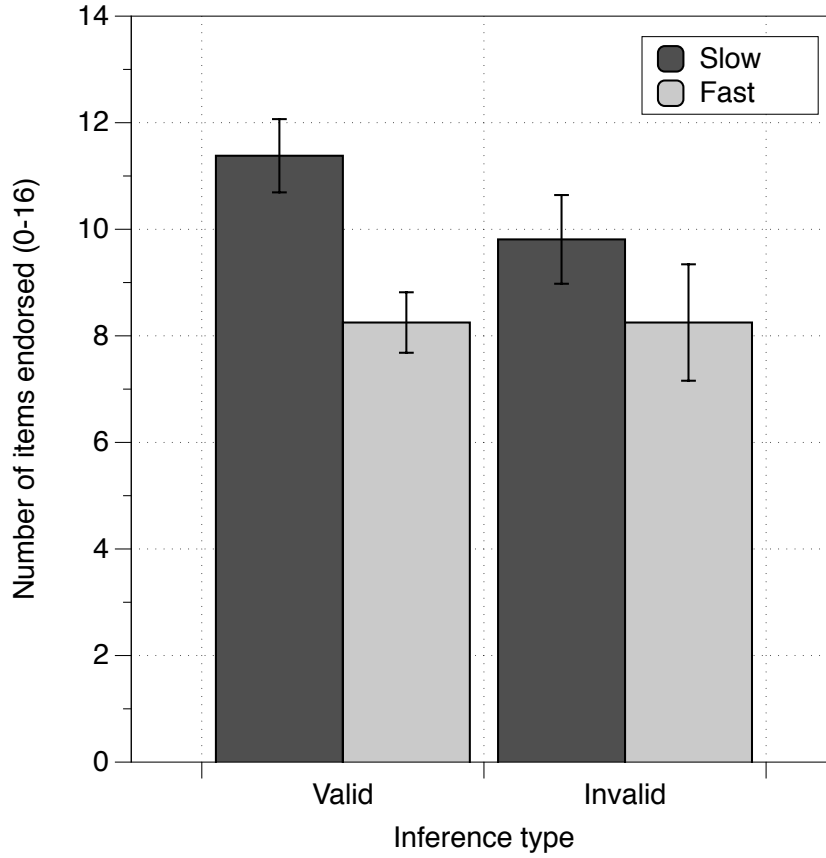


Figure 8.2: Endorsement rates for valid and invalid inferences in the fast and slow conditions for the non-mathematics group (error bars show ± 1 standard error of the mean).

There was also a significant interaction between Group and Interpretation, $F(3, 90) = 5.02, p = .003, \eta_p^2 = .143$. Planned t-tests revealed that the mathematics group had a significantly higher Material Conditional Index ($M=25.06, SD=5.85$) than the non-mathematics group ($M=19.13, SD=3.38$), $t(30) = 3.52, p = .001, d = 1.24$. The mathematics group also had a marginally higher Defective Conditional Index ($M=22.56, SD=1.19$) than the non-mathematics group ($M=18.75, SD=1.55$), $t(30) = 1.96, p = .060, d = 0.69$. A post hoc power analysis calculated the achieved power for this analysis as 0.61, suggesting that the sample size was too small despite the effect size being reasonably large. A sample of 94 participants would have been necessary to find a significant effect of this size with a power of 0.95. The mathematics group also had a marginally lower Conjunctive Index ($M=18.81, SD=3.66$) than the non-mathematics group ($M=21.00, SD=3.01$), $t(30) = 1.85, p = .075, d = -0.65$. There was no

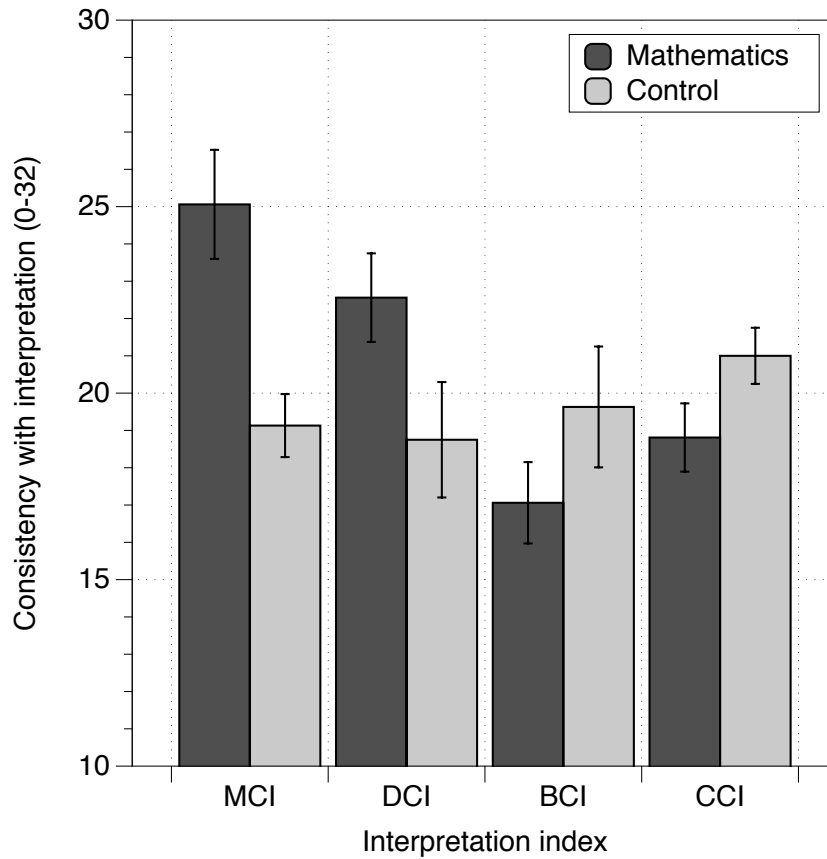


Figure 8.3: Mean interpretation indices in each group in the slow condition (error bars show ± 1 standard error of the mean).

difference between the groups in the Biconditional Index.

Although this pattern of results reflects those of Chapter 5, the interaction between Group and Interpretation lost significance once RAPM was included as a covariate, $F(3, 87) = 1.26, p = .293, \eta_p^2 = .04$. Hypothesis 2 is therefore partially supported; the mathematics group did have a marginally higher Defective Conditional Index than the non-mathematics group when RAPM was not accounted for, but they also had a significantly higher Material Conditional Index than the non-mathematics group, and with an effect size nearly twice as large ($d = 1.24$ versus $d = 0.69$). Moreover, the initial interaction lost significance once RAPM scores were accounted for, suggesting that the group differences in interpretation could not be disaggregated from the effect of differences in general cognitive ability. However, this analysis had lower power than the analyses in the AS level study.

One possible reason for the high rate of MCI in the mathematics students in

Group	Time	Material	Defective	Biconditional	Conjunctive
Maths	Fast	18.44 (4.11)	19.94 (4.91)	16.44 (3.95)	20.31 (2.82)
	Slow	25.06 (5.85)	22.56 (4.75)	17.06 (4.36)	18.81 (3.66)
Control	Fast	14.44 (3.97)	16.69 (4.19)	18.06 (4.09)	19.69 (3.28)
	Slow	19.13 (3.38)	18.75 (6.19)	19.63 (6.48)	21.00 (3.01)

Table 8.1: Mean index scores for each interpretation of the conditional statement in the fast and slow condition for the mathematics and non-mathematics (control) group (standard deviations in parentheses).

this study could be that they had a higher mean intelligence level than those in the other studies reported in this thesis. In the current group of mathematics students, the mean RAPM score was 10.31 with a 10 minute time limit, whereas in the undergraduate study the mean score at Time 1 was 11.10 with a 15 minute time limit (roughly equivalent to 7.40 under a 10 minute time limit). In the AS level students, there was again a 15 minute time limit and the mean RAPM score at Time 1 was 9.29 (roughly equivalent to 6.19 under a 10 minute time limit). These means suggest that the group of mathematics students in the current study may indeed have been more intelligent, and this in turn could account for their high MCI scores.

Hypotheses 3 and 4: Interpretations of the conditional under time pressure

Hypothesis 3 stated that if the heuristic level was responsible for the mathematics group's more defective reasoning compared to the non-mathematics group, then the difference would remain in the fast condition, and the mathematics group's Defective Conditional Index would not differ between the fast and slow conditions. Hypothesis 4 alternatively stated that if the heuristic level was not responsible, then the mathematics group would respond no more defectively than the non-mathematicians in the fast condition, and less defectively in the fast condition than in the slow condition.

To distinguish between these hypotheses, conditional inference scores were subjected to a $2 \times 4 \times 2$ ANOVA with two within-subjects factors: Time (slow, fast) and Interpretation (Material, Defective, Biconditional, Conjunctive), one between-subjects factor: Group (mathematics, non-mathematics) and one covariate: RAPM score. The means for each group's interpretation index under the fast and slow conditions are presented in Table 8.1. As expected, due to the non-independence of the interpretation scores, there was a main effect of interpretation, $F(3, 87) = 3.38, p = .022, \eta_p^2 = .104$. Surprisingly, there was no main

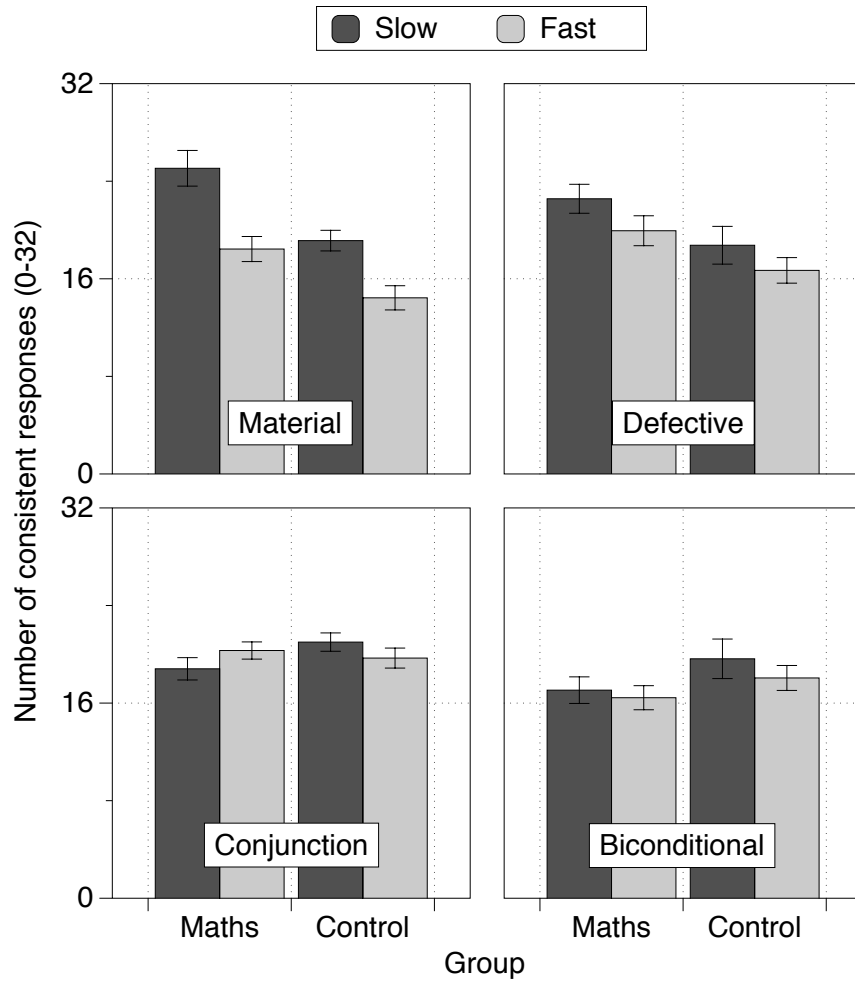


Figure 8.4: Consistency with each of the four interpretations for the mathematics and control group under the fast and slow conditions. Error bars show ± 1 s.e.m.

effect of Time, $F < 1$. However, there was a significant three-way interaction between Group, Time and Interpretation, $F(3, 87) = 2.89, p = .040, \eta_p^2 = .09$, as depicted in Figure 8.4.

This interaction was further investigated with 2×2 ANOVAs for each of the four interpretations, with one within-subjects factor: Time (fast, slow), and one between-subjects factor: Group (mathematics, non-mathematics). RAPM was also included as a covariate in the analyses of Material, Defective and Biconditional responding, given the correlations reported above. For the Material Conditional Index, there was no significant interaction between Time and Group,

$F < 1$. For the Defective Conditional Index, there was a marginally significant interaction between Time and Group, $F(1, 29) = 4.06, p = .053, \eta_p^2 = .12$, as was the case for the Biconditional Index, $F(1, 29) = 3.22, p = .083, \eta_p^2 = .10$, and the Conjunctive Index, $F(1, 30) = 3.91, p = .057, \eta_p^2 = .12$.

Given that the Defective Conditional Index is of most interest here (it increased alongside mathematical study in the AS level students in Chapter 5) and that it produced the smallest p -value ($p=.053$) and largest effect size ($\eta_p^2 = .12$) of the 2×2 ANOVA interactions reported above, the nature of the marginally significant interaction was investigated with t -tests. For the Defective Conditional Index, the mathematics group's scores were significantly higher in the slow condition ($M=22.56, SD=4.75$) than in the fast condition ($M=19.94, SD=4.91$), $t(16) = 2.49, p = .025, d = 0.54$, whereas the non-mathematics group's scores did not differ by Time ($p=.21$). The mathematics group's Defective Conditional Index remained marginally higher than the non-mathematics group's under both the fast, $t(30) = 2.02, p = .053, d = 0.71$, and slow, $t(30) = 1.96, p = .060, d = 0.69$, conditions. Hypotheses 3 and 4 were proposed as alternatives to each other, yet it seems that neither can be conclusively rejected.

Finally, each group's RTs in the slow condition were compared to test the hypothesis that mathematics students may rely more heavily on Type 1 processing than non-mathematicians. An independent samples t -test showed no significant difference between groups in RTs, $t(29) = .41, p = .685, d = 0.15$.

8.4 Discussion

The aim of this Chapter was to investigate whether the heuristic level of cognition, as described by Stanovich (2009a), could potentially be the source of the change found in mathematics students' reasoning behaviour described in Chapter 5. Mathematics students at A level were found to become more defective in their reasoning over time. This was characterised by a reduction in their endorsement of DA, AC and MT inferences. Evans et al. (2009) found that when participants were forced to rely on heuristic level processing, their endorsement of all four inferences decreased. It therefore seemed plausible that greater reliance on the heuristic level could be the source of mathematicians' defective reasoning.

A group of mathematics undergraduates and postgraduates and a group of non-mathematics undergraduates and staff at Loughborough University completed the Conditional Inference task twice: first with a total of 4 seconds to read and respond to each item, and second with a minimum of 6.5 seconds to read the problem before being allowed to respond. The fast time limit was shown

to be long enough for the heuristic level to process the problems (demonstrated by above-chance acceptance of MP inferences) and short enough to prevent the algorithmic level from influencing the response (demonstrated by a lack of correlation between RAPM and responding to the Conditional Inference task in the fast condition).

Hypothesis 1 stated that non-mathematicians would endorse fewer inferences in the fast condition than in the slow condition, as was the case with Evans et al.'s (2009) participants. This was partially supported: rather than a general decrease in endorsement rates, only valid items were endorsed less in the fast condition. This suggests that Type 2 processing is responsible for correctly endorsing MP and MT inferences when non-mathematicians reason without time pressure, while the heuristic level is responsible for incorrectly endorsing DA and AC inferences. This is also supported by the marginally higher endorsement of DA inferences in the fast condition.

Hypothesis 2 stated that mathematics students would respond more defectively than non-mathematicians in the slow condition. The mathematics students' Defective Conditional Index was marginally higher than the non-mathematicians', and the effect size was medium-large ($d = 0.69$). However, this difference was accounted for by the mathematics group's higher RAPM scores. The same was true for the Material Conditional Index, which was significantly higher in the mathematics group than in the non-mathematics group with a large effect size ($d = 1.24$), although only before RAPM had been taken into account. A comparison of RAPM scores in the mathematics group from this study and from the undergraduate and AS level studies suggested that the mathematics students investigated here had a higher mean level of intelligence, which could account for their high level of conformity to the material conditional.

