Debiasing Reasoning: A Signal Detection Analysis

Nicola Marie Crane MRes

Thesis submitted to for the degree of
Doctor of Philosophy

Department of Mathematics and Statistics
Lancaster University

July 2015

Abstract

This thesis focuses on deductive reasoning and how the belief bias effect can be reduced or ameliorated. Belief bias is a phenomenon whereby the evaluation of the logical validity of an argument is skewed by the degree to which the reasoner believes the conclusion. There has been little research examining ways of reducing such bias and whether there is some sort of effective intervention which makes people reason more on the basis of logic. Traditional analyses of this data has focussed on simple measures of accuracy, typically deducting the number of incorrect answers from the number of correct answers to give an accuracy score. However, recent theoretical developments have shown that this approach fails to separate reasoning biases and response biases. A reasoning bias, is one which affects individuals' ability to discriminate between valid and invalid arguments, whereas a response bias is simply the individual's tendency to give a particular answer, independent of reasoning. A Signal Detection Theory (SDT) approach is used to calculate measures of reasoning accuracy and response bias. These measures are then analysed using mixed effects models. Chapter 1 gives a general introduction to the topic, and outlines the content of subsequent chapters. In Chapter 2, I review the psychological literature around belief bias, the growth of the use of SDT models, and approaches to reducing bias. Chapter 3 covers the methodology, and includes a a thorough description of the calculation of the SDT measures, and an explanation of the mixed effects models I used to analyse these. Chapter 4 presents an experiment in which the effects of feedback on reducing belief bias is examined. In Chapter 5, the focus shifts in the direction of individual differences, and looks at the effect of different instructions given to participants, and Chapter 6 examines the effects of both feedback and specific training. Chapter 7 provides a general discussion of the implications of the previous three chapters.

Acknowledgements

I would like to thank my former supervisor, Professor Linden Ball, for often having more faith in my work than I did myself. Another former supervisor, Professor Tom Ormerod, for your enthusiasm and creativity, even if I did wonder what I'd said when you left a year after Linden did. A huge thank you to Professor Padraic Monaghan for taking over as my third supervisor in as many years. And finally, Dr David Lucy, for all the good advice, statistical and otherwise.

Thanks to Lacey, without whom none of the Amazon Mechanical Turk research would have happened.

I also need to thank Preston Roller Girls; although our relationship was a complicated one, I wouldn't have been submitting this thesis if it wasn't for the lessons you taught me in sheer bloody-minded determination and stress management.

Allie at The Yorkshire House, for keeping me grounded.

All my friends in Lancaster, too many to name, for keeping me (relatively!) sane during the past 4 years, or at least being there for a beer when sanity wasn't quite so forthcoming.

Barry, for being the best Dad ever, and for paying for my masters.

Mum, because "it doesn't matter what you do, as long as you're happy" helped me keep everything in perspective.

Padraic. For all the great ideas, feedback, and reassuring me that despite my worrying, everything was in fact, **not** on fire.

And finally, Matt. You shared in my victories and despairs, and let me ramble on about ideas that would have bored the pants off everyone else. You always just 'got it'.

Declaration

This thesis is my own work and no portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification at this or any other institute of learning.

Nicola Crane 25th July 2015

Contents

List of Figures

List of Tables

Chapter 1

Introduction

We come across vast amounts of information in everyday life, and even more so as the near-ubiquitous availability of internet access is becoming increasingly commonplace. According to Hilbert and López (2011), in 2008, people encountered five times as much information every day than they did in 1986, with this quantity continuously increasing. Given the amount of information which we process daily, the ability to distinguish what is true and relevant from what is not is an important skill if we are to make sense of an increasingly complex world.

This can have serious consequences; for example, the anti-vaccination movement gained a lot of support as the result of a now discredited paper published in 1998 which linked the MMR vaccine to autism spectrum disorders (Wakefield et al., 1998). This led to a drop in uptake of the MMR vaccine, despite numerous studies and large scale meta-analyses showing no evidence of a link between the vaccine and autism (e.g. Jefferson, Price, Demicheli, & Bianco, 2003; Doja & Roberts, 2006; DeStefano, 2007; Taylor, Swerdfeger, & Eslick, 2014). One of the consequences of this was a large outbreak of measles in Swansea in 2013, which led to over 1200 people falling ill and one death. A little knowledge is a dangerous thing.

Clearly, the ability to assess the veracity of ideas and arguments is important, and examining the reasoning process in order to evaluate how it works and potentially even improve it is a worthwhile research pursuit. Lilienfeld, Ammirati,

and Landfield (2009) even go as far as to discuss the potential of debiasing as an important tool in the reduction of idelogical extremism and conflict on a much larger scale. Modern reasoning research really picked up speed with the publication of an influential paper, "Judgement Under Uncertainty: Heuristics and Biases" (Tversky & Kahneman, 1974) and a subsequent volume of the same name a few years later containing the original paper and a collection of other similar research by a variety of authors (Kahneman, Slovic, & Tverky, 1982). This research suggested that many of the judgements made by humans are influenced by systematic biases, which can lead to illogicality. These biases are caused by the use of heuristics; quick and automatic processes which tend to be association-based and require little cognitive effort.

One such bias is the belief bias effect - whereby people tend to subject information which disagrees with their pre-existing beliefs to a higher degree of scrutiny compared to information that conforms with pre-existing beliefs. For example, Lord, Ross, and Lepper (1979) presented participants, who had completed prior surveys about their opinions regarding the death penalty, with information from fictional studies focussed on the use of the death penalty as a deterrent against crime. Both the pro- and anti- death penalty groups showed a tendency to argue that the study which corresponded with their personal belief had greater methodological and empirical validity.

Much of the subsequent research on belief bias has used syllogistic reasoning problems. Syllogisms are simple three statement arguments, containing two premises and a conclusion. They are a useful form of stimuli as they are short, so multiple syllogisms can be evaluated by a single participant in an experiment, and their structure is easy to replicate across multiple studies.

A syllogism might have a believable, unbelievable, or neutral conclusion. An example of each is shown below.

Believable:

Some metals are soft

No soft things are steel

Therefore, some metals are not steel

Unbelievable:

No animals are reptiles

Some reptiles are cats

Therefore, some cats are not animals

Neutral:

Some junarics are lizards

All lizards are panphids

Therefore, some panphids are junarics

Earlier theories of syllogistic reasoning were predominantly *single process* theories; that is, they explained differences in syllogistic reasoning performance as either due to the use of heuristics, quick "rules of thumb", which people apply automatically, or due to analytic processes involving rules and slower more effortful thinking. However, trends in more recent research on belief bias in syllogistic reasoning began with a study by Evans, Barston, and Pollard (1983) who found that, despite the brevity of syllogistic reasoning problems, the belief bias effect still affected participants' reasoning. Participants were asked to evaluate a number of syllogisms as either 'valid' or 'invalid'. Half of these syllogisms were valid - the conclusion did logically lead on from the premises, and half were invalid - the conclusion did not logically lead on from the premises. Within the valid and invalid conclusions, believability was also manipulated - half of the conclusions were believable and the other half unbelievable. The results of this study constitute one of the key findings in belief bias research - an interaction between validity and believability; whilst for valid problems, believability did not influence reasoning, belief bias did affect

the evaluation of invalid problems. For invalid problems, participants were much more likely to deem the conclusion as "valid" if it was believable. This finding of a believability by validity interaction sparked over two decades of research and theoretical development to explain the basis of the phenomenon, as well as the application of dual-process theory to belief bias research. Rather than focus on single-processes; dual-process theories suggest that this interaction is driven by a conflict between heuristic (quick and automatic) and analytic (slower and effortful) processes, although accounts differ in their explanations of exactly how this conflict is detected and resolved. A major methodological shift in this research began in 2010 when Dube, Rotello, and Heit (2010) provided evidence to show that the way in which indices of accuracy and bias had been calculated was flawed. Previously, accuracy and belief bias were measured by simply testing for the effects of validity, believability, and an interaction between the two, often examined through the use of a number of indices. These were linear in nature, calculated by adding and subtracting the number of problems of different types that had been deemed 'valid' by participants. Although theoretical accounts differed in their details, there was a broad agreement that the belief bias effect was due to superior reasoning on unbelievable problems. However, Dube et al. (2010) argued that belief bias primarily manifests as a response bias - participants' tendency to choose the option 'valid' or 'invalid' is independent of any reasoning. They argue that the traditional method of calculation fails to separate reasoning and response bias and instead posited that the use of Signal Detection Theory (SDT) based analyses led to a more accurate model of belief bias as they can properly distinguish these two forms of bias. Subsequent research (e.g. Trippas, Handley, & Verde, 2013) has extended this view, demonstrating that when complex reasoning takes place, for example in participants of a higher cognitive ability or when there are no time constraints, belief bias also has an effect upon reasoning. This constituted a major shift in focus and highlighted the importance of considering individual differences between reasoners in response style when examining belief bias. Later research, such as

that of Trippas, Verde, and Handley (2015) also examined the role of cognitive style; that is, whether individuals have the tendency to rely on heuristic or analytic processing, regardless of their actual ability to do so.

As discussed earlier, debiasing reasoning is an important research aim, and Stupple and Ball (2014) emphasise that understanding and improving reasoning are not mutually exclusive research goals. Thus, debiasing has two roles to play - primarily in identifying methods of reducing or eliminating the extent to which bias affects reasoning, but also to simply aid in further developing theoretical knowledge of the nature of the reasoning process itself. Given that individual differences affect the way in which belief bias affects reasoning, and that even strategy change can be affected by individual differences (e.g. Roberts & Newton, 2003), it is clear that such differences should be accounted for when developing debiasing strategies.

Little research has focussed on debiasing reasoning; the findings of earlier research which aimed to reduce belief bias (e.g. Evans, Newstead, & Allen, 1994; Newstead, Pollard, Evans, & Allen, 1992) has been disputed by more recent work such as Heit and Rotello (2014) who found that SDT analyses show that attempts to reduce reasoning bias only affected response bias. Although Ball (2013) found an effect of feedback on reducing belief bias, once again, the lack of SDT analyses makes these results unclear.

The limited amount of previous research on debiasing that exists has predominantly used traditional indices, making interpretation of these results problematic. The only empirical work which uses SDT is that of Heit and Rotello (2014), which does not examine individual differences. Thus, this thesis provides a novel contribution to the belief bias literature by using SDT analyses to explore potential methods of debiasing reasoning whilst accounting for individual differences.

In Chapter 2, I review a range of literature related to belief bias and syllogistic reasoning experiments. Firstly, belief bias is described in more detail and the structural components of syllogisms are explained, followed by a discussion of general theories of syllogistic reasoning. Following this, I go on to describe the

belief bias effect and compare the different theories which have been suggested to explain this phenomenon, leading into more recent debates regarding the importance of SDT analyses in accurately analysing belief bias data. After summarising the main points in this ongoing debate, Chapter 2 concludes with an overview of debiasing techniques which have been used in the past in order to attempt to improve thinking and reasoning, and reduce or eliminate the effects of bias.

In Chapter 3, I present a detailed discussion of the statistical methodology used to analyse the data in this thesis. Previous approaches to analysing belief bias data are outlined first, followed by an in-depth discussion of SDT models, beginning with theoretical background and moving on to how the various SDT measures are calculated. I explain the importance of using unequal-variance SDT models when they are warranted, along with the use of confidence rating Receiver Operating Characteristic (ROC) curves. Furthermore, I discuss the benefits conferred through use of mixed effects regression models and include formulae for the models contained in this thesis. This is followed by an explanation of the model selection techniques used.

Chapter 4 presents an experimental investigation as to whether the provision of feedback across multiple testing sessions is an effective intervention for debiasing reasoning. This chapter constitutes a novel contribution to the literature, in using an SDT approach to track changes in bias across time.

Chapter 5 is another empirical chapter, and provides a novel contribution by testing the hypothesis that individual differences between participants are crucial to consider when examining debiasing. Here, the effect of differing sets of instructions as a debiasing technique was investigated. This chapter also compares lab-based testing and online testing, and makes a case for the wider use of online experimentation, whilst considering the advantages and drawbacks of both methodologies.

Chapter 6 is the final experimental chapter and investigates the use of specific training and feedback, as a tool to reduce belief bias. Once again, the effects of

individual differences are considered. Also examined here is participants' ability to correctly discriminate between different types of syllogism.

Chapter 7 summarises the conclusions to the research questions presented throughout the thesis, and discusses the possible implications these findings have when considered in the wider context of the underlying theoretical assumptions surrounding the mechanisms underlying belief bias.

An examination of human reasoning cannot be without mention of normativism, the idea that reasoning either does or should conform to rules of logic. This topic is not without controversy and there has been recent debate as to whether it is appropriate to apply normative standards to human reasoning. For example, Elqayam and Evans (2011) argue that when examining human reasoning, using normative standards not only narrows the research questions which are to be asked but also makes unnecessary assumptions, whereas a descriptivist approach, which simply describes the system, could instead be taken. They do concede, however, that when the intention is to improve reasoning, some standard against which reasoning can be measured is necessary. Stupple and Ball (2014) provide a more reasonable middle ground between normativism and descriptivism, which they term "soft normativism", which allows for a normative approach as long as the problems which can be caused by normativism are accounted for, and individual differences between individuals are taken into account. Stupple and Ball (2014) also argue that accounts of reasoning can be strengthened by the use of triangulation; examining multiple measures simultaneously in order to better capture the finer nuances of responses. Question responses alone are an overly simplistic measure of reasoning; given that there is a wealth of evidence showing differences between individuals' strategies, we need to be able to distinguish between a rapid, automatic normatively incorrect response, and one which is slow and effortful yet still incorrect. Although both of these responses may appear identical if only individual responses are examined, the underlying processes behind each of them are likely to be quite different, and this may become more apparent if other information, such as response time, is

also examined. In summary, belief bias and debiasing are both methodologically and theoretically complex topics. In the next chapter, I will review the relevant literature surrounding these areas.

Chapter 2

Literature Review

Although the focus of this thesis is removing bias from reasoning, we first must examine the nature of belief bias and the theories of the cognitive processes thought to underlie the belief bias phenomenon. In addition, it is important to consider the ways in which belief bias is investigated and potential influencing factors which must be controlled for in any empirical investigation of the topic. In this chapter, I first outline the belief bias effect and discuss tasks which have been used to investigate it, focussing upon syllogistic reasoning tasks. This leads on to an in-depth discussion of these tasks, including properties of syllogisms, general theories of syllogistic reasoning, and then goes on to explore theories of the way in which belief bias affects syllogistic reasoning and how methods based around Signal Detection Theory (SDT) have added extra dimensions to the syllogistic reasoning debate. Finally I examine ways in which previous studies have attempted to reduce the effects of bias on reasoning, and categorise and compare the differing approaches.

2.1 Belief Bias and Syllogistic Reasoning Tasks

It has been widely demonstrated that when individuals engage in deductive reasoning, their competence in both forming and evaluating logically valid conclusions can be affected by a number of cognitive biases and errors. When these errors are

systematic in nature, that is, common across multiple individuals and settings, their examination allows us insight into the particulars of the underlying process behind them. One of these potential biases is known as belief bias, defined as when an individual's prior beliefs affects their ability to make an accurate logical judgement about facts that they have been presented with (Evans et al., 1983). The predominant methodology used to investigate this phenomenon involves syllogistic reasoning tasks; individuals are shown an argument containing two premises followed by a conclusion and are asked to evaluate whether the conclusion logically follows on from the premises. When logic and belief are in conflict, individuals' logical judgements are often impeded by the believability of the argument's conclusion (Ball, Phillips, Wade, & Quayle, 2006; Evans, 1989; Newstead et al., 1992; Stanovich & West, 1997; Stupple, Ball, Evans, & Kamal-Smith, 2011). This effect has been shown to be stronger with invalid content, with individuals typically being equally likely to accept valid conclusions whether believable or unbelievable, but more likely to incorrectly accept invalid believable than invalid unbelievable conclusions.

A number of different tasks have been used to investigate belief bias. One such example is an adaptation of the base-rate neglect problem (Kahneman & Tversky, 1973) used by De Neys and Franssens (2009). In this task, participants are given a short description of an individual person chosen at random from a pool of people. A number of characteristics of the person are described, and paint a picture which conforms to various stereotypes. Participants are then asked if the individual being described is more likely to be a member of a group coherent with these stereotypes or another group which differs. Participants are also told that the pool of people which the individual has been selected from contains a certain number of member of each group. An example of such a problem from De Neys and Franssens (2009) is shown below.

> "In a study 100 people were tested. Jo is a randomly chosen participant of this study. Among the 100 participants there were 5 men and 95

women. Jo is 23 years old and is finishing a degree in engineering. On Friday nights, Jo likes to go out cruising with friends while listening to loud music and drinking beer. What is most likely?

A. Jo is a man

B. Jo is a woman."

It has widely been shown that even when it is stated that the pool contains 5% stereotype-conforming and 95% stereotype-opposing individuals, participants will tend to ignore this information and be influenced by their prior beliefs rather than logical probability when choosing which group the individual is likely to be from.

Base-rate neglect problems are more commonly used in judgement and decision-making research, and deductive reasoning studies commonly use categorical syllogistic reasoning tasks. These tasks typically use a conclusion-evaluation paradigm; participants evaluate whether the conclusion to a categorical syllogism logically follows on from the two preceding premises, and are told to accept that the information contained in premises is true (e.g. Evans et al., 1983). This task is known as the conclusion-evaluation task, and the number of conclusions accepted for each of the different combinations of conclusion believability (believable/unbelievable) and validity (valid/invalid) are then submitted to statistical analysis. This gives us four different problem types; valid-believable (VB), valid-unbelievable (VU), invalid-believable (IB) and invalid-unbelievable (IU). Examples of all four types of categorical syllogism using the same content are shown below.

Valid-believable:

Some metals are soft

No soft things are steel

Therefore, some metals are not steel

Valid-unbelievable:

No soft things are metal

Some steel is soft

Therefore, some steel is not metal

Invalid-believable:

No metals are soft

Some soft things are steel

Therefore, some metals are not steel

Invalid-unbelievable:

Some soft things are metal

No steel is soft

Therefore, some steel is not metal

VB and IU problems are often referred to as 'non-conflict' problems (Stupple et al., 2011). This means that for these problems, whether responding on the basis of logic or on the basis of belief, the same response is likely to be given. However, on VU and IB problems, which can be termed 'conflict' problems, there is a clash between logic and belief, or rather, a belief-based response is incongruent with a logic-based response. Research has found that individuals find conflict problems more difficult than non-conflict problems, and are more likely to respond non-normatively to such stimuli (e.g De Neys, 2013). This effect is stronger for IB than VU problems, with many theoretical perspectives on belief bias arguing that is because these problem types are inherently different to each other. Whilst valid problems are determinately valid (i.e. the conclusion is necessitated by the premises), typically, invalid problems which are used in such studies are indeterminately invalid. However, the belief bias effect has also been demonstrated on determinately invalid problems (Klauer, Musch, & Naumer, 2000); this will be discussed in more detail later in this chapter.

A slightly different methodology, the conclusion-production paradigm, requires participants to generate their own conclusions; however, it has been shown that the belief bias effect persists regardless (Oakhill & Johnson-Laird, 1985; Markovits & Nantel, 1989; Oakhill & Johnson-Laird, 1989). Belief bias has also been demonstrated on relational reasoning tasks; in such tasks, participants are given a number of premises which are related on a temporal or spatial basis, and must evaluate the validity of a given conclusion on this basis, for example (from Roberts & Sykes,

2003):

> The Pharaohs ruled after the Romans
> The Pharaohs ruled before the Normans
> The Pyramids were built at the time of the Romans
> At the time of the Normans, William the Conqueror was king Therefore,
> William the Conqueror was king after the Pyramids were built

Despite the difference in overall structure, Roberts and Sykes (2003) found that participants showed similar patterns of results on relational reasoning tasks to those typically found on categorical syllogistic reasoning tasks. In addition, the effect of belief on conditional reasoning has been investigated. There are a number of variants of conditional reasoning problems. For example, Thompson, Prowse, and Pennycook (2011) used the following format:

> If the TV is plugged in, then it works The TV works, therefore, it is plugged in

However, Santamaría, García-Madruga, and Johnson-Laird (1998) used double conditionals, which bear more similarities to categorical syllogisms e.g.:

> If Pablo is at home, then he watches the news. If Pablo is at home, then he turns on the television. Therefore, if Pablo turns on the television, then he watches the news.

In both cases, regardless of the specific nature of the task completed, belief bias effect was still found to have an effect upon responses, providing further evidence for the generality of the belief bias effect. The widespread use of categorical syllogistic reasoning tasks can be attributed to the fact that categorical syllogisms are unambiguously valid or invalid, and their properties can be described in a standardised way, thus making different studies by different authors easily directly comparable. Although this type of task is somewhat removed from everyday

experience of deductive reasoning, the empirical benefits mentioned above along with the generality of the belief bias phenomenon to a variety of different contexts lead many to argue that, for these reasons, syllogistic reasoning tasks are an invaluable tool for investigating the underlying processes behind deductive reasoning (e.g. Oakhill & Johnson-Laird, 1985).