In Chapter 5, mathematics AS level students were found to become increasingly defective in their reasoning over time compared to a control group, and to a lesser extent they conformed more to a Material Conditional interpretation over time. Here, this trend seems to have changed. It may be the case that at the start of post-compulsory mathematics study (i.e. AS level) students become less biconditional and more defective in their reasoning. In other words, they begin to endorse fewer DA, AC and MT inferences. Later in their mathematical study (i.e. at undergraduate level), it may be that students move more towards a material interpretation of the conditional, whereby they still endorse fewer DA and AC inferences, but revert back to correctly accepting more MT inferences.

In line with this suggestion, the data presented in Chapter 6 showed that first year mathematics undergraduates became more material in their interpretation of the conditional over time, although the interaction between Time and Group

was not significant. Furthermore, the third year undergraduate mathematics students in Chapter 7 had a slightly higher MCI than DCI in the ‘if then’ condition (see Figure 7.2). Alternatively, the contradiction between the AS students in Chapter 5 and the undergraduate students here could be due to a selection effect, whereby (on the whole) only the most intelligent A level students go on to study at degree level. These hypotheses would be best tested with a longitudinal study of a much longer duration than those presented in this thesis, for example, from the beginning of AS level until the end of undergraduate degree level, or even into postgraduate study.

Hypothesis 3 stated that if the heuristic level was responsible for the mathematics group’s more defective reasoning compared to the non-mathematics group, then the difference would remain in the fast condition, and the mathematics group’s Defective Conditional Index would not differ between the fast and slow conditions. Hypothesis 4 alternatively stated that if the heuristic level was not responsible, then the mathematics group would respond no more defectively than the non-mathematicians in the fast condition, and less defectively in the fast condition than in the slow condition.

Neither of these hypotheses could be conclusively ruled out based on the data presented here. On the one hand, mathematics students responded less defectively under time pressure than they did without time pressure, suggesting that the algorithmic level is at least partly responsible for this tendency, in support of Hypothesis 4. On the other hand, mathematics students responded marginally more defectively than the non-mathematics group under the fast condition as well as under the slow condition, suggesting that the heuristic level is in some part responsible for their tendency to respond defectively in line with Hypothesis 3. However, the mathematics group was only marginally more defective than the non-mathematics group under both time limits, and this may suggest that Hypothesis 4 has more support: the difference between the mathematics students’ Defective Conditional Index under the fast and slow conditions was significantly different with an effect size of $d = 0.54$, indicating that defective responding was half of a standard deviation higher in the slow condition. This could reflect the influence of Type 2 processing in the slow condition, either stemming from the algorithmic or reflective level.

Perhaps the mixed results indicate that the heuristic level is partly, but not entirely, responsible for mathematics students’ defective reasoning behaviour. The tendency to consider not- p cases irrelevant (the trademark of defective conditional reasoning) may begin as a conscious process and gradually become ingrained into the heuristic level, similar to the skill of driving (Lewin, 1982; Newstead, 2000). Alternatively, it could be that repeated exposure to forward inferences (as discussed in Chapter 5) creates an unconscious habit to consider

not- p cases irrelevant. This tendency could be reinforced by Type 2 thinking through a justification process as described by Evans (2006, 2011), which could in turn foster the defective tendency in subsequent problems, leading to the higher Defective Conditional Index in the slow condition. If this is the case, it could be that increasing conscious reflection eventually triggers mathematics students to move more towards material conditional responding later in their mathematics education.

It was suggested based on the findings of Evans et al. (2009) that mathematics students' defective reasoning could be due to greater reliance on heuristic level processing as opposed to a change in the nature of heuristic level processing. If this were the case, we would expect mathematicians to respond faster than non-mathematicians even without time pressure. This was not the case: there was no significant difference between the two groups' mean RTs in the slow condition. This is consistent with the hypothesis that any heuristic level differences between groups was due to the nature of the processing rather than on choosing Type 1 over Type 2 processing.

8.4.1 Summary of novel findings

1. The heuristic level is not entirely responsible for mathematics undergraduates' defective reasoning compared to a control group: mathematics students responded significantly less defectively when restricted to heuristic level processing.
2. The heuristic level is not entirely redundant either: the mathematics students were marginally more defective than the control group even when restricted to heuristic level processing.
3. It seems that both Type 1 and Type 2 processing play roles in mathematics students' conditional reasoning behaviour.

In some sense it is necessarily the case that Type 1 and Type 2 processing play roles in mathematics students' reasoning behaviour; the heuristic level is continually working to direct our attention to relevant aspects of our environment, and when time is not restricted Type 2 processing is also always engaged to some extent, even if only to approve the heuristically-generated response (Evans, 2011).

The next chapter investigates the potential for executive functions, i.e., the efficiency of algorithmic level processing, to be responsible for the differences between mathematics and non-mathematics students' reasoning behaviour. A group of non-mathematics undergraduates' executive function skills were measured, along with their Conditional Inference behaviour, in order to see whether

different interpretations of the conditional statement are associated with better or worse working memory, inhibition, or shifting skills.

Chapter 9

The role of executive functions in conditional reasoning ability

9.1 Introduction

It was established in Chapters 5 and (to a lesser extent) 6 that studying mathematics at advanced levels is associated with changes in logical reasoning skills, as the TFD suggests. However, rather than reasoning more normatively, it was found that participants studying mathematics increasingly rejected DA, AC and MT inferences and increasingly accepted MP inferences, making their reasoning more defective. A remaining question is what the mechanism of these changes might be. In Chapter 2 Stanovich's (2009a) tri-process model of cognition was identified as a useful starting point for narrowing the possibilities down. The tri-process model proposed that cognition happens via three levels: the heuristic level, the algorithmic level and the reflective level. The focus of this chapter will be on the algorithmic level: the computational capacity and efficiency available for effortful, conscious thinking.

It was argued in the literature review that general intelligence and executive functions are constructs that form part of the algorithmic level. The longitudinal study reported in Chapter 5 suggested that intelligence was not a mechanism of the increased defective reasoning, but the aim of this chapter is to assess the potential of executive functions to be a mechanism.

Executive functions regulate how we use our cognitive resources in order to complete a task (Banich, 2009; Stanovich, 2009a). There are thought to be three executive functions: working memory, inhibition, and shifting (Banich,

2009; Handley, Capon, Beveridge, Dennis & Evans, 2004; Gilhooly & Fioratou, 2009; Miyake et al., 2000). Working memory is the ability to hold and update information in the conscious mind. Individuals vary in how much information they can hold in mind, how accurately they can monitor this information and how effectively they can remove information no longer needed and replace it with new information. Inhibition is the ability to refrain from making unwanted responses and can be physical or cognitive. An example of physical inhibition would be not looking at a distraction when asked to focus your vision on a set point. An example of cognitive inhibition would be reciting your new phone number soon after changing it from a long-standing old one. Shifting refers to the ability to shift attention between different tasks being carried out simultaneously, for example, safely changing a CD whilst driving, or cooking whilst watching a film without losing track of either task.

The three executive functions have been shown to be clearly separable from each other and from intelligence (Ardila et al., 2000; Arffa, 2007; Friedman et al., 2006; Handley et al., 2004; Miyake et al., 2000). The study reported here investigated the contribution of each executive function to conditional reasoning behaviour in a group of non-mathematics undergraduates. If any relationships were found, it would indicate a potential mechanism via which studying mathematics might change reasoning behaviour. In particular, it might be the case that studying mathematics improves one's working memory, inhibition or shifting ability, and that this in turn leads to changed reasoning ability. For this to be the case, it is necessary that individual differences in executive functions are related to individual differences in reasoning behaviour. The study reported here will not test whether executive functions *are* the mechanism behind the TFD, only whether it is *possible* that they are.

Previous research has shown that working memory capacity is implicated in various forms of reasoning, including conditional reasoning (De Neys, 2006; Garcia-Madruga, Gutierrez, Carriedo, Luzon & Vila, 2007; Verschueren et al., 2005). Early research investigated conditional reasoning from the view of Baddeley & Hitch's (1974, 1986) model of working memory. According to the model, working memory consists of three components: the central executive, the phonological loop and the visuo-spatial sketchpad. The latter two components are short-term storage systems for verbal and visuo-spatial information, respectively, while the central executive controls attention and moderates the flow of information to and from the slave systems. Toms, Morris and Ward (1993) found that only the central executive component was recruited for conditional reasoning, while the visuo-spatial sketch pad and phonological loop were unrelated. They suggested that conditional reasoning requires an abstract working memory medium as opposed to a verbal or visuo-spatial one. This finding of

a relationship between conditional reasoning and working memory has been conceptually replicated in various studies. De Neys, Schaeken and d’Ydewalle (2005) found that participants with high working memory spans reasoned more in line with the material conditional than participants with low working memory spans. Furthermore, Verschueren et al. (2005) found that working memory capacity was not only related to reasoning performance, but it determined whether a reasoner would use a probabilistic (low capacity) or counterexample (high capacity) strategy to solve problems. However, the relationship may not be straightforward: Handley, Capon, Copp and Harper (2002) found that performance on a conditional reasoning task was only related to verbal working memory, not spatial working memory. It appeared from their findings as though successful conditional reasoning was dependent on the simultaneous processing and storage of verbal representations, which contradicts Toms et al.’s (1993) findings.

While it is well established that working memory is in some way important for reasoning, research on the inhibition and shifting aspects of executive functions and their relation to reasoning ability is more sparse. To the best of my knowledge, the relationship between conditional reasoning and shifting ability has not yet been investigated and few studies have investigated the role of inhibition in reasoning. In a study of 10-year-old children’s reasoning ability, Handley et al. (2004) found that while reasoning with belief-based problems was related to both working memory and inhibition skills, belief-neutral reasoning was only related to working memory. Similarly, Markovits and Doyon (2004) found that ability to inhibit irrelevant information was related to success in solving thematic conditional reasoning problems, as well as, and separately from, working memory. Inhibition, then, may only be a necessary cognitive tool when a reasoner needs to decontextualise a problem in order to solve it, i.e. when they need to inhibit their prior beliefs. If this is the case then the abstract Conditional Inference task used throughout this thesis, and in the study reported in this chapter, may not require inhibition skills, but it seems likely that performance will be related to working memory.

Moving beyond overall Conditional Inference task performance, the study presented here offers an opportunity to investigate the Negative Conclusion and Affirmative Premise biases, and several hypotheses can be formulated. Negative Conclusion Bias (NCB) is the tendency to draw more inferences with negative conclusions than with positive conclusions and is most often observed on DA and MT inferences (Schroyens et al., 2001). NCB has been explained in two ways which can be tested here. One suggestion is that NCB is a heuristic bias, whereby reasoners assume that ‘not p ’ is more likely to be true than ‘ p ’ (there are more non-human things than there are humans, and more non-tables than there are tables, for example) and so are more willing to accept it (Pollard &

Evans, 1980; Oaksford et al., 2000). Alternatively, NCB may be a bias at the level of Type 2 processing. Evans et al. (1995) has proposed that the problem lies with the double negation inherent in DA and MT inferences with affirmative conclusions. When making an MT inference from the conditional ‘If A then 3’, the premise ‘not 3’ leads to the negative conclusion ‘not A’. However, when making the same inference from the conditional ‘If not A then 3’, an extra step is involved to reach the affirmative conclusion that is necessary: ‘not 3’ implies ‘not (not A)’, which needs to be converted into ‘A’. Evans et al. (1995) argued that reasoners do not easily see the equivalence of ‘not (not p)’ and ‘ p ’, and that this is the source of difficulty. The quadratic relationship found between NCI and DCI in Chapter 5 suggests that the latter explanation is more likely to be the case - NCI appeared to be a by-product of systematic reasoning rather than a heuristic process.

These two competing hypotheses make different predictions for the relationship between NCB and executive functions. If NCB stems from a heuristic level process, then better inhibition may be related to greater success at avoiding it. If, on the other hand, NCB stems from a struggle to complete the extra logical step necessary to make MT and DA inferences with negative conclusions, then we might expect better working memory to be associated with a lower rate of the bias.

The Affirmative Premise Bias (APB) is the tendency to accept more inferences from affirmative premises than from negative premises, particularly when the negative premise is implicit. For example, the inference ‘if not-A then 3; A; therefore not-3’ is accepted more often than the inference ‘if A then 3; D; therefore not-3’, even though they are both invalid DA inferences. This may be due to a matching bias in Type 1 processing, whereby the premise ‘A’ is more obviously related to the conditional than is the premise ‘D’ (Evans & Handley, 1999). As with NCB, if APB is indeed due to a Type 1 processing error, then the ability to avoid it may be related to inhibition.

To summarise, there are several predictions for the current study:

1. The three components of executive function will be unrelated.
2. Working memory will be positively correlated with the MCI.
3. Given the abstract nature of the task, inhibition will not be related to reasoning behaviour.
4. NCB will be related to inhibition OR working memory scores.
5. APB will be related to inhibition scores.

There is no suggestion from previous research that shifting scores will (or will not) be related to performance on an abstract Conditional Inference task, so this aspect is exploratory.

9.2 Method

9.2.1 Participants

Ninety-four undergraduate students from various engineering courses took part unpaid during a lab session for a second year introductory statistics module. They later analysed the data for their coursework assignment. All participants gave informed consent for their data to be used for research purposes.

9.2.2 Measures

There were four measures, one for each of working memory, inhibition, shifting, and conditional reasoning. All tasks were programmed and administered in E-Prime version 2.0 (Schneider et al., 2002).

Working memory

Working memory was measured using a 2-back task. Participants saw a string of letters presented sequentially and were instructed to press a ‘Target’ key when the letter on screen matched the one presented two letters back. For all other letters, they pressed a ‘Not a target’ key. An example sequence of trials is shown in Figure 9.1. There were 90 trials, 30 of which were targets, and the task was preceded by a practice session of 20 trials. Letters were presented for 500ms with 1000ms gaps between letters. The measure taken was mean accuracy.

Inhibition

A version of the Stroop task (Stroop, 1935) was used as a measure of cognitive inhibition. In a no conflict condition, participants saw strings of @ symbols presented in coloured fonts and were instructed to identify the colour of the font. In a conflict condition, participants saw colour names presented in a different coloured font, for example, the word ‘blue’ presented in a red font, and were again instructed to identify the colour of the font, ignoring the conflicting word. Conditions were blocked and each consisted of 40 trials preceded by 10 practice trials. Five colours were used - red, blue, green, yellow and purple. Participants responded to each trial by pressing one of five keys which were identified with coloured stickers. Participants were instructed to respond as quickly and accurately as possible and the stimuli was displayed until response.

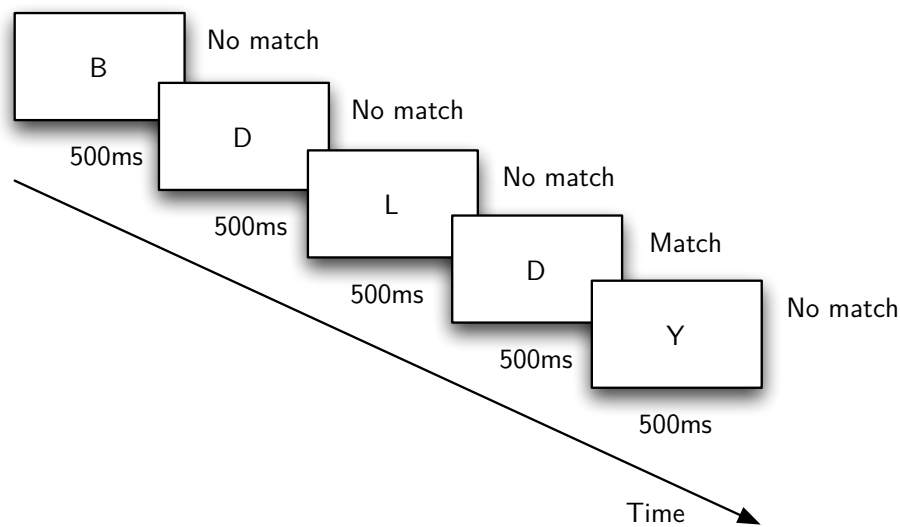


Figure 9.1: Example sequence of trials from the 2-back task

Inhibition ability was calculated as total time taken to complete the conflict list minus total time taken to complete the no conflict list (Gilhooly & Fioratou, 2009; Miyake et al., 2000). Higher scores therefore reflect a greater cost of dealing with conflict, and hence poorer inhibition ability.