2.2 Properties of Syllogisms

Syllogisms are logical arguments which, typically, contain two premises followed by a conclusion. Depending on the structure and content of the premises, the conclusion may or may not be valid. The structure of a syllogism can be described in terms of figure and mood (Evans et al., 1983). The figure is the order in which the individual terms within the syllogism are presented. An example of a syllogism is: "some metals are soft; no soft things are steel; therefore, some metals are not steel". Here, 'metals' are the predicate of the conclusion, 'soft things' are the middle term, and 'steel' is the subject of the conclusion. For the purpose of simplicity, these terms are often denoted as 'A' (predicate of the conclusion), 'B' (the middle term), and 'C' (subject of the conclusion). The predicate and subject of the conclusion each appear once in either the first or second premise, but not together, whereas the middle term appears once in both. The different combinations of the terms results in four possible figures for the premises; ABBC (Figure 1), BACB (Figure 2), ABCB (Figure 3), and BABC (Figure 4). These pairs of premises then will be followed by a conclusion of either AC or CA figure. In the example above, the premises are in the ABBC figure, and the conclusion is in the AC figure.

The use of the phrasing 'some...are...' in the first premise of the example syllogism above is termed the affirmative existential mood. There are four different possible moods, which are commonly denoted by the letters A, E, I and O (Evans et al., 1983); affirmative universal ('all X are Y' - A), negative universal ('no X are Y' - E), affirmative existential ('some X are Y' - I), and negative existential ('some X are not Y' - O).

The combination of the possible different moods and figures gives a total of 512 different potential structures for a syllogism, with 64 different pairs of premises. However, out of these 64, not all have a valid conclusion that logically leads on from the premises. The traditional Aristotelian approach identifies 24 valid categorical syllogisms, as shown in Table 2.1.

Figure			
1	2	3	4
AAI	**AAA**	**AEE**	***AAI***
AEE	*AAI*	*AEO*	**AII**
AEO	**AII**	**AOO**	***EAO***
EAO	**EAE**	**EAE**	**EIO**
EIO	*EAO*	*EAO*	**IAI**
IAI	**EIO**	**EIO**	**OAO**

Table 2.1: Valid Syllogisms: Aristotelian definition

It should be noted that the enumeration of syllogistic figure in Aristotelian terms differs from the convention used in the modern syllogistic reasoning literature (Johnson-Laird & Byrne, 1991). For the sake of simplicity, the modern enumeration has been used. The syllogisms in italics are ones which, if a category they mention is empty, are invalid. In other words, if any of the categories mentioned in a premise don't actually contain any members, the conclusion drawn is no longer valid. The syllogisms in bold italics in Table 1 are ones which can be said to be containing weakened moods, i.e. a more definite conclusion can be drawn. For example, given an AA premise pair with a valid A conclusion, one could argue that an I conclusion could technically be also correct as it is a subset of A.

Johnson-Laird and Bara (1984), however, identify 27 possible valid premise pairs, but 29 valid premise and conclusion combinations as some premises have more than one valid conclusion. This number is not taking into account the order in which the terms in the conclusion are presented; whether in AC order or CA order. The difference from the Aristotelian total comes from the fact that Johnson-Laird and Bara (1984) exclude syllogisms which can have weakened moods, and include syllogisms which can be concluded with AC conclusions (the Aristotelian method

only includes those which can be concluded in CA order). These syllogisms can be found in Table 2.2. Nine of these syllogisms are valid in both AC and CA conclusion order, leading to a total of 48 unique valid syllogisms when mood and figure for the premises and conclusion order are taken into account.

Figure			
1	2	3	4
AAA*	AAA^	AEE	AAI
AAI^	AAI*	AOO^	AEO*
AEE	AII	EAE	AII
EAO^	AEO*	EIO^	AOO*
EIO^	EAE	IEO*	EAO^
IAI	EIO^	OAO*	EIO^
IEO*	IEO*		IAI
			IEO*
			OAO^

Table 2.2: Valid syllogisms, Johnson-Laird & Bara (1984) * =valid only in AC mood ^=valid only in CA mood

2.3 Theories of Syllogistic Reasoning

Before discussing theories of belief bias in syllogistic reasoning, we must first examine the more general theories of syllogistic reasoning. There are a number of consistent effects and patterns of responses in terms of typical errors which reasoners make which have been shown to occur in both conclusion generation and conclusion evaluation paradigms. These errors appear routinely across numerous studies, and some categorical syllogisms are easily solved or evaluated whereas others have been found repeatedly to be difficult. A number of theories outline typical errors, and give an account of potential underlying causes.

It has been shown that *figural effects* can have an impact upon individuals' competence at assessing whether or not a syllogism has a valid conclusion or not. Research has demonstrated that individuals find some figures easier to reason with than others (e.g. Johnson-Laird, 1975; Johnson-Laird & Steedman, 1978). Participants tend to give conclusions in the order AC for Figure 1 premises, and

CA for Figure 2 premises, with a smaller preference for AC conclusions for Figures 3 and 4 (Khemlani & Johnson-Laird, 2012). Johnson-Laird and Bara (1984) argue that this is because reasoning is simplest when the middle 'B' terms are contiguous (i.e. adjacent to one another). Therefore, Figure 1 is the easiest, with Figure 2 being slightly more complex, requiring reordering of the premises to bring the middle terms together. Figures 3 and 4 are more difficult still, with reordering of the terms within one of the premises also being required in order for the middle terms to become contiguous. If a valid conclusion requires the terms to be given in an order which goes against the direction of this effect, participants find it difficult to produce a correct answer.

The *atmosphere* account of syllogistic reasoning (Woodworth & Sells, 1935; Revlis, 1975) argues that individuals' perceptions of the validity of the conclusion of a syllogism depends heavily upon the mood of the premises. If the two premises are in the same mood as each other, the conclusion drawn is likely to also be in that mood. When different moods are present, if a negative mood (i.e. E or O) is used in one or more of the premises, reasoners are more likely to accept a conclusion which also has a negative mood. In addition, if an existential mood (i.e. I or O) is used in one or more of the premises, the most likely conclusion to be drawn will be in the existential mood.

Similarly, the *matching* account (Wetherick & Gilhooly, 1995) agrees that the mood of a syllogism will influence individuals' selection of conclusion. However, this account argues that the atmosphere account is flawed due to problems with the methodology used by Sells (1936) on which much of this theory is based. Instead, Wetherick and Gilhooly (1995) argue that it is the number of category members that the mood describes which is important in determining variation in reasoners ability to correctly deduce valid conclusions. Wetherick and Gilhooly (1995) argue that the individuals will tend to match the conclusion to the most conservative premise (i.e. the one which makes inferences about the smallest number of category members). The negative universal mood ("no X are Y") is the most conservative, followed by

the negative existential ("some X are not Y") and affirmative existential ("some X are not Y") moods, which are equally conservative, and then the affirmative universal mood ("all X are Y"), which is the least conservative. Wetherick and Gilhooly (1995) do not argue that matching is the only way in which individuals engage in syllogistic reasoning, but rather clarify that this is just one of many possible strategies (Khemlani & Johnson-Laird, 2012).