Shifting

Shifting ability was measured with a letter-number categorisation task. Participants saw letter-number pairs and in one list identified whether the letter was a vowel or consonant, in another list identified whether the number was odd or even, and in a third list switched between the two tasks. Each list consisted of 64 trials preceded by 5 practice trials. Each stimulus was displayed until response with a gap of 500ms between trials. A measure of shifting ability was calculated as mean response time (RT) on switching trials minus mean RT on non-switching trials. Higher scores therefore reflect less efficient shifting ability.

Conditional Reasoning

The Conditional Inference task (Evans et al., 1995; Inglis & Simpson, 2009a) was used as a measure of reasoning ability for consistency with the longitudinal studies in which development alongside mathematical study was found (see Section 4.3.2 for details of the task). In this study, the task was administered by computer as opposed to pen and paper. Items were presented sequentially and the order was randomised between participants. Each item remained on

screen until response with 500ms gaps between trials. Participants responded by pressing the ‘S’ or ‘L’ key on a standard keyboard, labelled as ‘No’ and ‘Yes’ respectively. Four interpretation indices, MCI, DCI, BCI and CCI, and two bias indices, NCI (the measure of NCB) and API (the measure of APB), were calculated from the task as in previous chapters.

9.2.3 Procedure

Participants took part in a computer lab in groups of approximately 30. They worked alone and in silence. The reasoning task was always completed first so that the speeded nature of the executive function tasks would not interfere with performance on the non-speeded reasoning task. The executive function tasks were presented in a set order: working memory, inhibition, then shifting. The sessions lasted approximately 35 minutes.

9.3 Results

9.3.1 Data cleaning

One participant’s score on the inhibition task was deleted for being more than three standard deviations above the mean (representing unusually poor performance). All other data fell between the task mean and ± 3 standard deviations.

9.3.2 Task performance

Performance across all tasks was as expected. Note that on the working memory task higher scores reflect better performance, whereas with the inhibition and shifting tasks higher scores reflect a greater cost of conflict/shifting, and hence poorer skills. Shifting scores, $M = 712.09ms$, $SD = 350.30$, were significantly above zero, $t(93) = 19.76$, $p < .001$, indicating that on average participants were slower on the switching trials than the non-switching trials, as expected (Figure 9.2). Inhibition scores, $M = 4829.10ms$, $SD = 6009.47$, were also significantly above zero, $t(92) = 7.75$, $p < .001$, demonstrating that the word/font colour conflict slowed participants’ responses down, again as expected (Figure 9.3). Lastly, working memory scores (proportion of items correctly categorised), $M = 0.81$, $SD = 0.08$, were also significantly above the 50% level, $t(93) = 39.02$, $p < .001$ (Figure 9.4). This set of results demonstrates that participants were engaging with the tasks and showing the patterns of performance expected from the literature.

On the conditional inference task, the interpretation index with the highest mean was the BCI ($M = 20.28$, $SD = 5.99$), followed by the CCI ($M =$

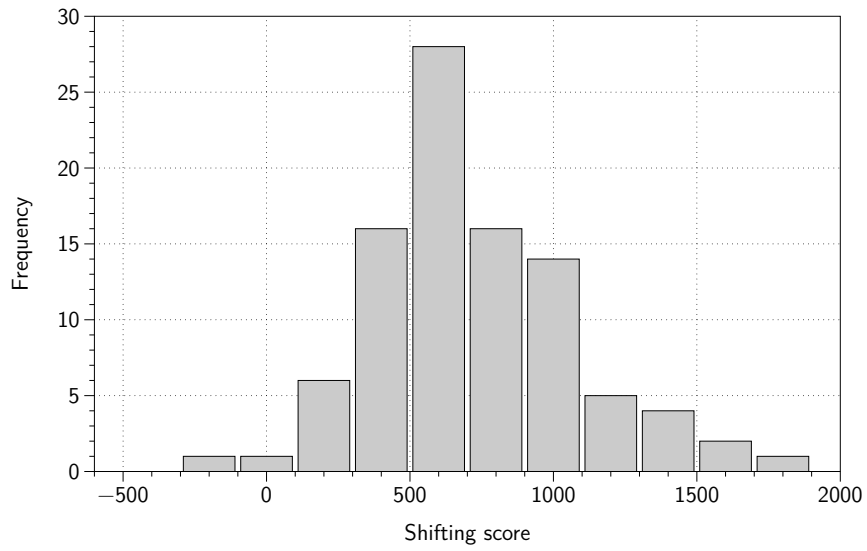


Figure 9.2: Distribution of scores on the shifting task.

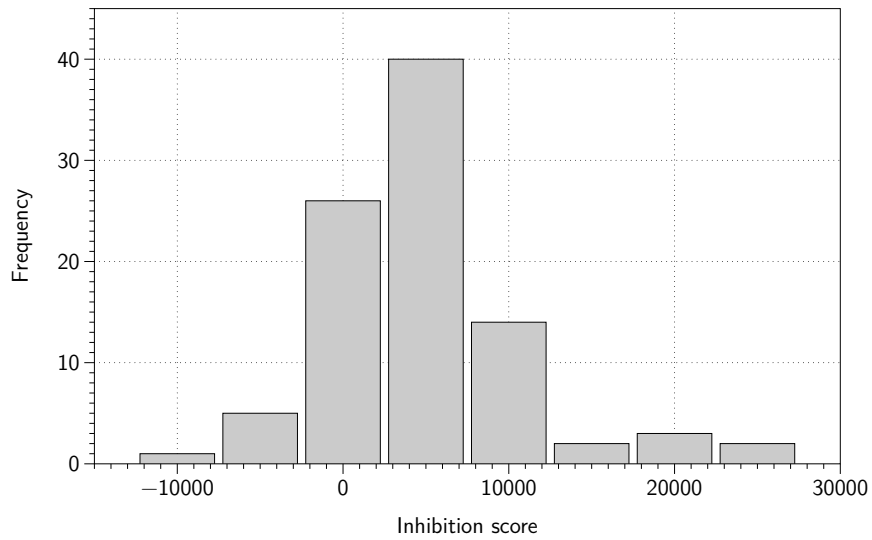


Figure 9.3: Distribution of scores on the inhibition (Stroop) task.

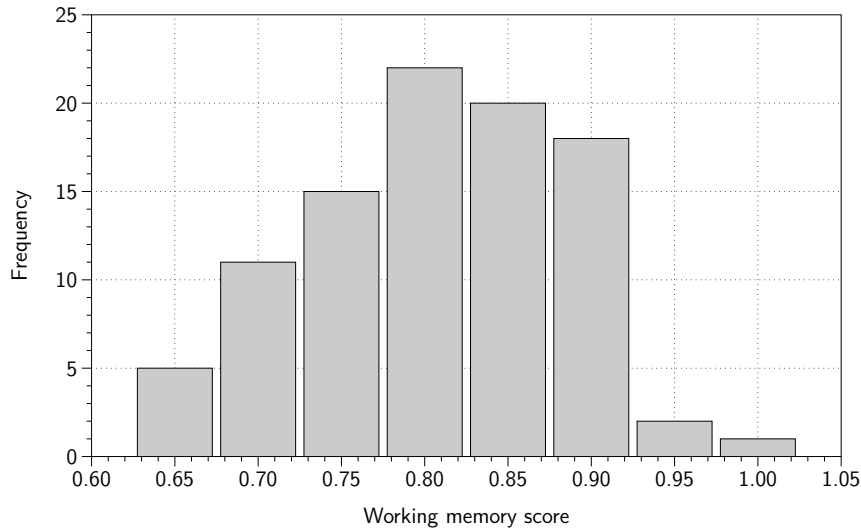


Figure 9.4: Distribution of scores on the working memory (2-back) task.

19.81, $SD = 3.13$), then the MCI ($M = 19.47, SD = 3.37$), and lastly the DCI ($M = 18.04, SD = 5.61$). This is not surprising given previous findings in this thesis that non-mathematics students tend to show a more biconditional interpretation than either a defective or material interpretation.

9.3.3 Relationships between executive functions

The relationships between the three executive function measures were analysed with Pearson correlations and are summarised in Table 9.1. Hypothesis 1 stated that the three components of executive function would not be related to each other. Working memory was not significantly correlated with inhibition ability, $r(93) = .02, p = .827$, nor with shifting ability, $r(94) = -.11, p = .294$. Inhibition and shifting scores were also not significantly correlated, $r(93) = .17, p = .104$. This is consistent with previous findings that the three components of executive function are separable skills (Ardila et al., 2000; Arffa, 2007; Friedman et al., 2006; Handley et al., 2004; Miyake et al., 2000).

	Inhibition	Shifting
Working Memory	.02	-.11
Inhibition	-	.17

Table 9.1: Correlations between Executive Functions. All $ps > .1$.

	Inhibition	Shifting	Working Memory
MCI	-.13	-.12	.34**
DCI	-.18 [†]	-.18 [†]	.12
CCI	-.13	.01	-.12
BCI	.21*	.22*	-.08

Table 9.2: Correlations between Executive Functions and interpretation indices. [†] $p < .1$, * $p < .05$, ** $p < .01$.

9.3.4 Executive functions and Conditional Reasoning

Interpretation Indices

The relationships between the three executive function measures and four interpretation indices were investigated with Pearson correlations and are summarised in Table 9.2. Hypothesis 2 stated that working memory scores would be positively correlated with the MCI, and this was supported by the data, $r(94) = .34, p = .001$. Those with better working memory had higher MCI scores (Figure 9.5).

Hypothesis 3 stated that inhibition scores would not be related to the interpretation indices due to the abstract nature. Contrary to this, inhibition scores were significantly positively correlated with the BCI, $r(93) = .21, p = .041$, and marginally negatively correlated with the DCI, $r(93) = -.18, p = .083$, suggesting that better inhibition (represented by a lower score) was associated with a lower BCI and a marginally higher DCI. These two correlation coefficients were marginally different, $t(91) = 1.93, p = .056$. This could be because in non-mathematicians the biconditional interpretation stems from the heuristic level of cognition, and when inhibited it is replaced with a defective conditional interpretation. In previous studies there was no indication of a relationship between inhibition and abstract conditional reasoning behaviour, but this may be because previous studies only looked at consistency with the material interpretation and didn't consider a BCI or DCI. It may be the case that taking a material interpretation depends on other factors (e.g. working memory) whereas the BCI and DCI are related to inhibition in the manner suggested above.

There were no firm predictions about the relationship between shifting and conditional reasoning, but shifting was found to be significantly positively correlated with the BCI, $r(94) = .22, p = .036$, and marginally negatively correlated with the DCI, $r(94) = -.18, p = .077$. This suggests that better shifting ability (represented by a lower score) was associated with a less biconditional interpretation of the conditional, and a marginally more defective interpretation, mirroring the relationships found for inhibition.

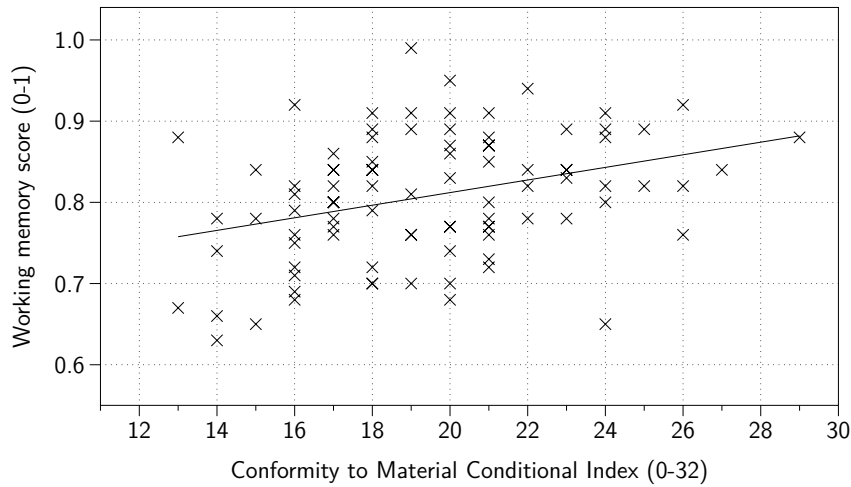


Figure 9.5: Correlation between MCI and working memory scores.

Biases in Conditional Reasoning

Lastly, the relationships between participants' three executive functions and their NCI and API were examined, again with Pearson correlations. Hypothesis 4 stated that NCB would be related to inhibition or working memory scores, and Hypothesis 5 stated that APB would be related to inhibition scores.

NCIs were significantly negatively correlated with shifting scores, $r(93) = -.26, p = .013$, indicating, rather counterintuitively, that those with better shifting ability had *higher* NCIs. An intriguing possibility is that the quadratic relationship between DCI and NCI found in Chapter 5 was driven in some part by shifting skills, i.e. better shifting ability is associated with higher NCIs and a more defective interpretation of the conditional (marginally significant here) meaning that NCIs tend to be higher in people with a more defective interpretation of the conditional, at least up to a point.

NCIs were not correlated with working memory scores, $r(93) = .08, p = .472$, nor with inhibition scores, $r(92) = -.07, p = .493$, contrary to Hypothesis 4. APIs were significantly negatively correlated with shifting ability, $r(94) = -.21, p = .043$, where those with better shifting ability showed higher APIs, similarly to NCIs. APIs were marginally negatively correlated with inhibition, $r(92) = .19, p = .066$, suggesting that those with better inhibition ability tend to show a greater API, contrary to prediction. Lastly, APIs were not correlated with working memory scores, $r(93) = .07, p = .485$.

9.4 Discussion

The aim of this chapter was to investigate the potential for the algorithmic level of cognition to be a source of change in conditional reasoning behaviour. It was suggested that the algorithmic level might be a mechanism via which the TFD could operate, but for this to be the case it would be necessary for there to be a relationship between measures of the algorithmic level and measures of reasoning behaviour. The study presented in Chapter 5 suggested that intelligence, at the algorithmic level, was not a mechanism for the changes found in the mathematics students' reasoning behaviour. Here, the three executive functions of working memory, inhibition and shifting were used as further measures of the algorithmic level. The Conditional Inference task used in Chapters 5 and 6, which mathematics students were found to significantly change on during AS level, was used as a measure of reasoning ability. Performance on the executive function tasks was as expected: the distributions of scores were in line with previous studies and the three skills were found to be clearly separable (Hypothesis 1).

Hypothesis 2 predicted that working memory would be positively correlated with the MCI, and this was the case. Better working memory was associated with a more material interpretation of the conditional, confirming the findings of previous studies (e.g. De Neys et al., 2005; Verschueren et al., 2005). The other interpretations did not appear to be related to working memory and MCI was not correlated with inhibition or shifting skills. This could reflect that, rather than relying on the inhibition of intuitive responses in favour of an available alternative interpretation, a material interpretation relies on having the necessary working memory capacity to apply or calculate the alternative.

Inhibition and shifting were found to be positively correlated with the BCI and marginally negatively correlated with the DCI. This could indicate that the BCI is the intuitive interpretation of the conditional statement, and that those with better inhibition and shifting ability are more able to inhibit this response in favour of a defective interpretation. This has implications for the TFD: it is possible that when mathematics students become less biconditional and more defective in their reasoning patterns (as found in the AS level students in Chapter 5), it is because their inhibition and shifting skills have improved. This is not incompatible with the hypothesis put forward in Chapter 5 which suggested that exposure to forward ('if then') inferences encourages students to consider not- p cases irrelevant. Perhaps it is this exposure which trains mathematics students to inhibit a biconditional interpretation.

The results related to the NCI and API scores were surprising. For NCI it was hypothesised that either inhibition (if NCB is a Type 1 bias) or working

memory scores (if NCB is a Type 2 bias) would be correlated, yet neither was. For API it was hypothesised that inhibition scores would be correlated, and they were not. However, it may be the case that the type of inhibition measured here is not the same type of inhibition required to avoid NCB and APB. In executive function tasks participants are told how to perform optimally and the measure reflects how well they can do so (Stanovich, 2009a). In the case of the inhibition task used here, participants were told to inhibit the interference of font colour on their responses and the measure was the RT cost of doing so. The spontaneous inhibition of Type 1 errors by Type 2 processing intervention when solving a reasoning task may have an altogether different nature. This likely depends on the thinking dispositions of the reasoner at the reflective level of cognition rather than the efficiency of implementing inhibition at the algorithmic level (Stanovich, 2009a). A more effective way to test for the relationship between inhibition and NCB and APB would be to use a measure of reflective-level inhibition such as the Cognitive Reflection Test (Toplak et al., 2011).

Such an analysis is possible in the data presented in Chapter 6 where another set of undergraduate participants completed the CRT and the Conditional Inference Task. A reanalysis of this data showed that reversed intuitive scores on the CRT were significantly negatively correlated with APB, $r(55) = -.30, p = .028$, indicating that those with better inhibition of intuitive responses to the CRT had lower APB scores. Conversely, there was not a significant correlation between reversed intuitive scores to the CRT and NCB scores, $r(55) = -.18, p = .187$. These additional analyses suggest that APB is indeed a heuristic bias, while there is no evidence of this being the case for NCB.

It is important to note that the results of this chapter do not support the idea that executive functions *are* responsible for changes to mathematics students' reasoning behaviour, simply that they *could* be. The longitudinal study reported in Chapter 5 suggested that thinking disposition (measured by CRT scores) was not a factor in changes to reasoning behaviour, but that it was a predictor of starting reasoning ability. It may be the case that shifting and inhibition play a similar role, influencing a person's baseline reasoning ability but not playing a role in development alongside mathematical study. In order to investigate this further, another longitudinal study would be required in which mathematics and non-mathematics students' executive functions are measured at the start and end of a period of study. Nevertheless, the results of this chapter do put forth some evidence that executive functions could be an interesting place to look for the mechanisms of reasoning change.

9.4.1 Summary of novel findings

1. Inhibition and shifting were positively correlated with conformity to a biconditional interpretation and marginally negatively correlated with conformity to the defective conditional interpretation of conditional statements.
2. Affirmative premise bias was not related to an executive function measure of inhibition, but it was related to a reflective-level measure of inhibition, the CRT, in a reanalysis of a previous dataset.

Chapter 10

Conclusions

10.1 Introduction

For millenia it has been assumed that studying mathematics improves general reasoning skills (Plato, 2003/375B.C; Locke, 1971/1706; Smith, 2004; Walport, 2010; Oakley, 1949). This idea is known as the Theory of Formal Discipline (TFD). Previous research suggested that there may be a relationship between advanced mathematics and conditional reasoning skills (Lehman & Nisbett, 1990; Inglis & Simpson, 2008), although it remained unclear whether the relationship was developmental (Inglis & Simpson, 2009a). This thesis aimed to do two things: establish whether studying mathematics at advanced levels is associated with changes to reasoning skills, and if so, to investigate some potential mechanisms for the changes found. The findings of my studies suggest that studying mathematics is associated with a less biconditional and a more material, and in particular, more defective, interpretation of conditional statements. This appeared to be limited to abstract ‘if then’ conditional reasoning. The source of the change is discussed further below, after a summary of each study’s findings.

10.2 Overview of findings and interpretations

Chapter 5 documented a one-year longitudinal study with students taking mathematics and English literature AS levels. In the UK, AS level is the first stage of post-compulsory study, and students can choose which subjects they would like to study (typically four). The students who had chosen to study mathematics and those who had chosen English literature did not differ on a Conditional Inference task (validating deductions from statements of the form ‘if p then q ’) at the start of their studies. By the end of their year of study, however,

the mathematics students had become significantly less biconditional and significantly more material and defective in their interpretation of the conditional statement, while the English literature students did not change. An examination of the effect sizes revealed that the change in the defective conditional index ($d = .880$) was notably greater than either of the other changes. Because general intelligence and thinking disposition were controlled for, it seems probable that these changes were driven by the participants' mathematical study. It seems that studying mathematics teaches students to assume p , and to therefore consider not- p cases irrelevant. As discussed in Section 2.3 on rationality, the material conditional interpretation is the normative standard (considered correct by logicians), and the defective interpretation, while not being the normative model, could be considered as 'better' than the biconditional interpretation. In this sense it could be said that the mathematics students improved in their reasoning with conditional statements. They did not, however, change in their reasoning with thematic syllogisms, which further supports the idea that the change was in the students' interpretation of 'if'.

The explanation proposed for this finding was that mathematics regularly exposes students to implicit conditional statements (the AS level syllabus contains no explicit reference to conditional logic) where they are expected to assume p and reason about q . Houston (2009) argued that most mathematical statements are of the form 'if statement A is true, then statement B is true', even if they are heavily disguised (p. 63). He also argued that in mathematics A is assumed to be true, even when it is clearly not true (for example, in the case of contradiction proofs), and the truth or falsity of B is then deduced. For this reason, Hoyles and Kuchemann (2002) argued that the defective interpretation is actually more appropriate for mathematics classrooms than the material interpretation. In line with this idea, Inglis and Simpson (2009b) found that a group of successful undergraduate mathematics students tended to have a more defective than material interpretation of conditionals.

The AS study was modified and repeated with first year undergraduate students in Chapter 6, and the pattern of change was replicated. Mathematics students' interpretation of the conditional became more material ($d = .55$), more defective ($d = .66$), and less biconditional ($d = -.46$) over time. However, psychology undergraduates showed a similar pattern of change and there was also an issue of low statistical power. As a result, the pattern of change from the AS level students could only be detected in the undergraduate students with t-tests comparing each interpretation index across time.

Rather than a thematic syllogisms task, the participants in Chapter 6 completed a thematic Conditional Inference task. This allowed the breadth of the change in a defective interpretation to be investigated. Intriguingly, the math-

ematics students showed no changes in any of the interpretation indices over time. Although statistical power was again a problem, there were not even trends for change. This suggests that studying mathematics may only change students' interpretation of abstract conditional statements.

In Chapter 7, third year undergraduate mathematics students completed the Conditional Inference task with the conditional phrased as either 'if p then q ' or as ' p only if q '. These two forms of the conditional are logically equivalent, but previous research had shown non-mathematics students to treat them differently (Evans, 1977). Chapters 5 and 6 suggested that the effect of studying mathematics on reasoning skills may be limited to abstract conditional statements. It was hypothesised that this could be because students are repeatedly exposed to (implicit) 'if p then q ' statements where they are to assume that p is true and reason about q . Evans (1977) argued that 'if then' and 'only if' statements are treated differently because necessity and sufficiency are differentially emphasised in each. While 'if p then q ' statements emphasise the sufficiency of p , ' p only if q ' emphasises the necessity of q . If mathematics students do learn to assume that 'if p then q ' means that p is true and they need to reason about q , as opposed to holding a general defective interpretation of conditional statements, then the necessity of q emphasised in 'only if' statements could disrupt their reasoning and cause them to respond less defectively. This was the effect found. The students who completed the 'only if' version of the task responded more biconditionally and less defectively and materially than their peers in the 'if then' condition. The mathematics students' endorsement rates in the 'only if' condition appeared to be similar to the non-mathematics students' endorsement rates in Evans's (1977) study. This suggests that mathematics students' defective interpretation of the conditional is limited to 'if then' statements, and may indeed stem from exposure rather than from a general change in their understanding of conditionals.

The potential of the heuristic level of cognition to be the mechanism of the change in mathematics students' reasoning skills was investigated in Chapter 8. Mathematics and non-mathematics undergraduate students completed the standard abstract Conditional Inference task twice: once with a very short time limit for each item and once with as much time as they liked. The time limit was shown to be long enough for the heuristic level to generate a response but short enough to prevent Type 2 thinking from interfering. Under this restriction, mathematics students responded significantly less in line with a defective interpretation of the conditional than under the free time condition, but they were still marginally more defective than the non-mathematics students. This suggests that the defective interpretation does stem, in part, from the heuristic level, but that Type 2 thinking plays a significant role in encouraging it.

These findings fit with the hypothesis proposed above, that mathematics students are exposed to implicit ‘if then’ statements which means they learn to assume p and as a consequence hold a defective understanding of the conditional. Repeated exposure to an implicit stimulus could over time bring about changes in the heuristic level, in the same way that word and meaning comprehension in general becomes automatic when we learn to read (LaBerge & Samuels, 1974). An unconscious habit to consider not- p cases irrelevant could be reinforced by Type 2 thinking through a justification process (Evans, 2006, 2011), whereby the output of the heuristic level is justified rather than scrutinised. This in turn could foster the defective tendency in subsequent problems (“I responded like this before so I probably think the same about this question”), leading to the higher Defective Conditional Index in the slow condition.

This leaves open the possibility that increasing conscious reflection or added ‘rules’ (such as the material definition of the conditional) eventually cause the mathematics students to move more towards material conditional responding later in their mathematics education. An alternative explanation is that the material mathematicians actually have a defective interpretation but are able to use a contradiction proof to endorse MT inferences, making their responding conform to the material conditional.

Finally, Chapter 9 investigated the role of executive functions (at the algorithmic level of cognition) in conditional reasoning. Although the participants were non-mathematics students, this allowed an indication of whether there was any potential for executive functions (shown by a relationship with conditional reasoning behaviour) to be the mechanism of change. Firstly, it was shown that working memory was related to the material conditional index, which replicates previous findings. The novel findings were that better inhibition and shifting were related to less biconditional responding and marginally more defective responding. It was suggested that this could be because the biconditional interpretation is the intuitive one, and when this can be inhibited it tends to be replaced by the defective interpretation.

This seems to contradict the findings from Chapter 8, by suggesting that the defective interpretation is implemented at the level of Type 2 thinking rather than at the heuristic level. However, the proposal outlined above was that studying mathematics drives the defective interpretation into the heuristic level, and the participants in the executive function study were from non-mathematics degree courses. It could be that in the general population the biconditional interpretation is the intuitive one, and that high IQ participants are able to respond more defectively (Evans et al., 2007) because they are more successful at inhibiting the heuristic response and implementing a Type 2 level response. In mathematicians, the biconditional interpretation might gradually be replaced by

a defective interpretation as the intuitive standard, and this is simply reinforced at the level of Type 2 thinking.

The findings of the executive function study suggest that the change in mathematicians might initially come about through greater inhibition of the biconditional interpretation. Perhaps exposure to ‘if then’ statements encourages students to inhibit a biconditional interpretation because it is proved inappropriate through the examples they are exposed to. The repeated experience of assuming p and reasoning about q also begins to replace the biconditional interpretation held in the heuristic mind. Eventually, the defective interpretation becomes intuitive and is further reinforced with Type 2 thinking through a justification process. At some point, explicit instruction in conditional logic may provide students with the knowledge necessary for a material interpretation, and inhibition may again play a role in inhibiting the defective interpretation in favour of the newly learned material interpretation.

10.3 Future research

The research presented here has opened up several questions that require further research. Firstly, there was no evidence that studying mathematics has any impact on reasoning beyond abstract ‘if then’ conditional reasoning. Nevertheless, it would be worthwhile to investigate further types of deductive reasoning skills, such as disjunctive reasoning, and informal reasoning skills, such as application of the law of large numbers and other skills described in Chapter 4. It is possible that different areas of mathematics impact on different types of reasoning skills, and some of the skills affected were not measured by the conditional and syllogisms tasks used here.

In a similar vein, it would also be worthwhile to isolate the impact of different areas of mathematics on abstract conditional reasoning skills. It might be the case that, for example, geometry learning is not related to changes in conditional reasoning behaviour while calculus learning is. A first attempt at this difficult task could be to examine the development of conditional reasoning skills across different curriculums, for example, in Cyprus where there is a large emphasis on geometry-based mathematics.

An important avenue of investigation for future work is to see whether the findings from the AS level study would replicate in an educational system where it is compulsory for students to study some form of mathematics until the age of 18. The students in the AS level (and undergraduate) study had chosen to stay in education and had chosen which subjects to study. It could be that this disposition is essential for the effects found. This results from the quasi-experimental design of the study, which was unavoidable in this case.

However, replicating the study in a compulsory-mathematics cohort would help to alleviate the confound of the desire to study mathematics.

Based on the findings from this thesis it was suggested that there may be a developmental trend in the relationship between studying mathematics and changes in conditional reasoning behaviour, whereby students go from having a biconditional view of the conditional to a defective view, and finally to a material view. This trend, along with the proposed mechanisms (changes to the heuristic level of cognition which are reinforced at the algorithmic level of cognition), should be tested. This could be done either with a longitudinal study following students from AS level up to the end of degree level, and where possible, into postgraduate study. Alternatively, and far more economically, the trend could be investigated with a cross-sectional study comparing groups of mathematicians and non-mathematicians from every stage between AS level and academic staff level. This would allow us to see whether, and at which stage, mathematicians change from having a defective interpretation of the conditional to a material interpretation of the conditional.

10.4 The Theory of Formal Discipline revisited

This thesis was prompted by the TFD, which has claimed, without evidence, that studying mathematics improves a person's general reasoning skills. The evidence found here is very limited in its support of the TFD. While mathematics students did 'improve' in abstract conditional reasoning based on 'if then' statements, they did not become straightforwardly more normative, and they were not found to improve on thematic 'if then' reasoning, abstract conditional reasoning of the form 'only if', or on thematic syllogisms. Based on this evidence the TFD appears to have been greatly overstated in the past. Such quotes as "Through mathematics we also wish to teach logical thinking – no better tool for that has been found so far" from Amitsur (Sfard, 1998, p. 453) and "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" from Oakley (1949, p. 19) are not supported by the research presented here: the terms 'logical thinking' and 'rationality' are surely meant to refer to behaviour beyond abstract 'if then' reasoning. This also has implications for mathematics education policy. The TFD has been used as an argument for mathematics to be prioritised in the UK National Curriculum in several ways, for example by Smith (2004), who said that studying mathematics "disciplines the mind, develops logical and critical reasoning, and develops analytical and problem-solving skills to a high degree" (p. 11). Further claims of this nature might best be withheld until they can be supported by evidence.

Publications List

- Alter, A. L., Oppenheimer, D. M., Epley, N. & Eyre, R. N. (2007). Overcoming intuition: Metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General*, 136, 569-576.
- American Council on Education. (1953). *Test of critical thinking: Instructor's manual*. Washington, D.C: American Council on Education.
- Anderson, A. R. & Belnap, N. D. (1975). *Entailment: The logic of relevance and necessity*. Princeton, NJ: Princeton University Press.
- Ardila, A., Pineda, D. & Rosselli, M. (2000). Correlation between intelligence test scores and executive function measures. *Archives of Clinical Neuropsychology*, 15, 31-36.
- Arffa, S. (2007). The relationship of intelligence to executive function and non-executive function measures in a sample of average, above average, and gifted youth. *Archives of Clinical Neuropsychology*, 22, 969-978.
- Aronson, J., Lustina, M. J., Good, C. & Keough, K. (1999). When white men can't do math: necessary and sufficient factors in stereotype threat. *Journal of Experimental Social Psychology*, 35, 29-46.
- Ashenfelter, O. & Rouse, C. (1999, January). *Schooling, intelligence, and income in America: Cracks in the Bell curve* (Working Paper No. 6902). Princeton: National Bureau of Economic Research. Available from <http://www.nber.org/papers/w6902>
- Baddeley, A. & Hitch, G. (1974). The psychology of learning and motivation: advances in research and theory. In G. H. Bower (Ed.), (Vol. 8, p. 47-89). New York: Academic Press.
- Baddeley, A. D. (1986). *Working memory*. Oxford: Clarendon.
- Banich, M. T. (2009). Executive function: the search for an integrated account. *Current Directions in Psychological Science*, 18, 89-94.
- Baron, J. & Hershey, J. C. (1988). Outcome bias in decision evaluation. *Journal of Personality and Social Psychology*, 54, 569-579.
- Beth, E. W. & Piaget, J. (1966). *Mathematical epistemology and psychology*. Dordrecht: D. Reidel.