The *conversion* account (Chapman & Chapman, 1959; Revlis, 1975) postulates that errors in syllogistic reasoning can be accounted for by illicit conversion. The terms in premises in moods E ("no X are Y") and I ("some X are Y") can legitimately be reordered as they are logically equivalent and so terms X and Y are interchangeable here. However, this is not the case for moods A ("All X are Y") and O ("Some X are not Y"); for example, in the case of the former, although all instances of X must also be Y, this does not preclude the existence of instances of Y which are not X. Thus, erroneous conversion of premises can lead to incorrect conclusions being drawn. Although evidence has shown that when participants are given instructions explaining that this sort of conversion can lead to errors then their performance improves (Dicksten, 1975), it has also been highlighted by Johnson-Laird and Byrne (1991) that if conversion was a universal error which all reasoners engaged in on all syllogisms then there should be no evidence for figural effects. To illustrate this point, take the example of a syllogism in Figure 1 (ABBC), with the second premise in the O ("some X are not Y") mood, and the first premise in a legitimately convertible mood. If universal conversion were the case, there should be no difference between whether this syllogism is presented in Figure 1 (ABBC) or Figure 3 (ABCB), given that conversion will occur anyway. However, systematic differences in correct solution rate dispute this.

Chater and Oaksford's (1999) Probability Heuristics Model has a few similarities to the matching account in that it ranks moods in terms of informativeness. The term "informativeness" comes from Information Theory (Shannon & Weaver, 1949) and the informativeness of a given statement has an inverse relationship to its

probability, and is related to how often a statement is likely to be true in natural language. Unlike the matching account, however, Chater and Oaksford (1999) employ a different ranking, with "all X are Y" premises ranked as the most informative, followed by "some X are Y", then "no X are Y", and finally "some X are not Y". This model has a greater number of complexities than other heuristics-based models of syllogistic reasoning. Chater & Oaksford discuss a number of heuristics which reasoners use, which predict how likely a conclusion is to be chosen. Firstly, the min-heuristic leads to reasoners drawing conclusions in moods which are the same as the premise with the least amount of information contained within them. The p-entailment heuristic means that the conclusion which is probabilistically entailed by the conclusion, for example, conclusions in mood A entail those in mood I as "some" is entailed by "all", and so the conclusion next most likely to be selected by reasoners is the one which is p-entailed by the min-heuristic. The attachment heuristic determines the ordering of terms in the conclusion; if the min-premise, that is, the premise which is least informative, has term A or C as the subject of the sentence, then this term will appear as the subject in the conclusion. If it does not appear as the subject of the min-premise, then the A or C term which is the subject of the max-premise (most informative premise) becomes the subject of the conclusion. Chater & Oaksford also discuss a further two heuristics which are suggested to aid reasoners in assessing how accurate the conclusion generated according to the above principles is. The max-heuristic leads to higher confidence for conclusions with more informative premises and lower confidence for premises with less informative premises. The O-heuristic leads reasoners to be less likely to choose a conclusion in the O mood, as this is the least informative mood. The probability heuristics model has been lauded by Khemlani and Johnson-Laird (2012) as it is a detailed explanation of syllogistic reasoning, which accounts for figural effects, and can be extend further than simply problems with A, E, I and O moods to describe reasoning about problems which contain more specific quantifiers such as "most", "few", and others. However, it has also been criticised by the same

authors for not fully explaining the role of logic when participants give normative answers that diverge from these heuristics, and cannot account for improvements in performance over time. If reasoners are mainly employing a particular heuristic, one would expect their performance to remain stable over time, and to generally show low levels of normative responding.

Some theories of syllogistic reasoning places much more emphasis on the role of logic rather than heuristics in reasoning. Such theories argue that logical reasoning is an innate human ability, and Rips (1994) argues that reasoners reason in accordance with formal inferential rules including those such as modus ponens and modus tollens. Modus ponens is a rule, whereby, if the statement "if A, then B" is given, then if the existence or truth of A is stated, the existence or truth of B must logically follow. The same statement, but with the non-existence of B, logically A cannot exist either. Some rules operate in a forwards direction, linking premises to conclusions, others operate in a backwards direction, linking conclusions back to their premises, and others can work in either direction. Errors in syllogistic reasoning are argued to stem from improper or non-existent application of the relevant rule, and difficulty is said to be based upon the number of rules which are needed to be used. This theory has been used to construct a computational model PSYCOP which simulates these tendencies.

Many of the theories discussed above focus mainly on how people reason about syllogisms. However, other theories place more emphasis on examining how syllogisms are mentally represented in order to make inferences about the specific nature of the reasoning process. Johnson-Laird & Byrne (1991) argue that convincing theories explaining syllogistic reasoning rely upon the individual constructing mental models or representations of the premises in order to assess whether the associated conclusion is valid, or to generate their own conclusion. They cite previous research which has posited that these representations are in the form or Euler circles or Venn diagrams. Johnson-Laird & Byrne argue that these theories do not explain variations in difficulty; premise pairs which, when mapped out using

Euler circles, should generate a larger number of possible representations are not necessarily those found to be most difficult, and individuals struggle with some premise pairs which have a very small number of valid diagrammatic representations. Johnson-Laird & Byrne (1991) give the example of the premises "Some A are B, All B are C" and "All B are A, No C are B". They argue that to fully map out all the possible conclusions for these examples with Euler circles would require 16 combinations for the former and 6 for the latter, and so the second premise pairs should be simpler to solve. However, in fact, they cite evidence which shows that 88% of people correctly solve the first syllogism, but only 8% correctly solve the second, and conclude that this seriously undermines the credibility of these representations.

Johnson-Laird & Byrne (1991) argue that the commonalities of most model-based theories are that they all generally agree that reasoners begin by generating models representing the premises in the simplest possible way. They then go on to draw a conclusion based on this information, and then add in further complex details to these representations during a search for counterexamples, in an attempt to falsify the conclusion they are testing out. If the conclusion is still found to hold true after such comparisons, it is deemed to be valid. Johnson-Laird & Byrne (1991) discuss their *mental models* theory of syllogistic reasoning and argue that individuals use symbolic representations of instances of category members described in the premises in order to assess or construct a conclusion. These representations denote members of the categories mentioned in the premises and aim to represent various possible combinations of group membership or exclusion. Extra premises are incorporated by simply adding them on to the representation of the existing premises in any way as long as it does not invalidate them. Johnson-Laird & Byrne argue that their theory better accounts for difficulty than those dependent on Euler circle or Venn diagrams, as in their theory, difficulty is related to the number of possible models that can be used to represent the premises. Additionally, difficult syllogisms are those in which the valid conclusion is only found in one of

the multiple representations.