- Bors, D. A. & Vigneau, F. (2003). The effect of practice on Raven's Advanced Progressive Matrices. *Learning and Individual Differences*, 13, 291-312.
- Braine, M. D. S. (1978). On the relation between the natural logic of reasoning and the standard logic. *Psychological Review*, 85, 1-21.
- Brainerd, C. J. & Reyna, V. F. (2002). Fuzzy-trace theory and false memory. *Current Directions in Psychological Science*, 11, 164-169.
- Bramall, S. (2000). Rethinking the place of mathematical knowledge in the curriculum. In S. Bramall & J. White (Eds.), *Why learn maths?* London: Institute of Education University of London.
- Bramall, S. & White, J. (Eds.). (2000). *Why learn maths*. London: Institute of Education University of London.
- Cacioppo, J. T., Petty, R. E., Feinstein, J. A. & Jarvis, W. B. G. (1996). Individual difference in cognitive motivation: the life and times of individuals varying in need for cognition. *Psychological Bulletin*, 119, 197-253.
- Cacioppo, J. T., Petty, R. E. & Kao, C. F. (1984). The efficient assessment of Need for Cognition. *Journal of Personality Assessment*, 48, 306-307.
- Chen, S. & Chaiken, S. (1999). Dual process theories in social psychology. In S. Chaiken & Y. Trope (Eds.), (p. 73-96). New York: Guilford.
- Cheng, P. W. & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology*, 17, 391-416.
- 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.
- Christensen, L. B. (2000). *Experimental methodology* (Seventh ed.). Allyn and Bacon.
- Cosmides, L. (1989). The logic of social exchange: has natural selection shaped how humans reason? studies with the Wason selection task. *Cognition*, 31, 187-316.
- Cosmides, L. & Tooby, J. (1992). Cognitive adaptations for social exchange. In J. H. Barkow, L. Cosmides & J. Tooby (Eds.), *The adapted mind*. Oxford: Oxford University Press.
- Cummins, D. D. (1995). Naive theories and causal deduction. *Memory & Cognition*, 23, 646-658.
- De Neys, W. (2006). Automatic-heuristic and executive-analytic processing during reasoning: Chronometric and dual-task considerations. *The Quarterly Journal of Experimental Psychology*, 59, 1070-1100.
- De Neys, W., Schaeken, W. & d'Ydewalle, G. (2005). Working memory and everyday conditional reasoning: retrieval and inhibition of stored counter-examples. *Thinking & Reasoning*, 11, 349-381.
- Deary, I. J. (2001). *Intelligence: a very short introduction*. Oxford: Oxford

University Press.

- Deary, I. J. (2008). Why do intelligent people live longer? *Nature*, 456, 175-176.
- Diener, E. & Crandall, R. (1978). *Ethics in social and behavioural research*. Chicago: University of Chicago Press.
- Dolton, P. J. & Vignoles, A. (2002). The return on post-compulsory school mathematics study. *Economica*, 69, 113-141.
- Elias, S. M. & Loomis, R. J. (2002). Utilizing Need for Cognition and perceived self-efficacy to predict academic performance. *Journal of Applied Social Psychology*, 32, 1687-1702.
- Erwin, T. D. (1981). *Manual for the scale of intellectual development*. Harrisburg, VA: Developmental Analytics.
- ESRC. (2009). *ESRC strategic plan 2009-2014*. Swindon: ESRC.
- Ethics Committee of the British Psychological Society. (2009). *Code of Ethics and Conduct*. Leicester: The British Psychological Society.
- 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. (1993). The mental model theory of conditional reasoning: critical appraisal and revision. *Cognition*, 48, 1-20.
- Evans, J. St. B. T. (1998). 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 Sciences*, 7, 454-459.
- Evans, J. St. B. T. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin & Review*, 13, 378-395.
- Evans, J. St. B. T. (2007). *Hypothetical thinking: dual processes in reasoning and judgement*. Hove: Psychology Press.
- Evans, J. St. B. T. (2011). Dual-process theories of reasoning: Contemporary issues and developmental applications. *Developmental Review*, 31, 86-102.
- 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. *The Quarterly Journal of Experimental Psychology*, 48A, 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. & Handley, S. J. (1999). The role of negation on conditional inference. *The Quarterly Journal of Experimental Psychology*, 52A, 739-769.
- Evans, J. St. B. T., Handley, S. J. & Bacon, A. M. (2009). Reasoning under time pressure: A study of causal conditional inference. *Experimental Psychology*, 56, 77-83.
- Evans, J. St. B. T., Handley, S. J., Neilens, H. & Over, D. (2010). The influence of cognitive ability and instructional set of causal conditional inference. *The Quarterly Journal of Experimental Psychology*, 63, 892-909.
- Evans, J. St. B. T., Handley, S. J., Neilens, H. & Over, D. E. (2007). Thinking about conditionals: a study of individual differences. *Memory & Cognition*, 35, 1772-1784.
- Facione, P. A. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction. Executive Summary "The Delphi Report"*. Millbrae, CA: California Academic Press.
- Farsides, T. & Woodfield, R. (2003). Individual differences and undergraduate academic success: the roles of personality, intelligence, and application. *Personality and Individual Differences*, 34, 1225-1243.
- Fong, G. T., Krantz, D. H. & Nisbett, R. E. (1986). The effects of statistical training on thinking about everyday problems. *Cognitive Psychology*, 18, 253-292.
- Fong, G. T. & Nisbett, R. E. (1991). Immediate and delayed transfer of training effects in statistical reasoning. *Journal of Experimental Psychology: General*, 120, 34-45.
- Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19, 25-42.
- Friedman, N. P., Miyake, A., Corley, R. P., Young, S. E., DeFries, J. C. & Hewitt, J. K. (2006). Not all executive functions are related to intelligence. *Psychological Science*, 17, 172-179.
- Frisch, D. (1993). Reasons for framing effects. *Organizational Behaviour and Human Decision Processes*, 54, 399-429.
- Garcia-Madruga, J. A., Gutierrez, F., Carriedo, N., Luzon, J. M. & Vila, J. O. (2007). Mental models in propositional reasoning and working memory's central executive. *Thinking & Reasoning*, 13, 370-393.
- George, D. & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference. 11.0 update* (4th ed.). Boston: Allyn & Bacon.
- Gigerenzer, G. & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, 62, 451-482.

- Gilhooly, K. J. & Fioratou, E. (2009). Executive functions in insight versus non-insight problem solving: An individual differences approach. *Thinking & Reasoning*, 15, 355-376.
- Gillard, E. (2009). *Dual processes in the psychology of mathematics education and beyond*. Unpublished doctoral dissertation, Katholieke Universiteit Leuven.
- Gillard, E., Van Dooren, W., Schaeken, W. & Verschaffel, L. (2009a). Processing time evidence for a default-interventionist model of probability judgements. In *Proceedings of the 31st Annual Meeting of the Cognitive Science Society* (p. 1792-1797). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Gillard, E., Van Dooren, W., Schaeken, W. & Verschaffel, L. (2009b). Proportional reasoning as a heuristic-based process. *Experimental Psychology*, 56, 92-99.
- Giroto, V., Kimmelmair, M., Sperber, D. & van der Henst, J. B. (2001). Inept reasoners or pragmatic virtuosos? relevance and the deontic selection task. *Cognition*, 81, 69-76.
- Giroto, V. & Legrenzi, P. (1989). Mental representation and hypothetico-deductive reasoning: The case of the THOG problem. *Psychological Research*, 51, 129-135.
- Giroto, V. & Legrenzi, P. (1993). Naming the parents of THOG: mental representation and reasoning. *Quarterly Journal of Experimental Psychology*, 46A, 701-713.
- Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M. & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgement. *Science*, 293, 2105-2108.
- Greenwald, A. G. & Banaji, M. R. (1995). Implicit social cognition: attitudes, self-esteem and stereotypes. *Psychological Review*, 102, 4-27.
- Handley, S. J., Capon, A., Beveridge, M., Dennis, I. & Evans, J. St. B. T. (2004). Working memory, inhibitory control and the development of children's reasoning. *Thinking & Reasoning*, 10, 175-195.
- Handley, S. J., Capon, A., Copp, C. & Harper, C. (2002). Conditional reasoning and the tower of hanoi: the role of spatial and verbal working memory. *British Journal of Psychology*, 93, 501-518.
- Harman, G. (1995). Thinking: invitation to cognitive science. In E. E. Smith & D. N. Osherson (Eds.), (Vol. 3, p. 175-211). Cambridge: MIT Press.
- Heim, A. (1969). *AH5 group test of intelligence*. London: National Foundation of Educational Research.
- Heiman, G. W. (2002). *Research methods in psychology*. Boston, MA: Houghton Mifflin.

- Heit, E. & Rotello, C. M. (2010). Relations between inductive reasoning and deductive reasoning. *Journal of Experimental Psychology: Learning, memory and Cognition*, 36, 805-812.
- Hergenhahn, B. R. & Olson, M. H. (2004). *An introduction to theories of learning* (7th ed.). New Jersey: Prentice Hall.
- Houston, K. (2009). *How to think like a mathematician: a companion to undergraduate mathematics*. Cambridge: Cambridge University Press.
- Hoyles, C. & Kuchemann, D. (2002). Students' understanding of logical implication. *Educational Studies in Mathematics*, 51, 193-223.
- Huckstep, P. (2000). Mathematics as a vehicle for 'mental training'. In S. Bramall & J. White (Eds.), *Why learn maths?* London: Institute of Education University of London.
- Inglis, M. (2012). Views on the theory of formal discipline. *Unpublished manuscript*.
- Inglis, M., Attridge, N., Batchelor, S. & Gilmore, C. (2011). Non-verbal number acuity correlates with symbolic mathematics achievement: But only in children. *Psychonomic Bulletin & Review*, 18, 1222-1229.
- Inglis, M., Palipana, A., Trenholm, S. & Ward, J. (2011). Individual differences in students' use of optimal learning resources. *Journal of Computer Assisted Learning*, 27, 490-502.
- Inglis, M. & Simpson, A. (2004). Mathematicians and the selection task. In M. Johnsen & A. B. Fuglestad (Eds.), *Proceedings of the 28th conference of the international group for the psychology of mathematics education* (Vol. 3, p. 89-96).
- Inglis, M. & Simpson, A. (2008). Conditional inference and advanced mathematical study. *Educational Studies in Mathematics*, 67, 187-204.
- Inglis, M. & Simpson, A. (2009a). Conditional inference and advanced mathematical study: Further evidence. *Educational Studies in Mathematics*, 72, 185-198.
- Inglis, M. & Simpson, A. (2009b). The defective and material conditionals in mathematics: does it matter? In *33rd Conference of the International Group for the Psychology of Mathematics Education, PME 33* (p. 225-232).
- Inhelder, B. & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence*. New York: Basic Books.
- Jensen, A. R. (1998). *The g factor: the science of mental ability*. Westport, CT: Greenwood.
- Johnson-Laird, P. N. (2008). *How we reason*. Oxford: Oxford University Press.
- Johnson-Laird, P. N. & Byrne, R. M. J. (1991). *Deduction*. Hillsdale, NJ: Lawrence Erlbaum Associates.

- Johnson-Laird, P. N. & Byrne, R. M. J. (2002). Conditionals: A theory of meaning, pragmatics, and inference. *Psychological Review*, 109, 646-678.
- Johnson-Laird, P. N., Byrne, R. M. J. & Schaeken, W. (1992). Propositional reasoning by model. *Psychological Review*, 99, 418-439.
- Johnson-Laird, P. N., Legrenzi, P. & Legrenzi, M. S. (1972). Reasoning and a sense of reality. *British Journal of Psychology*, 63, 395-400.
- Johnson-Laird, P. N. & Wason, P. C. (1970). A theoretical analysis of insight into a reasoning task. *Cognitive Psychology*, 1, 134-148.
- Jones, W., Russell, D. & Nickel, T. (1977). Belief in the paranormal scale: an objective instrument to measure belief in magical phenomena and causes. *JSAS Catalog of Selected Documents in Psychology*, 7, 1-32.
- Judge, T. A., Higgins, C. A., Thoresen, C. J. & Barrick, M. R. (1999). The big five personality traits, general mental ability, and career success across the life span. *Personnel Psychology*, 52, 621-652.
- Kahneman, D. (1991). Judgment and decision making: A personal view. *Psychological Science*, 2, 142-145.
- Kahneman, D. & Tversky, A. (1972). On prediction and judgement. *Oregon Research Institute Bulletin*, 12, 4.
- Kilpatrick, J. (1983). Research problems in mathematics education. *For the Learning of Mathematics*, 4(1), 45-46.
- Kosonen, P. & Winne, P. H. (1995). Effects of teaching statistical laws on reasoning about everyday problems. *Journal of Educational Psychology*, 87, 33-46.
- LaBerge, D. & Samuels, S. K. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, 6, 293-323.
- Landsberger, H. A. (1958). *Hawthorne revisited*. Ithaca, NY: Cornell University.
- Larrick, R. P., Nisbett, R. E. & Morgan, J. N. (1993). Who uses the cost-benefit rules of choice? implications for the normative status of microeconomic theory. *Organizational Behaviour and Human Decision Processes*, 56, 331-347.
- Larsen, L., Hartmann, P. & Nyborg, H. (2008). The stability of general intelligence from adulthood to middle-age. *Intelligence*, 36, 29-34.
- Lawson, D. (1997). What can we expect from a level mathematics students? *Teaching Mathematics and its Applications*, 16, 151-156.
- Lawson, D. (2003). Changes in student entry competences 1991-2001. *Teaching Mathematics and its Applications*, 22, 171-175.
- Lehman, D. R., Lempert, R. O. & Nisbett, R. E. (1988). The effects of graduate training on reasoning: Formal discipline and thinking about everyday-life events. *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.
- Lehmann, I. J. (1963). Changes in critical thinking, attitudes, and values from freshman to senior years. *Journal of Educational Psychology*, 54(6), 305-315.
- Lewin, I. (1982). Driver training: a perceptual-motor skill approach. *Ergonomics*, 25, 917-924.
- Locke, J. (1971/1706). *Conduct of the Understanding*. New York: Burt Franklin.
- Locurto, C. (1990). The malleability of IQ as judged from adoption studies. *Intelligence*, 14, 275-292.
- Manktelow, K. (1999). *Reasoning and thinking*. Hove, East Sussex: Psychology Press.
- Markovits, H. (1985). Incorrect conditional reasoning among adults: competence or performance? *British Journal of Psychology*, 76, 241-247.
- Markovits, H. & Doyon, C. (2004). Information processing and reasoning with premises that are empirically false: interference, working memory, and processing speed. *Memory & Cognition*, 32, 592-601.
- Markovits, H. & Nantel, G. (1989). The belief-bias effect in the production and evaluation of logical conclusions. *Memory and Cognition*, 17, 11-17.
- Miller, P. H. (2011). *Theories of developmental psychology* (Vol. 4th). New York, NY: Worth.
- Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H. & Howerter, A. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: a latent variable analysis. *Cognitive Psychology*, 41, 49-100.
- Mook, D. G. (1983). In defense of external invalidity. *American Psychologist*, 38, 379-387.
- Moutier, S., Angeard, N. & Houdé, O. (2002). Deductive reasoning and matching-bias inhibition training: evidence from a debiasing paradigm. *Thinking & Reasoning*, 8, 205-224.
- Nair, K. U. & Ramnarayan, S. (2000). Individual differences in Need for Cognition and complex problem solving. *Journal of Research in Personality*, 34, 305-328.
- National Research Council. (2001). *Adding it up: Helping children learn mathematics*. Washington, D.C: National Academy Press.
- Neilens, H. L., Handley, S. J. & Newstead, S. E. (2009). Effects of training and instruction on analytic and belief-based reasoning processes. *Thinking and Reasoning*, 15, 37-68.
- Neisser, U. (1967). *Cognitive psychology*. New York, NY: Meredith.

- Neisser, U., Boodoo, G., Bouchard, T. J., Boykin, A. W., Brody, N., Ceci, S. J. et al. (1996). Intelligence: knowns and unknowns. *American Psychologist*, 51, 77-101.
- Newstead, S. E. (2000). Are there two different types of thinking? (Peer commentary on "Individual differences in reasoning: implications for the rationality debate?" by K. E. Stanovich and R. F. West). *Behavioural and Brain Sciences*, 23, 645-726.
- Newstead, S. E., Girotto, V. & Legrenzi, P. (1995). The THOG problem and its implications for human reasoning. In S. E. Newstead & J. St. B. T. Evans (Eds.), *Perspectives on thinking and reasoning. Essays in honour of Peter Wason*. Hove, UK: Lawrence Erlbaum Associates.
- Nisbett, R. E. (2009). *Can reasoning be taught?* Cambridge, MA: American Academy of Arts and Sciences.
- Nisbett, R. E., Fong, G. T., Lehman, D. R. & Cheng, P. W. (1987). Teaching reasoning. *Science*, 238, 625-631.
- Nisbett, R. E., Krantz, D. H., Jepson, C. & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review*, 90, 339-363.
- Novick, M. R. & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika*, 32, 1-13.
- Oakley, C. O. (1949). Mathematics. *The American Mathematical Monthly*, 56, 19.
- Oaksford, M., Chater, N. & Larkin, J. (2000). Probabilities and polarity biases in conditional inference. *Journal of Experimental Psychology: Learning, memory and Cognition*, 26, 883-899.
- O'Brien, D. (2009). Human reasoning includes a mental logic. *Behavioural and Brain Sciences*, 32, 96-97.
- O'Brien, D. & Manfrinati, A. (2010). The mental logic theory of conditional propositions. In M. Oaksford & N. Chater (Eds.), *Cognition and conditionals: probability and logic in human thinking*. Oxford: Oxford University Press.
- Owen, A. M., Hampshire, A., Grah, J. A., Stenton, R., Dajani, S., Burns, A. S. et al. (2010). Putting brain training to the test. *Nature*, 465, 775-779.
- Pacini, R. & Epstein, S. (1999). The relation of rational and experiential information processing styles to personality, basic beliefs, and the ratio-bias phenomenon. *Journal of Personality and Social Psychology*, 76, 972-987.
- Petty, R. E. & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In *Advances in experimental social psychology* (Vol. 19, p. 123-205). New York: Academic Press.
- Piaget, J. (1928). *Judgement and reasoning in the child*. London: Routledge.

- Piaget, J. (1960). *Logic and psychology*. New York: Basic Books.
- Piaget, J. (1970). Piaget's theory. In P. Mussen (Ed.), *Carmichael's manual of child psychology*. New York: Wiley.
- Plato. (2003/375B.C). *The Republic* (D. Lee, Ed.). London: Penguin.
- Pollard, P. & Evans, J. St. B. T. (1980). The influence of logic on conditional reasonign performance. *Quarterly Journal of Experimental Psychology*, 32, 605-624.
- Polya, G. (1954). *Induction and analogy in mathematics*. New Jersey: Princeton University Press.
- Pyszczynski, T., Greenberg, J. & Solomon, S. (1999). A dual-process model of defense against conscious and unconscious death-related thoughts: an extension of terror management theory. *Psychological Review*, 106, 835-845.
- Raven, J., Raven, J. C. & Court, J. H. (1998). *Manual for raven's advanced progressive matrices and vocabulary scales*. San Antonio: Pearson.
- Reeve, C. L. & Lam, H. (2005). The psychometric paradox of practice effects due to retesting: measurement invariance and stable ability estimates in the face of observed score changes. *Intelligence*, 33, 535-549.
- Reyna, V. F. & Brainerd, C. J. (1995). Fuzzy-trace theory: an interim synthesis. *Learning and Individual Differences*, 7, 1-75.
- Rips, L. (1989). The psychology of knights and knaves. *Cognition*, 31(2), 85-116.
- Rokeach, M. (1960). *The open and closed mind*. New York: Basic Books.
- Rönnlund, M. & Nilsson, L.-G. (2006). Adult life-span patterns in WAIS-R block design performance: cross-sectional versus longitudinal age gradients and relations to demographic factors. *Intelligence*, 34, 63-78.
- Ruston, J. P. & Jensen, A. R. (2010). The rise and fall of the Flynn Effect as a reason to expect a narrowing of the Black-White iq gap. *Intelligence*, 38, 213-219.
- Sá, W. C., West, R. F. & Stanovich, K. E. (1999). The domain specificity and generality of belief bias: Searching for a generalisable critical thinking skill. *Journal of Educational Psychology*, 91, 497-510.
- Sanz de Acedo Lizarraga, M. L., Sanz de Acedo Baquedano, M. T. & Soria Oliver, M. (2010). Psychological intervention in thinking skills with primary education students. *School Psychology International*, 31, 131-145.
- Schneider, W., Eschman, A. & Zuccolotto, A. (2002). *E-Prime reference guide*. Pittsburgh: Psychology Software Tools, Inc.
- Schroyens, W., Schaeken, W. & d'Ydewalle, G. (2001). The processing of negations in conditional reasoning: A meta-analytical case study in mental

- model and/or mental logic theory. *Thinking and Reasoning*, 7, 121-172.
- Sfard, A. (1998). A mathematician's view of research in mathematics education: An interview with Shimson A. Amitsur. In A. Sierpiska & J. Kilpatrick (Eds.), *Mathematics education as a research domain: A search for identity* (Vol. 2, p. 445-458). Dordrecht: Kluwer.
- Shafir, E. (1994). Uncertainty and the difficulty of thinking through disjunctions. *Cognition*, 50, 403-430.
- Shih, M., Pittinsky, T. L. & Ambady, N. (1999). Stereotype susceptibility: identity salience and shifts in quantitative performance. *Psychological Science*, 10, 80-83.
- Skinner, B. F. (1938). *The behaviour of organisms*. New York: Appleton-Century-Crofts.
- Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, 119, 3-22.
- Smith, A. (2004). *Making mathematics count: The report of Professor Adrian Smith's inquiry into post-14 mathematics education*. London: The Stationery Office.
- Spearman, C. (1927). *The abilities of of man*. New York: Macmillan.
- Spencer, S. J., Steele, C. M. & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, 35, 4-28.
- Sperber, D., Cara, F. & Girotto, V. (1995). Relevance theory explains the selection task. *Cognition*, 57, 31-96.
- Sperber, D. & Girotto, V. (2002). Use of misuse of the selection task? Rejoinder to Fiddick, Cosmides, and Tooby. *Cognition*, 85, 277-290.
- Stanovich, K. E. (1999). *Who is rational? Studies of individual difference in reasoning*. Mahwah, New Jersey: Lawrence Erlbaum Associates.
- Stanovich, K. E. (2004). *The robot's rebellion: Finding meaning in the age of Darwin*. Chicago: Univeristy of Chicago Press.
- Stanovich, K. E. (2009a). In two minds. In J. St. B. T. Evans & K. Frankish (Eds.), (p. 55-88). Oxford: Oxford University Press.
- Stanovich, K. E. (2009b). *What intelligence tests miss: The psychology of rational thought*. Yale: Yale University Press.
- Stanovich, K. E. & Cunningham, A. E. (1992). Studying the consequences of literacy within a literate society: the cognitive correlates of print exposure. *Memory & Cognition*, 20, 51-68.
- Stanovich, K. E. & West, R. F. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of Educational Psychology*, 89, 342-357.
- Stanovich, K. E. & West, R. F. (1998). Individual differences in rational thought.

- Journal of Experimental Psychology: General*, 127, 161-188.
- Stanovich, K. E. & West, R. F. (2000). Individual differences in reasoning: implication for the rationality debate. *Behavioural and Brain Sciences*, 23, 645-726.
- Stanovich, K. E. & West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. *Personality Processes and Individual Differences*, 94, 672-695.
- Stich, S. (1990). *The fragmentation of reason*. Cambridge, MA: MIT Press.
- Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18, 643-662.
- The British Psychological Society. (2010). *Code of human research ethics*. Leicester: The British Psychological Society.
- Thompson, V. A. (2009). Dual process theories: a metacognitive perspective. In J. St. B. T. Evans & K. Frankish (Eds.), *In two minds: dual processes and beyond*. Oxford: Oxford University Press.
- Thompson, V. A. (2010). Towards a metacognitive dual process theory of conditional reasoning. In M. Oaksford & N. Chater (Eds.), *Cognition and conditionals: probability and logic in human thinking*. Oxford: Oxford University Press.
- Thompson, V. A., Prowse Turner, J. A. & Pennycook, G. (2011). Intuition, reason and metacognition. *Cognitive Psychology*, 63, 107-140.
- Thorndike, E. L. (1924). Mental discipline in high school studies. *The Journal of Educational Psychology*, 15, 1-22.
- Thorndike, E. L. & Woodworth, R. S. (1901). The influence of improvement in one mental function upon the efficiency of other functions. *Psychological Review*, 8, 247-261.
- Tobacyk, J. & Milford, G. (1983). Belief in paranormal phenomena. *Journal of Personality and Social Psychology*, 44, 1029-1037.
- Toms, M., Morris, N. & Ward, D. (1993). Working memory and conditional reasoning. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, 46, 679-699.
- Toplak, M. E. & Stanovich, K. E. (2002). The domain specificity and generality of disjunctive reasoning: searching for a generalizable critical thinking skill. *Journal of Educational Psychology*, 94, 197-209.
- Toplak, M. E., West, R. F. & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, 39, 1275-1289.
- Troldhal, V. & Powell, F. (1965). A short-form dogmatism scale for use in field studies. *Social Forces*, 44, 211-215.
- Tversky, A. & Kahneman, D. (1971). Belief in the law of small numbers.

- Psychological Bulletin*, 76, 105-110.
- Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124-1131.
- Tversky, A. & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211, 453-458.
- Tversky, A. & Kahneman, D. (1982). Evidential impact of base rates. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgement under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.
- Tversky, A. & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgement. *Psychological Review*, 90, 293-315.
- Vamvakoussi, X., Van Dooren, W. & Verschaffel, L. (2012a). Educated adults are still affected by intuitions about the effect of arithmetical operation: evidence from a reaction-time study. *Educational Studies in Mathematics, Advance online publication*. DOI: 10.1007/s10649-012-9432-8.
- Vamvakoussi, X., Van Dooren, W. & Verschaffel, L. (2012b). Naturally biased? in search of reaction time evidence for a natural number bias in adults. *The Journal of Mathematical Behavior*, 31, 344-355.
- Van Breukelen, G. J. P. (2006). Ancova versus change from baseline had more power in randomized studies and more bias in nonrandomized studies. *Journal of Clinical Epidemiology*, 59, 920-925.
- Venet, M. & Markovits, H. (2001). Understanding uncertainty with abstract conditionals. *Merrill-Palmer Quarterly*, 47, 74-99.
- Verschueren, N., Schaeken, W. & d'Ydewalle, G. (2005). Everyday conditional reasoning: a working memory-dependent tradeoff between counterexample and likelihood use. *Memory & Cognition*, 33, 107-119.
- Walport, M. (2010). *Science and mathematics secondary education for the 21st century: Report of the Science and Learning Expert Group*. London: Crown.
- Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New horizons in psychology*. Harmondsworth: Penguin.
- Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, 20(3), 273-281.
- Wason, P. C. & Brooks, P. J. (1979). THOG: the anatomy of a problem. *Psychological Research*, 41, 79-90.
- Wason, P. C. & Johnson-Laird, P. N. (1972). *Psychology of reasoning: structure and content*. London: Batsford.
- Wason, P. C. & Shapiro, D. (1971). Natural and contrived experience in a reasoning problem. *Quarterly Journal of Experimental Psychology*, 23, 63-71.

- West, R. F., Toplak, M. E. & Stanovich, K. E. (2008). Heuristics and biases as measures of critical thinking: Associations with cognitive ability and thinking dispositions. *Journal of Educational Psychology*, 100, 930-941.
- White, E. E. (1936). A study of the possibility of improving habits of thought in school children by a training in logic. *British Journal of Educational Psychology*, 6, 267-273.
- Zepp, R. A. (1987). Logic in everyday language and in mathematics. *Chinese Univeristy Education Journal*, 15, 45-49.

Appendix A

Ravens Advanced Progressive Matrices

The Raven's Advanced Progressive Matrices subset used throughout this thesis consisted of items 12 and 14-30 from Set II. The instructions used are presented below. Participants were given 15 minutes to complete the task in Chapters 5 and 6 and 10 minutes in Chapter 8.

Pattern completion

In this section you will see grids and each one will have a pattern with a piece missing. Your task is to decide which of the numbered pieces correctly completes the grid. In each problem, circle the piece that you think is correct. There are 18 items and you will have **15 (10) minutes** to do as many as you can. First, take a look at the example problem below. The piece labelled 8 is correct because it fits the pattern both down and across the grid.

Appendix B

Conditional Inference Task

On the following pages is the full conditional inference task, including instructions, as it was presented to participants (but without randomisation of the items).

Logic

Please read the following instructions carefully.

This section is concerned with peoples ability to reason logically with sentences in various forms. You will be presented with a total of 32 problems on the following pages. In each case you are given two statements together with a conclusion which may or may not follow from these statements.

Your task in each case is to decide whether or not the conclusion *necessarily* follows from the statements. **A conclusion is *necessary* if it must be true, given that the statements are true.**

Each problem concerns an imaginary letter-number pair and contains an initial statement or rule which determines which letters may be paired with which numbers. An example of a rule of similar form to those used would be:

- ☐ If the letter is B then the number is not 7.

In each case you must assume that the rule holds and then combine it with the information given in the second statement. This will concern either the letter or the number of an imaginary pair, for example:

- ☐ The letter is Y.
☐ The number is not 4.

If the information concerns the letter the conclusion will concern the number and vice-versa.

A full problem looks something like:

If the letter is X then the number is 1.
The letter is X.
Conclusion: The number is 1.
☐ YES
☐ NO

If you think the conclusion necessarily follows then please tick the YES box, otherwise tick the NO box. Please work through the problems in order and make sure you do not miss any. Do not return to a problem once you have finished and moved on to another.

If you think the conclusion necessarily follows please tick YES, otherwise tick NO. Do not return to a problem once you have finished and moved on to another. Answer all questions.

-
1. If the letter is A then the number is 3.

The letter is A.

Conclusion: The number is 3.

- ☐ YES
☐ NO

-
2. If the letter is T then the number is 5.

The letter is not T.

Conclusion: The number is not 5.

- ☐ YES
☐ NO

-
3. If the letter is F then the number is 8.

The number is 8.

Conclusion: The letter is F.

- ☐ YES
☐ NO

-
4. If the letter is D then the number is 4.

The number is not 4.

Conclusion: The letter is not D.

- ☐ YES
☐ NO

-
5. If the letter is G then the number is not 6.

The letter is G.

Conclusion: The number is not 6.

- ☐ YES
☐ NO

-
6. If the letter is R then the number is not 1.

The letter is not R.

Conclusion: The number is 1.

- ☐ YES
☐ NO

7. If the letter is K then the number is not 3.

The number is not 3.

Conclusion: The letter is K.

☐ YES

☐ NO

8. If the letter is U then the number is not 9.

The number is 9.

Conclusion: The letter is not U.

☐ YES

☐ NO

9. If the letter is not B then the number is 5.

The letter is not B.

Conclusion: The number is 5.

☐ YES

☐ NO

10. If the letter is not S then the number is 6.

The letter is S.

Conclusion: The number is not 6.

☐ YES

☐ NO

11. If the letter is not V then the number is 8.

The number is 8.

Conclusion: The letter is not V.

☐ YES

☐ NO

12. If the letter is not H then the number is 1.

The number is not 1.

Conclusion: The letter is H.

☐ YES

☐ NO

13. If the letter is not F then the number is not 3.

The letter is not F.

Conclusion: The number is not 3.

☐ YES

☐ NO

14. If the letter is not L then the number is not 9.

The letter is L.

Conclusion: The number is 9.

☐ YES

☐ NO

15. If the letter is not J then the number is not 8.

The number is not 8.

Conclusion: The letter is not J.

☐ YES

☐ NO

16. If the letter is not V then the number is not 7.

The number is 7.

Conclusion: The letter is V.

☐ YES

☐ NO

17. If the letter is D then the number is 2.

The letter is D.

Conclusion: The number is 2.

☐ YES

☐ NO

18. If the letter is Q then the number is 1.

The letter is K.

Conclusion: The number is not 1.

☐ YES

☐ NO

19. If the letter is M then the number is 4.

The number is 4.

Conclusion: The letter is M.