In their *verbal reasoning theory*, Polk and Newell (1995) argue that, given the relative sophistication of humans' linguistic processing ability, reasoning is more verbally based than alternative theories suggest. Deductive reasoning relies on the use of syntactic rules, and re-encoding of the premises a number of times, until a conclusion can be generated or verified. If the conclusion is consistent with the model generated by the encoding of the premises, it is accepted as valid. Computational models VR1-VR3 were constructed to explain this process. Model VR1 is the simplest; any information which cannot be inferred indirectly from the premises is not included. Because of this, this model has been found to lead to a large proportion of "no valid conclusion" responses. VR2 extends VR1 by allowing the legitimate conversions of "some X are Y" to "some Y are X" and "no X are Y" to "no Y are X". VR3 allows more of what is termed "indirect knowledge", for example, if we were to state that "all Y are Y", then we can infer that "no things that are not X are Y". This theory has some aspects in common with mental models theory, discussed below; however, it explicitly rejects the possibility of reasoners constructing counterexamples to falsify potential conclusions. A later version of this theory, *modified verbal reasoning theory* (Thompson, Striemer, Reikoff, Gunter, & Campbell, 2003), also adds in further conditions, for example, that reasoners allocate a finite amount of time to the encoding and re-encoding process before giving up. However, these theories differ from mental models theory, as it does not account for the possibility that individuals may construct counterexamples.

Although mental rules and logic based theories have been criticised for underplaying the role of heuristic processes (Khemlani & Johnson-Laird, 2012), it is not necessarily the case that such theories argue that it is reliance on mental rules alone that influence reasoning (Braine & OBrien, 1991). There is, however, some evidence that reasoners, to some extent, use rules-based techniques in reasoning. Ford (1995) found support for this theory using protocols in which participants were asked to discuss their answers with the experimenter, and found that some

engaged in verbal substitutions, in which terms were interrelated or substituted from one premise to another. Errors occur when reasoners use substitutions which are not logically consistent with the ways in which the terms in the premises are interrelated. Revising earlier ideas dismissed by mental models theorists, Ford (1995) has argued that other reasoners may use a different version of Euler circles to those described by Johnson-Laird and Byrne (1991). This version does not suffer the problem of the number of possible representations differing from the perceived difficulty. Johnson-Laird (2005) argues that this theory is indistinguishable from their mental models theory. However, Ford (1995), makes a few key distinctions. Firstly, Ford criticises the way in which Johnson-Laird & Byrne (1991) categorise syllogisms as single or multiple-model syllogisms, arguing that there are inconsistencies between their definitions and those in earlier work, such as that of Johnson-Laird & Bara (1984). Ford specifically highlights the example of a model which is originally classed as a single model syllogism, but in later work, reclassified as a multiple model syllogism. Ford argues that Johnson-Laird & Byrne (1991) give an explanation for model construction for this syllogism which is inconsistent with explanations they give for other similar examples, and implies that this explanation has been given as the authors are unable to explain participants relatively low accuracy rates on what should be a simple single model syllogism. Ford goes on to detail how this syllogism is the only single-model syllogism in which the conclusion is not in the same mood as either of its premises, and given that reasoners have a tendency to produce or select conclusions with a mood matching one of the premises, this would explain the low accuracy rate on this particular problem, which is in line with accuracy rates for multiple-model problems, all of which are only valid with moods which do not correspond to one of the premises.

Ford also argues that despite very little evidence to support it, mental models theory suggests that individuals' mental representations are of individual members of a group, whereas alternative theories which posit individuals constructing representations of sets as a whole may be more accurate. Ford provides evidence for

this claim from an experiment in which the conclusion generation paradigm was used. Participants were shown the 27 premise pairs which led to a valid conclusion, and were instructed to derive a conclusion for each. They were instructed to think out loud as they completed the problems, and were also given a pen and paper with which they were allowed to make notes. Once they had completed all 27 problems, participants gave a written explanation of their answers. Ford found that some participants, whom she termed 'spatial reasoners' reasoned using diagrammatic representations, many similar to Euler circles, which tended to represent entire sets rather than particular members of a set. Other participants, labelled 'verbal reasoners', used verbal substitution techniques which Ford argues worked on algebraic and rules-based techniques, with some using a fairly basic version of this, and others employing more complicated techniques. The distinction between these two types of reasoners also accounted for differences in solution rates for different types of syllogisms, with spatial and verbal reasoners showing wide gulfs in proficiency on differing syllogisms. Ford argues that these results show that the differences between spatial and verbal reasoners must be accounted for when examining syllogistic reasoning data. However, others such as Rips (2002) suggests that people only use diagrammatic methods like those used by the spatial reasoners if they have been taught how to use them in the past. In addition, (Stenning & Oberlander, 1997) highlight how it can be shown that Euler circles can be seen to be equivalent to mental models theory in terms of how people use representations to reason about syllogisms at a computational level. A similar inference is made by Stenning and Yule (1997) comparing Euler circles and mental rules.

There are a large number of current theories which attempt to account for difference in reasoners' performance on different syllogisms, which seem in part to explore different aspects of syllogistic reasoning, for example, what drives normative responding and what drives erroneous responses. This may appear to make comparison between some of these theories somewhat difficult. A meta-analysis conducted by Khemlani & Johnson-Laird (2012) found some support for heuristic

theories of syllogistic reasoning, and also greater support for mental-models based theories than rules-based theories. Thus, they concluded that syllogistic reasoning is likely to operate under the constraint of a number of heuristics which determines the content of what is mentally represented, but the mental representations themselves deliberated between and reasoned about using some sort of set or model. They also highlight how, as strategies may vary between different individuals, that it may not be as simple as a single one of the current theories being able to account for all syllogistic reasoning. Research on the belief bias effect seems to tie these two areas together, and shifts the focus away from the specific nature of heuristics and analytic processes, and towards the interplay between the two in providing a response. I will now go on to consider theories of belief bias, including dual-process theories, and the extra insight which they give us into the deductive reasoning process.

2.4 Theories of Belief Bias

Some earlier theories of belief bias adapt more general theories of syllogistic reasoning to explain this phenomenon. Revlin, Leirer, Yopp, and Yopp (1980) argues that belief bias can be beneficial to reasoning, as it can prevent conversion of the premises from occurring. Individuals are much more likely to erroneously convert the terms in premises containing abstract terms than make the same mistake with sentences containing increased semantic content such as "all cats are animals" and "all animals are cats". However, subsequent research (e.g. Evans et al., 1983; Oakhill & Johnson-Laird, 1985) controlled for conversion by using premises in mood E (i.e. "some X are y") and mood I (i.e. "no X are Y"), for which conversion does not alter the validity of potential conclusions to be drawn. The belief bias effect was found to persist under such conditions, indicating that the conversion account was not an adequate explanation.