☐ YES

☐ NO

10. If the letter is V then the number is 5.

The number is 2.

Conclusion: The letter is not V.

☐ YES

☐ NO

21. If the letter is S then the number is not 8.

The letter is S.

Conclusion: The number is not 8.

☐ YES

☐ NO

22. If the letter is B then the number is not 3.

The letter is H.

Conclusion: The number is 3.

☐ YES

☐ NO

23. If the letter is J then the number is not 2.

The number is 7.

Conclusion: The letter is J.

☐ YES

☐ NO

24. If the letter is U then the number is not 7.

The number is 7.

Conclusion: The letter is not U.

☐ YES

☐ NO

25. If the letter is not E then the number is 2.

The letter is R.

Conclusion: The number is 2.

☐ YES

☐ NO

26. If the letter is not A then the number is 6.

The letter is A.

Conclusion: The number is not 6.

☐ YES

☐ NO

27. If the letter is not C then the number is 9.

The number is 9.

Conclusion: The letter is not C.

☐ YES

☐ NO

28. If the letter is not N then the number is 3.

The number is 5.

Conclusion: The letter is N.

☐ YES

☐ NO

29. If the letter is not A then the number is not 1.

The letter is N.

Conclusion: The number is not 1.

☐ YES

☐ NO

30. If the letter is not C then the number is not 2.

The letter is C.

Conclusion: The number is 2.

☐ YES

☐ NO

31. If the letter is not W then the number is not 8.

The number is 3.

Conclusion: The letter is not W.

☐ YES

☐ NO

32. If the letter is not K then the number is not 1.

The number is 1.

Conclusion: The letter is K.

☐ YES

☐ NO

Appendix C

Belief Bias Syllogisms Task

On the following pages is the belief bias syllogisms task and its instructions as presented to participants. The task is divided into two parts as described in Chapter 5. Part 1 is questions 1-12 and part 2 is questions 13-24.

Reasoning with information

Please read the following instructions carefully.

In the following problems, you will be given two premises, which you must assume are true. A conclusion from the premises then follows. *You must decide whether the conclusion follows logically from the premises or not. You must suppose that the premises are all true and limit yourself only to the information contained in the premises.* This is very important. Decide if the conclusion follows logically from the premises, assuming the premises are true, and tick your response.

If you think the conclusion logically follows, assuming that the premises are true, please tick YES, otherwise tick NO. Answer all questions.

Part 1:

1. Premises: All things that are smoked are good for the health.
Cigarettes are smoked.
Conclusion: Cigarettes are good for the health.

☐ YES

☐ NO

2. Premises: All things made of wood can be used as fuel.
Gasoline is not made of wood.
Conclusion: Gasoline cannot be used as fuel.

☐ YES

☐ NO

3. Premises: All lapitars wear clothes.
Podips wear clothes.
Conclusion: Podips are lapitars.

☐ YES

☐ NO

4. Premises: All nuts can be eaten.
Rocks cannot be eaten.
Conclusion: Rocks are not nuts.

☐ YES

☐ NO

5. Premises: All poor people are unemployed.
Rockefeller is not poor.
Conclusion: Rockefeller is not unemployed.

☐ YES

☐ NO

6. Premises: All guns are dangerous.
Rattlesnakes are dangerous.
Conclusion: Rattlesnakes are guns.

☐ YES

☐ NO

7. Premises: All things with four legs are dangerous.
Poodles are not dangerous.
Conclusion: Poodles do not have four legs.

☐ YES

☐ NO

8. Premises: All ramadions taste delicious.
Gumthorps are ramadions.
Conclusion: Gumthorps taste delicious.

☐ YES

☐ NO

9. Premises: All living things need water.
Roses need water.
Conclusion: Roses are living things.

☐ YES

☐ NO

10. Premises: All selacians have sharp teeth.
Snorlups do not have sharp teeth.
Conclusion: Snorlups are not selacians.

☐ YES

☐ NO

11. Premises: All fish can swim.

Tuna are fish.

Conclusion: Tuna can swim.

☐ YES

☐ NO

12. Premises: All hudon are ferocious.

Wampets are not hudon.

Conclusion: Wampets are not ferocious.

☐ YES

☐ NO

Part 2:

13. Premises: All opprobines run on electricity.
Jamtops run on electricity.
Conclusion: Jamtops are opprobines.

☐ YES
☐ NO

14. Premises: All things that are alive drink water.
Televisions do not drink water.
Conclusion: Televisions are not alive.

☐ YES
☐ NO

15. Premises: All bats have wings.
Hawks are not bats.
Conclusion: Hawks do not have wings.

☐ YES
☐ NO

16. Premises: All mammals walk.
Whales are mammals.
Conclusion: Whales walk.

☐ YES
☐ NO

17. Premises: All large things need oxygen.
Mice need oxygen.
Conclusion: Mice are large things.

☐ YES
☐ NO

18. Premises: All African countries are hot.

Canada is not an African country.

Conclusion: Canada is not hot.

☐ YES

☐ NO

19. Premises: All things that move love water.

Cats do not love water.

Conclusion: Cats do not move.

☐ YES

☐ NO

20. Premises: All tumpers lay eggs.

Sampets are tumpers.

Conclusion: Sampets lay eggs.

☐ YES

☐ NO

21. Premises: All things that have a motor need oil.

Automobiles need oil.

Conclusion: Automobiles have motors.

☐ YES

☐ NO

22. Premises: All snapples run fast.

Alcomas do not run fast.

Conclusion: Alcomas are not snapples.

☐ YES

☐ NO

23. Premises: All birds have feathers.
Robins are birds.
Conclusion: Robins have feathers.

- ☐ YES
☐ NO
-

24. Premises: All argomelles are kind.
Magsums are not argomelles.
Conclusion: Magsums are not kind.

- ☐ YES
☐ NO
-

Appendix D

Need for Cognition Scale

Thinking style

Please rate the following statements according to this scale:

	1	2	3	4	5	6	7	8	9
	very strongly disagree		moderately disagree		neither agree nor disagree		moderately agree		very strongly agree
I would prefer complex to simple problems.	1	2	3	4	5	6	7	8	9
I like to have the responsibility of handling a situation that requires a lot of thinking.	1	2	3	4	5	6	7	8	9
Thinking is not my idea of fun.	1	2	3	4	5	6	7	8	9
I would rather do something that requires little thought than something that is sure to challenge my thinking abilities.	1	2	3	4	5	6	7	8	9
I try to anticipate and avoid situations where there is likely a chance I will have to think in depth about something.	1	2	3	4	5	6	7	8	9
I find satisfaction in deliberating hard and for long hours.	1	2	3	4	5	6	7	8	9
I only think as hard as I have to.	1	2	3	4	5	6	7	8	9
I prefer to think about small, daily projects to long-term ones.	1	2	3	4	5	6	7	8	9
I like tasks that require little thought once I've learned them.	1	2	3	4	5	6	7	8	9

The idea of relying on thought to make my way to the top appeals to me.	1	2	3	4	5	6	7	8	9
I really enjoy a task that involves coming up with new solutions to problems.	1	2	3	4	5	6	7	8	9
Learning new ways to think doesnt excite me very much.	1	2	3	4	5	6	7	8	9
I prefer my life to be filled with puzzles that I must solve.	1	2	3	4	5	6	7	8	9
The notion of thinking abstractly is appealing to me.	1	2	3	4	5	6	7	8	9
I would prefer a task that is intellectual, difficult, and important to one that is somewhat important but does not require much thought.	1	2	3	4	5	6	7	8	9
I feel relief rather than satisfaction after completing a task that required a lot of mental effort.	1	2	3	4	5	6	7	8	9
Its enough for me that something gets the job done; I dont care how or why it works.	1	2	3	4	5	6	7	8	9
I usually end up deliberating about issues even when they do not affect me personally.	1	2	3	4	5	6	7	8	9

Appendix E

AS level Mathematics Test

The mathematics manipulation check given to the AS level students in Chapter 5 is presented below along with the instructions.

Questions 1-12 were taken from the Woodcock-Johnson III Calculation subtest. Nine of them had shown an average accuracy of less than 55% and correlated with performance on the whole test at .86 in a previous dataset with mixed-discipline undergraduate students (Inglis, Attridge et al., 2011) and the remaining three were taken from the lower range to prevent floor effects in the English literature students.

Questions 13-15 were the most difficult questions on the Loughborough University diagnostic test for new mathematics undergraduates based on performance in 2008 and 2009. The diagnostic test is designed to assess students' capability with material covered in AS level mathematics, and these three items were included to prevent ceiling effects in the mathematics students at the second time point whilst ensuring that the content was not inappropriately advanced. Questions were presented in a set order that was intended to be progressive.

Calculation

Some of the questions in this section are **very difficult**, and some you will **never even have come across** before. We do not expect you to be able to attempt all of the questions, please just answer as many as you can. If you get stuck, do not worry, just move on to the next section.

1. $9 + 7 =$

2. $8 \times 5 =$

3. $48 - 19 =$

4. $1.05 \times .2 =$

5. $\left(\frac{4b}{3y}\right)\left(\frac{-4y}{12b^2}\right) =$

6. $\log_b 81 = 4$
 $b =$

7. $f(x) = 6x^3$
 $f'(x) =$

8. $\cos \theta = \frac{\sqrt{3}}{2}$
 $\theta =$

9. $2y = 6x + 8$
Slope=
 y -intercept=

10. Evaluate:

$$\begin{vmatrix} 8 & 2 \\ -4 & 1 \end{vmatrix}$$

11. $\int_1^3 3x^2 dx =$

12. $\tan \theta = 1$

$$\sin \theta =$$

13. When expressing $\frac{x}{(x+1)^2(x^2+2)}$ in partial fractions, the appropriate form is

- ☐ $\frac{A}{x+1} + \frac{Bx+C}{x^2+2}$
- ☐ $\frac{A}{x+1} + \frac{B}{x^2+2}$
- ☐ $\frac{A}{(x+1)^2} + \frac{B}{x+1} + \frac{C}{x^2+2}$
- ☐ $\frac{A}{(x+1)^2} + \frac{B}{x+1} + \frac{Cx+D}{x^2+2}$

14. $\int xe^x dx$ is

- ☐ $x^2e^x + c$
- ☐ $xe^x - e^x + c$
- ☐ $\frac{x^2}{2}e^x + c$
- ☐ $\frac{x^2}{2}e^{x+1} + c$

15. If the substitution $u = 5x - 7$ is used to find the integral $\int \frac{1}{(5x-7)^2} dx$ the result is

- ☐ $-\frac{1}{5(5x-7)} + c$
- ☐ $\frac{1}{5x-7} + c$
- ☐ $-\frac{1}{5x-7} + c$
- ☐ $-\frac{5}{5x-7} + c$

Appendix F

Thematic Conditional Inference Task

Here, the thematic conditional inference task that was given to the undergraduate students in Chapter 6 is presented with its instructions. There was no time limit.

Real world logic

Please read the following instructions carefully:

This study is concerned with people's ability to reason logically with everyday sentences. You will be presented with a total of 16 problems on the following pages. In each case you are given two statements together with a conclusion which may or may not follow from these statements.

Your task in each case is to evaluate the logic of the sentences, independently of the content, and decide whether or not the conclusion *necessarily* follows from the statements. **A conclusion is *necessary* if it must be true, given that the statements are true.**

The problems look something like this:

Assume the following is true:

If the material is wood, then it is hard.

Given that the following premise is also true:

The material is wood.

is it necessary that:

The material is hard.

- ☐ YES
- ☐ NO

In each case you must assume that the rule holds and combine it with the information in the second statement in order to decide whether the conclusion necessarily follows.

If you think the conclusion does necessarily follow then please tick the YES box, otherwise tick the NO box. Please work through the problems in order and make sure you do not miss any. Do not return to a problem once you have finished and moved on to another.

If you think the conclusion necessarily follows please tick YES, otherwise tick NO. Do not return to a problem once you have finished and moved on to another. Answer all questions.

1. *Assume the following is true:*

If oil prices continue to rise then UK petrol prices will rise.

Given that the following premise is also true:

Oil prices continue to rise.

is it necessary that:

UK petrol prices rise.

☐ YES

☐ NO

2. *Assume the following is true:*

If car ownership increases then traffic congestion will get worse.

Given that the following premise is also true:

Traffic congestion does not get worse.

is it necessary that:

Car ownership does not increase.

☐ YES

☐ NO

3. *Assume the following is true:*

If more people use protective sun cream then cases of skin cancer will be reduced.

Given that the following premise is also true:

Cases of skin cancer are not reduced.

is it necessary that:

More people do not use protective sun cream.

☐ YES

☐ NO

4. *Assume the following is true:*

If Sony release a PlayStation 4 then their company profits will rise.

Given that the following premise is also true:

Sony release a PlayStation 4.

is it necessary that:

The company profits rise.

☐ YES

☐ NO

5. *Assume the following is true:*

If oil prices continue to rise then UK petrol prices will rise.

Given that the following premise is also true:

Oil prices do not continue to rise.

is it necessary that:

UK petrol prices do not rise.

☐ YES

☐ NO

6. *Assume the following is true:*

If car ownership increases then traffic congestion will get worse.

Given that the following premise is also true:

Traffic congestion gets worse.

is it necessary that:

Car ownership increases.

☐ YES

☐ NO

7. *Assume the following is true:*

If more people use protective sun cream then cases of skin cancer will be reduced.

Given that the following premise is also true:

Cases of skin cancer are reduced.

is it necessary that:

More people use protective sun cream.

☐ YES

☐ NO

8. *Assume the following is true:*

If Sony release a PlayStation 4 then their company profits will rise.

Given that the following premise is also true:

Sony do not release a Playstation 4.

is it necessary that:

The company profits do not rise.

☐ YES

☐ NO

9. *Assume the following is true:*

If more new houses are built then the amount of homeless people will increase.

Given that the following premise is also true:

More new houses are built.

is it necessary that:

The amount of homeless people increases.

☐ YES

☐ NO

10. *Assume the following is true:*

If third world debt is cancelled then world poverty will worsen.

Given that the following premise is also true:

World poverty does not worsen.

is it necessary that:

Third world debt is not cancelled.

☐ YES

☐ NO

11. *Assume the following is true:*

If fast food is taxed then childhood obesity will increase.

Given that the following premise is also true:

Childhood obesity does not increase.

is it necessary that:

Fast food is not taxed.

☐ YES

☐ NO

12. *Assume the following is true:*

If EU quarantine laws are strengthened then rabies will spread to the UK.

Given that the following premise is also true:

EU quarantine laws are strengthened.

is it necessary that:

Rabies spreads to the UK.

☐ YES

☐ NO

13. *Assume the following is true:*

If more new houses are built then the amount of homeless people will increase.

Given that the following premise is also true:

More new houses are not built.

is it necessary that:

The amount of homeless people does not increase.

☐ YES

☐ NO

14. *Assume the following is true:*

If third world debt is cancelled then world poverty will worsen.

Given that the following premise is also true:

World poverty worsens.

is it necessary that:

Third world debt is cancelled.

☐ YES

☐ NO

15. *Assume the following is true:*

If fast food is taxed then childhood obesity will increase.

Given that the following premise is also true:

Childhood obesity increases.

is it necessary that:

Fast food is taxed.

☐ YES

☐ NO

16. *Assume the following is true:*

If EU quarantine laws are strengthened then rabies will spread to the UK.

Given that the following premise is also true:

EU quarantine laws are not strengthened.

is it necessary that:

Rabies does not spread to the UK.

☐ YES

☐ NO

Appendix G

Undergraduate Mathematics Test

The mathematics test given to undergraduate students in Chapter 6 consisted of 11 questions. Seven of these were taken from the Woodcock-Johnson III Calculation subtest, two were the most difficult questions on the Loughborough University diagnostic test for new mathematics undergraduates based on performance in 2008 and 2009, and the final two were based on the first year mathematics degree syllabus.

Calculation

Some of the questions in this section are **very difficult**, and some you will **never even have come across** before. We do not expect you to be able to attempt all of the questions, please just answer as many as you can. If you get stuck, do not worry, just move on to the next section. You may make notes but please do not use a calculator.