Khemlani and Johnson-Laird (2012) highlight how, in other areas of human cognition, many theories have been concerned with the role of Type 1 and Type 2

process, and later theories of belief bias predominantly fall within this dual-process framework. These theories stipulate that the belief bias effect arises from the interplay between two different types of processes. The terminology used to describe these processes varies from one account to another, with Type 1 processes also being labelled as heuristic, implicit, associative, impulsive, automatic, experiential, non-conscious, intuitive, or reflexive, and Type 2 processes being described as logical, explicit, rule-based, reflective, controlled, rational, conscious, analytic, or reflective (Evans & Stanovich, 2013). It must be noted that some theorists draw a distinction between accounts which characterise these different sorts of cognitions as belonging to different systems, and those which describe them as simply being the result of separate processes. Evans and Stanovich (2013) caution against the 'system' designation, given that there is some evidence to suggest overlap between multiple systems and multiple systems contributing towards output. In addition, they argue that it is crucial not to confuse the concept of features which are defining properties of a system with features that are often correlated but do not necessitate cognition being due to one specific process or the other. Evans & Stanovich argue that the most important defining feature of Type 1 processing is its autonomous nature. Whether or not a Type 1 response is given, Type 1 processing will always be engaged in, given the relevant stimuli. A further feature of Type 1 processing is the lack of demands it places on working memory resources. On the other hand, Type 2 processing is characterised by its higher working memory demands, and its use in "cognitive decoupling", which will be discussed in more detail below. Longer response latencies on problems have been thought to be indicative of Type 2 processing, although this assumption has been challenged. For example, Handley, Newstead, and Trippas (2011) found that when participants were asked to respond on the basis of believability, they took longer to make such judgements than judgements made on the basis of logical validity. Dual-process theories differ to each other in terms of the point at which belief bias is thought to affect a reasoner's response to a reasoning problem they are trying to solve. Whereas earlier accounts

such as conversion (Revlin et al., 1980) argue believability affects the encoding of problems, other accounts place it in the reasoning stage, response stage or both. Although traditional analyses cannot easily make this distinction, SDT analyses allow us to get insight into where in the reasoning process is affected by belief bias. In addition, accounts differ in that some theories argue that believability affects *amount* of Type 2 processing which occurs, whereas others argue that it has a strong effect on the *type of* strategy employed by participants. I will now review the a number of theories of belief bias, and the empirical evidence for and against them.

Dual process theories of belief bias began with work by Evans et al. (1983). They criticised earlier accounts of belief bias, such as that of Revlin et al. (1980), for using inappropriate methodology and insufficiently exploring or controlling for other things thought to have an impact on reasoning, such as conversion of the premises, figural bias, atmosphere, and validity or determinacy. In their empirical work, Evans and colleagues accounted for these shortcomings, and their data led them to develop the *selective scrutiny* account, which suggests that individuals adopt different strategies on the basis of conclusion believability. If a conclusion is believable, it is accepted, with no further analysis. However, in the case of an unbelievable conclusion, the problem is analysed logically, with valid problems being accepted and invalid-believable problems being rejected. Evans et al. (1983) support this theory by highlighting qualitative findings from their reasoning-aloud protocol. Participants who were incorrectly endorsing invalid problems tended to make more reference to information not relevant to the task, whereas those who were reasoning normatively discussed the premises of the syllogism, suggesting a difference in the degree to which analytic processing was engaged. However, the selective scrutiny account has been challenged by more recent work examining response latencies, that has shown that participants show longer response times, which is associated with Type 2 processing, on invalid-believable problems (Ball et al., 2006; Thompson, Morley, & Newstead, 2011), which would not be the case if

believable conclusions were simple accepted without any further analysis.

An alternative explanation of these results, also proposed by Evans et al. (1983), has been termed the *misinterpreted necessity* model. In this explanation, it is argued that reasoning operates prior to the effect of logic. Reasoners first assess whether the conclusion falsifies the premises (i.e. if it is determinately invalid). If it does, it is rejected outright. However, if not, the reasoner then goes on to attempt to decide if the conclusion is determined by the premises (i.e. if it is determinately valid). If this is the case, the conclusion is accepted as valid. However, if it is not determinately valid, then the conclusion believability is used to resolve this ambiguity, with believable conclusions being accepted and unbelievable conclusions being rejected. As valid syllogisms are determinately valid, they cannot be falsified, and thus have a generally high acceptance rate. However, the invalid syllogisms used in the study were indeterminately invalid; their conclusion is possible, but is not necessitated by the premises, and therefore the use of believability to guide reasoning on invalid problems leads to higher numbers of invalid-believable conclusions and low number of invalid-unbelievable conclusions being accepted. Support for this theory comes from evidence to show that when participants are given instructions that emphasise the importance of logical necessity, the belief by validity interaction is diminished (Evans et al., 1983, Experiment 3). However, it has been criticised as it cannot explain the presence of the belief bias effect on determinately invalid syllogisms (Newstead et al, 1992).

The *mental models* account (Oakhill, Johnson-Laird & Garnham, 1989) extends the earlier mental models theory of syllogistic reasoning to incorporate belief bias. This theory argues that individuals construct a single mental model incorporating the premises of a syllogism. If the mental model they have constructed is inconsistent with the given conclusion, the conclusion is deemed to be invalid. However, if is consistent, then believability is examined. At this stage in the reasoning process, believable conclusions are accepted as valid, whereas unbelievable conclusions are subjected to further analysis in which the individual attempts to construct all

possible model of the premises. If the given conclusion is true in all models, it is accepted, otherwise it is rejected. This theory explains the difference in difficulty found between single-model and multiple-model syllogisms as being a consequence of the increased demands on working-memory memory resources for multiple-model syllogisms. However, Ball et al (2006) found that response latencies did not support the claim that unbelievable problems result in a greater degree of logical analysis. Ball and colleagues got participants to respond to categorical syllogisms whilst attached to equipment which tracked the direction of their gaze. This methodology allowed the timing of attention paid to specific parts of the syllogism. Whilst mental models theory would have predicted longer processing times for unbelievable problems, due to the extra effort spent on model construction for such problems. However, the data showed that participants actually spent longer responding to believable problems, in direct opposition to this prediction.

The *metacognitive uncertainty* account (Quayle & Ball, 2000) is similar to both the misinterpreted necessity and mental models accounts, in that it agrees that reasoners construct mental models of syllogisms, and that belief operates after some reasoning has taken place. However, it differs in that it argues that when the reasoner's working-memory capacity is exhausted, due to a larger number of generated models for invalid conclusions, a state of uncertainty in induced. If their confidence is below a certain threshold, this will prompt them to rely on conclusion believability to guide their response. This is reflected in lower confidence ratings given by participants when responding to invalid than to valid syllogisms (Experiment 1), and participants with lower scores on a working memory test being more influenced by belief bias than higher scoring participants (Experiment 3). However, eye tracking data provided by Ball et al (2006) was inconsistent with predictions of this theory, regarding time spent examining premises and conclusions. Whilst metacognitive uncertainty theory would have predicted longer premise inspection times for invalid problems, Ball and colleagues found that problem validity was not predictive of response time.

Many of the aforementioned accounts suggest that belief bias affects the amount of logical reasoning that occurs. In these theories, the interplay between heuristic and analytic systems is determined by whether Type 2 processing occurs at all, or whether the output of Type 2 processing is deemed sufficient information for a response to be given. On the other hand, selective processing theory (Evans, 2000; Klauer et al., 2000) argues that all individuals engage in Type 2 processing, but do it in qualitatively different ways depending upon conclusion believability. Selective processing theory is somewhat in agreement with mental models theory in that individuals are thought to construct a single model of the premises. However, it posits that qualitatively different models are constructed on the basis of believability; a confirming model is constructed for believable conclusions, and a disconfirming model is constructed for unbelievable conclusions. Empirical evidence involving response latencies (Ball et al., 2006; Stupple & Ball, 2008; Thompson et al., 2003) and time-limited responses (Evans & Curtis-Holmes, 2005) support this theory. For example, Ball et al (2006), using an eye-tracking methodology were able to analyse how much time was spent examining the separate parts of a syllogism. They found no evidence for any differences in response times before the conclusion was viewed, and differences only arose subsequently. This led to the conclusion that, at least in a conclusion evaluation paradigm, this provides strong support for conclusion-to-premises theories such as selective processing theory, but refutes premises-to-conclusion theories, such as mental-models theory.