1. $8 \times 5 =$

2. $48 - 19 =$

3. $1.05 \times .2 =$

4. $\left(\frac{4b}{3y}\right)\left(\frac{-4y}{12b^2}\right) =$

5. $f(x) = 6x^3$
 $f'(x) =$

6. $2y = 6x + 8$
Slope=
 y -intercept=

7. $\int_1^3 3x^2 dx =$

8. When expressing $\frac{x}{(x+1)^2(x^2+2)}$ in partial fractions, the appropriate form is

☐ $\frac{A}{x+1} + \frac{Bx+C}{x^2+2}$

☐ $\frac{A}{x+1} + \frac{B}{x^2+2}$

☐ $\frac{A}{(x+1)^2} + \frac{B}{x+1} + \frac{C}{x^2+2}$

☐ $\frac{A}{(x+1)^2} + \frac{B}{x+1} + \frac{Cx+D}{x^2+2}$

9. $\int xe^x dx$ is

- ☐ $x^2e^x + c$
- ☐ $xe^x - e^x + c$
- ☐ $\frac{x^2}{2}e^x + c$
- ☐ $\frac{x^2}{2}e^{x+1} + c$

10. Given $f(x, y) = e^{xy}$ find $\frac{\partial^2 f}{\partial x^2}$

11. Express the complex number $\frac{1}{1+i}$ in the form $x + iy$ where $i^2 = -1$, and x and y are real numbers.

List of Abbreviations

- A level** Advanced level. An optional two year qualification that is usually taken after school but before university in the UK.
- AC** Affirmation of the consequent. An inference of the structure ‘if p then q ; q ; therefore p ’.
- ANOVA** Analysis of Variance. A statistical test for comparing means across groups.
- ANCOVA** Analysis of Covariance. A statistical test for comparing means across groups while controlling for a covariate.
- APB** Affirmative premise bias. The conditional inference bias towards endorsing more inferences with affirmative minor premises than with negative minor premises.
- API** Affirmative premise index. A behavioural measure of the affirmative premise bias.
- AS level** Advanced subsidiary level. The first year of an A level course.
- BCI** Biconditional index. A behavioural measure of the tendency to assume a biconditional interpretation of conditional statements.
- CCI** Conjunctive conditional index. A behavioural measure of the tendency to assume a conjunctive interpretation of conditional statements.
- CRT** Cognitive Reflection Test. A measure of thinking disposition.
- DA** Denial of the antecedent. An inference of the structure ‘if p then q ; not p ; therefore not q ’.
- DCI** Defective conditional index. A behavioural measure of the tendency to assume a defective interpretation of conditional statements.
- IT** If then. A phrasing of conditional statements, for example, ‘If the letter is G then the number is 5’.
- MCI** Material conditional index. A behavioural measure of the tendency to assume a material interpretation of conditional statements.
- MP** Modus Ponens. An inference of the structure ‘if p then q ; p ; therefore q ’.

MT Modus Tollens. An inference of the structure ‘if p then q ; not q ; therefore not p ’.

NCB Negative conclusion bias. The conditional inference bias towards endorsing more inferences with negative conclusions than with affirmative conclusions.

NCI Negative conclusion index. A behavioural measure of the negative conclusion bias.

NFC Need for Cognition. A measure of thinking disposition.

OI Only if. A phrasing of conditional statements, for example, ‘The letter is G only if the number is 5’.

RAPM Raven’s Advanced Progressive Matrices. A measure of intelligence.

Author Index

- Alter, A. L., 38, 189
 Ambady, N., 52, 199
 American Council on Education, 24, 189
 Anderson, A. R., 5, 26, 82, 85, 189
 Angeard, N., 44, 196
 Ardila, A., 121, 170, 177, 189
 Arffa, S., 121, 170, 177, 189
 Aronson, J., 52, 189
 Ashenfelter, O., 13, 42, 189
 Attridge, N., 102, 103, 194, 221
- Bacon, A. M., 32, 192
 Baddeley, A., 189
 Baddeley, A. D., 189
 Banaji, M. R., 33, 193
 Banich, M. T., 44, 169, 189
 Baron, J., 68, 189
 Barrick, M. R., 13, 195
 Barston, J. L., 9, 191
 Batchelor, S., 102, 194
 Belnap, N. D., 5, 26, 82, 85, 189
 Beth, E. W., 16, 189
 Beveridge, M., 170, 193
 St B T Evans, J., 170, 193
 Boodoo, G., 197
 Bors, D. A., 43, 190
 Bouchard, T. J., 197
 Boykin, A. W., 197
 Braine, M. D. S., 5, 26, 82, 85, 190
 Brainerd, C. J., 33, 190, 198
 Bramall, S., 1, 47, 190
 Brody, N., 197
 Brooks, P. J., 70, 71, 201
 Burns, A. S., 197
 Byrne, R. M. J., 16, 28–30, 194, 195
- Cacioppo, J. T., 33, 45, 86, 101, 190, 197
 Capon, A., 170, 171, 193
 Cara, F., 20, 199
- Carriedo, N., 170, 192
 Ceci, S. J., 197
 Chaiken, S., 34, 190
 Chater, N., 9, 197
 Chen, S., 34, 190
 Cheng, P. W., 19, 20, 24, 31, 32, 118, 190, 197
 Christensen, L. B., 50, 53, 55, 56, 58–60, 62, 63, 120, 190
 Clibbens, J., 2, 191
 Cohen, J. D., 33, 193
 Copp, C., 171, 193
 Corley, R. P., 192
 Cosmides, L., 13, 76, 190
 Court, J. H., 43, 198
 Crandall, R., 50, 191
 Cummins, D. D., 7, 190
 Cunningham, A. E., 86, 199
 Curtis-Holmes, J., 85, 155, 191
- Dajani, S., 197
 Darley, J. M., 33, 193
 De Neys, W., 170, 171, 180, 190
 Deary, I. J., 13, 42, 190, 191
 DeFries, J. C., 192
 Dennis, I., 170, 193
 Diener, E., 50, 191
 Dolton, P. J., 47, 191
 Doyon, C., 171, 196
 d'Ydewalle, G., 7, 9, 171, 190, 198, 201
- Elias, S. M., 45, 191
 Emerson, M. J., 44, 196
 Epley, N., 38, 189
 Epstein, S., 33, 197
 Erwin, T. D., 45, 191
 Eschman, A., 157, 198
 ESRC, 50, 191
 Ethics Committee of the British Psychological Society, 51, 52, 191

- Evans, J. St. B. T., vi, x, 2, 6, 7, 9, 13, 32–36, 73, 75, 81, 85, 99, 100, 126, 143–147, 149, 151, 154–156, 159, 165–168, 172, 174, 185, 186, 191, 192, 198
- Eyre, R. N., 38, 189
- Facione, P. A., 85, 192
- Farsides, T., 13, 42, 192
- Feinstein, J. A., 101, 190
- Fioratou, E., 170, 174, 193
- Fong, G. T., 21, 24, 192, 197
- Frederick, S., 13, 37, 38, 86, 94, 192
- Friedman, N. P., 44, 121, 170, 177, 192, 196
- Frisch, D., 68, 192
- Gaissmaier, W., 8, 192
- Garcia-Madruga, J. A., 170, 192
- George, D., 61, 192
- Gigerenzer, G., 8, 192
- Gilhooly, K. J., 170, 174, 193
- Gillard, E., vi, 33, 38, 39, 45, 85, 155, 193
- Gilmore, C., 102, 194
- Giroto, V., 20, 21, 26, 71, 72, 77, 193, 197, 199
- Good, C., 52, 189
- Grahn, J. A., 197
- Greenberg, J., 33, 198
- Greene, J. D., 33, 193
- Greenwald, A. G., 33, 193
- Gutierrez, F., 170, 192
- Hampshire, A., 197
- Handley, S. J., 6, 9, 32, 65, 126, 170–172, 177, 192, 193, 196
- Harman, G., 10, 193
- Harper, C., 171, 193
- Hartmann, P., 42, 195
- Heim, A., 26, 193
- Heiman, G. W., 61, 62, 193
- Heit, E., 155, 194
- Hergenbahn, B. R., 15, 194
- Hershey, J. C., 68, 189
- Hewitt, J. K., 192
- Higgins, C. A., 13, 195
- Hitch, G., 189
- Holyoak, K. J., 19, 31, 32, 190
- Houdé, O., 44, 196
- Houston, K., 3, 5, 82, 118, 151, 152, 184, 194
- Howerter, A., 44, 196
- Hoyles, C., 118, 184, 194
- Huckstep, P., 48, 194
- Inglis, M., 5, 26–28, 32, 42, 56, 78–80, 82, 84, 85, 89, 102, 103, 117–119, 124, 141, 174, 183, 184, 194, 221
- Inhelder, B., 16, 30, 194
- Jarvis, W. B. G., 101, 190
- Jensen, A. R., 41–43, 86, 194, 198
- Jepson, C., 65, 197
- Johnson-Laird, P. N., 16, 17, 28–31, 75, 194, 195, 201
- Jones, W., 45, 195
- Judge, T. A., 13, 42, 195
- Kahneman, D., 8, 9, 64–67, 195, 200, 201
- Kao, C. F., 45, 190
- Kemmelmeir, M., 77, 193
- Keough, K., 52, 189
- Kilpatrick, J., 1, 195
- Kosonen, P., 21, 24, 195
- Krantz, D. H., 21, 65, 192, 197
- Kuchemann, D., 118, 184, 194
- Kunda, Z., 65, 197
- LaBerge, D., 186, 195
- Lam, H., 42, 198
- Landsberger, H. A., 23, 195
- Larkin, J., 9, 197
- Larrick, R. P., 12, 195
- Larsen, L., 42, 195
- Lawson, D., 99, 195
- Legrenzi, M. S., 31, 195
- Legrenzi, P., 31, 71, 72, 193, 195, 197
- Lehman, D. R., vi, 20, 24–28, 124, 141, 183, 195, 197
- Lehmann, I. J., 22–24, 54, 196
- Lempert, R. O., 20, 195
- Lewin, I., 167, 196
- Lewis, C., 61, 197
- Locke, J., 1, 84, 183, 196
- Locurto, C., 42, 196
- Loomis, R. J., 45, 191
- Lustina, M. J., 52, 189
- Luzon, J. M., 170, 192
- Mallery, P., 61, 192
- Manfrinati, A., 30, 197
- Manktelow, K., 74, 75, 80, 81, 196
- Markovits, H., 7, 100, 171, 196, 201
- Milford, G., 45, 200
- Miller, P. H., 62, 63, 196
- Miyake, A., 44, 170, 174, 177, 192, 196

- Mook, D. G., 55, 62, 63, 196
 Morgan, J. N., 12, 195
 Morris, N., 170, 200
 Moutier, S., 44, 196

 Nair, K. U., 45, 196
 Nantel, G., 100, 196
 National Research Council, 1, 196
 Neilens, H., 6, 126, 192
 Neilens, H. L., 65, 196
 Neisser, U., 16, 42, 62, 196, 197
 Newstead, S. E., 65, 72, 167, 196, 197
 Nickel, T., 45, 195
 Nilsson, L.-G., 42, 198
 Nisbett, R. E., vi, 12, 15, 19–21, 24–28, 65, 124, 141, 183, 190, 192, 195, 197
 Novick, M. R., 61, 197
 Nyborg, H., 42, 195
 Nystrom, L. E., 33, 193

 Oakley, C. O., 47, 84, 140, 183, 197
 Oaksford, M., 9, 172, 197
 O'Brien, D., 30, 197
 Oliver, L. M., 19, 190
 Olson, M. H., 16, 194
 Oppenheimer, D. M., 38, 189
 Over, D., 127, 192
 Over, D. E., 6, 192
 Owen, A. M., 19, 197

 Pacini, R., 33, 197
 Palipana, A., 89, 194
 Pennycook, G., 154, 200
 Petty, R. E., 33, 45, 101, 190, 197
 Piaget, J., 16, 30, 189, 194, 197, 198
 Pineda, D., 121, 189
 Pittinsky, T. L., 52, 199
 Plato, 1, 15, 84, 140, 183, 198
 Pollard, P., 9, 171, 191, 198
 Polya, G., 5, 26, 198
 Powell, F., 45, 200
 Prowse Turner, J. A., 154, 200
 Pyszczynski, T., 33, 198

 Quinn, D. M., 51, 199

 Ramnarayan, S., 45, 196
 Raven, J., 43, 86, 198
 Raven, J. C., 43, 198
 Reeve, C. L., 42, 198
 Reyna, V. F., 33, 190, 198
 Rips, L., 30, 198
 Rokeach, M., 45, 198
 Rönnlund, M., 42, 198

 Rood, B., 2, 191
 Rosselli, M., 121, 189
 Rotello, C. M., 155, 194
 Rouse, C., 13, 42, 189
 Russell, D., 45, 195
 Ruston, J. P., 43, 198

 Sá, W. C., 9, 13, 21, 24, 35, 80, 81, 83, 85, 88, 100, 101, 127, 198
 Samuels, S. K., 186, 195
 Sanz de Acedo Baquedano, M. T., 22, 198
 Sanz de Acedo Lizarraga, M. L., 22, 24, 198
 Schaeken, W., 7, 9, 16, 33, 155, 171, 190, 193, 195, 198, 201
 Schneider, W., 157, 173, 198
 Schroyens, W., 9, 171, 198
 Sfard, A., 47, 188, 199
 Shafir, E., vi, 23, 199
 Shapiro, D., 76, 201
 Shih, M., 52, 199
 Simpson, A., 5, 26–28, 32, 42, 56, 78, 79, 82, 84, 85, 103, 117–119, 124, 141, 174, 183, 184, 194
 Skinner, B. F., 16, 199
 Sloman, S. A., 33, 34, 199
 Smith, A., xiii, 1, 47, 83, 84, 183, 199
 Solomon, S., 33, 198
 Sommerville, R. B., 33, 193
 Soria Oliver, M., 22, 198
 Spearman, C., 41, 199
 Spencer, S. J., 51, 199
 Sperber, D., 20, 21, 26, 31, 76, 77, 82, 118, 193, 199
 Stanovich, K. E., vi, xiii, 8–14, 22–24, 32, 33, 40–42, 44, 45, 68, 85–87, 101, 112, 121, 125, 127, 128, 154, 157, 165, 169, 181, 198–200, 202
 Steele, C. M., 51, 199
 Stenton, R., 197
 Stich, S., 10, 200
 Stroop, J. R., 173, 200

 The British Psychological Society, 51, 200
 Thompson, V. A., 36, 154, 200
 Thoresen, C. J., 13, 195
 Thorndike, E. L., 16, 18, 19, 21, 22, 24, 200
 Tobacyk, J., 45, 200
 Toms, M., 170, 171, 200
 Tooby, J., 13, 76, 190

- Toplak, M. E., vi, 14, 22–24, 41, 68, 69, 86, 101, 128, 181, 200, 202
- Trenholm, S., 89, 194
- Troldhal, V., 45, 200
- Tversky, A., 8, 9, 65–67, 195, 200, 201
- Vamvakoussi, X., 33, 201
- Van Breukelen, G. J. P., 57, 201
- van der Henst, J. B., 77, 193
- Van Dooren, W., 33, 155, 193, 201
- Venet, M., 7, 201
- Verschaffel, L., 33, 155, 193, 201
- Verschueren, N., 7, 170, 171, 180, 201
- Vigneau, F., 43, 190
- Vignoles, A., 47, 191
- Vila, J. O., 170, 192
- Walport, M., xiii, 1, 47, 84, 183, 201
- Ward, D., 170, 200
- Ward, J., 89, 194
- Wason, P. C., 17, 64, 70, 71, 75, 195, 201
- West, R. F., 9, 11–14, 40, 41, 45, 65, 68, 81, 83, 85, 86, 157, 198–200, 202
- White, E. E., 20, 202
- White, J., 1, 47, 190
- Winne, P. H., 21, 24, 195
- Witzki, A. H., 44, 196
- Woodfield, R., 13, 42, 192
- Woodworth, R. S., 16, 18, 19, 21, 22, 24, 200
- Young, S. E., 192
- Zepp, R. A., 7, 202
- Zuccolotto, A., 157, 198

Nina Attridge
Mathematics Education Centre
Loughborough University
`n.attridge@lboro.ac.uk`

Compiled on 20th March 2013.