Nevertheless, this version of selective processing theory cannot account for instances in which a smaller subset of participants show almost perfectly normative responses. Thus, Stupple et al (2011) posit an extended version of the theory. The key distinction between the original model and the revised model is that the latter allows analytical reasoning, if heuristic responding is suppressed, to go down one of two pathways. Individuals either perform an exhaustive search or a *satisficing* search. The word 'satisficing' is a portmanteau of 'satisfying and 'sufficing' and this kind of search does exactly that it looks for a model that satisfies the criteria of

either rejecting or accepting the conclusion, and does not perform a comprehensive search, but instead conducts one which is sufficient to yield a response within time and memory constraints. It may be that individuals choose to reject or accept the given conclusion as soon as they find a model which is consistent with their initial heuristic response. An exhaustive search involves a more in-depth and wide-ranging search of possible models to examine whether the conclusion in question is necessarily valid or not.

It has been highlighted that reasoners may be more aware of their own biased responses than the above theories might suggest. De Neys, Moyens, and Vansteenwegen (2010) found that participants showed higher level of autonomic arousal, measured via skin conductance rates, when they encountered conflict problems. It is tempting to suppose that this could be accounted for by the *parallel process* model suggested by Sloman (1996), in which it is argued that both Type 1 and Type 2 reasoning take place in parallel. Conflict between heuristic and analytic output is detected when the two types of processing give different responses; however, due to the speed of Type 1 responses, a Type 2 response is less likely to be given. This model, however, has been criticised as there would be no benefit of employing Type 2 processing when Type 1 would suffice, in the case of non-conflict problems (Handley & Trippas, 2015).

De Neys and Bonnefon (2013) argues that differences in the extent of belief bias are due to differences in storage, monitoring, or inhibition. A storage failure is one which is present due to the lack of knowledge of formal rules of logic. Monitoring failure accounts suggest that belief bias arises from the inappropriate use of heuristic thinking resulting in non-normative responding. Earlier theories of belief bias such as selective scrutiny can be seen as monitoring failure accounts; these theories do not allow for reasoners being able to detect a conflict between logic and belief. Inhibition failure accounts argue that the vast majority of individuals are capable of formal reasoning. However, differences in the ability to suppress a heuristic response lead to belief based responding.

De Neys and Bonnefon (2013) argues that this override mechanism is logical in nature, but also implicit. The logical intuitions account of belief bias is a hybrid of serial and parallel process models, and suggests that heuristic and logical processes, which are both intuitive in nature, are initially activated. However, it is only when a conflict between these processes is detected that logical deliberative processing happens. They argue that this intuitive logic is acquired early in human development, and highlight some studies which show high levels of logically normative responding in very young children (e.g. Brainerd & Reyna, 2001).

Handley et al. (2011) highlight a paradox within the default interventionist approach. Such theories characterise heuristics as a default response, which can be overridden by analytical processing. When a conflict is detected between belief and validity, this can lead to the additional Type 2 processing. However, the detection of the conflict in the first place already requires some type of Type 2 processing to have already taken place, thus making the theories logically inconsistent. However, research by Thompson, Prowse-Turner, et al. (2013) suggests that rather than being some sort of Type 2 process, this mechanism is a separate entity entirely. Thompson and colleagues asked participants to given an initial quick response to a number of reasoning problems, which was followed by the opportunity to change this response. The term *answer fluency* was used to define the amount of time it took to give this initial response. Answer fluency was found to be predictive of how likely a participant was to engage in Type 2 thinking, operationalised by Thompson and colleagues as the amount of time spent rethinking an answer, or the chance of changing an answer. It was not, however, predictive of accuracy. It was argued that it is metacognitive cues that are based upon the ease with which an initial heuristic response can be given that determines conflict detection. When a heuristic response is given with great ease, and thus high confidence, an accompanying analytic response is less likely to occur. This also once again raises the question of when in the reasoning process belief bias occurs.

In summary, earlier theories of belief bias argued that this phenomenon was the

result of either belief operating prior to reasoning (e.g. selective scrutiny), belief resolving uncertainties in reasoning (e.g. misinterpreted necessity, metacognitive uncertainty), belief affecting the extent of processing which takes place (e.g. mental models). However, more recent theories, such as selective processing theory argue that belief affects the type of reasoning which takes place. Evidence concerning autonomic arousal and confidence suggests that even when reasoners give a biased response, they have some sort of underlying awareness of this (e.g. De Neys et al., 2010; De Neys, 2013). Other evidence, such as the inspection-time findings of Ball et al (2006) further supports this view that reasoners are competent at detecting conflict between logic and belief, and thus that typically both Type 1 and Type 2 processing take place. Despite this evidence in favour of belief bias taking effect during the reasoning process, there is reason to believe it also manifests as a response bias, and these theories need not be mutually exclusive (Trippas et al., 2013). This will be discussed in the next section. As shown in the research of Thompson et al. (2003) and Ball et al. (2006) amongst others, the use of a wider range of methodology than simply calculating the number of correct responses has led to developments in theorising about belief bias. More recent methodological developments have been concerned with the statistics used to measure performance on syllogistic reasoning tasks. In the next section, I will discuss the MPT model of Klauer et al. (2000), and go on to examine SDT models, such as those of Dube et al. (2010) and Trippas et al. (2013).

2.4.1 Separating Reasoning and Response Bias

In developing their selective processing model of belief bias, Klauer et al. (2000) highlighted an important nuance of belief bias that had been previously overlooked; the importance of considering the effects of bias at both the reasoning stage of processing and the response stage. Typically, belief bias is measured using a number of indices. The logic index, which measures the degree of logical responding is calculated by deducting the number of invalid problems endorsed by participants as valid

from the number of valid problems endorsed. Belief based responses are measured using the belief index, calculated by deducting the number of unbelievable problems endorsed as valid from believable problems endorsed. Finally, the interaction index is designed to assess the degree of the belief bias effect, that is, the number of conflict problems endorsed minus the number of non-conflict problems endorsed. These indices have been criticised for conflating response bias and reasoning bias (Klauer et al., 2000; Dube et al., 2010; Trippas et al., 2013). Klauer and colleagues argue that whenever individuals are presented with a task in which a single response must be chosen from a number of given responses, it is crucial that response bias is accounted for, and so a model which can accommodate response bias is necessitated. Therefore, an alternative approach was used by Klauer et al, who constructed a multinomial processing tree (MPT) model, which separated reasoning biases from response biases. MPT models allow parameters to be mapped onto separate psychological processes in order to calculate the probability of a given response, and the particular process that led to that response. To illustrate this more clearly, Figure 2.1 shows the MPT model fitted by Klauer and colleagues.

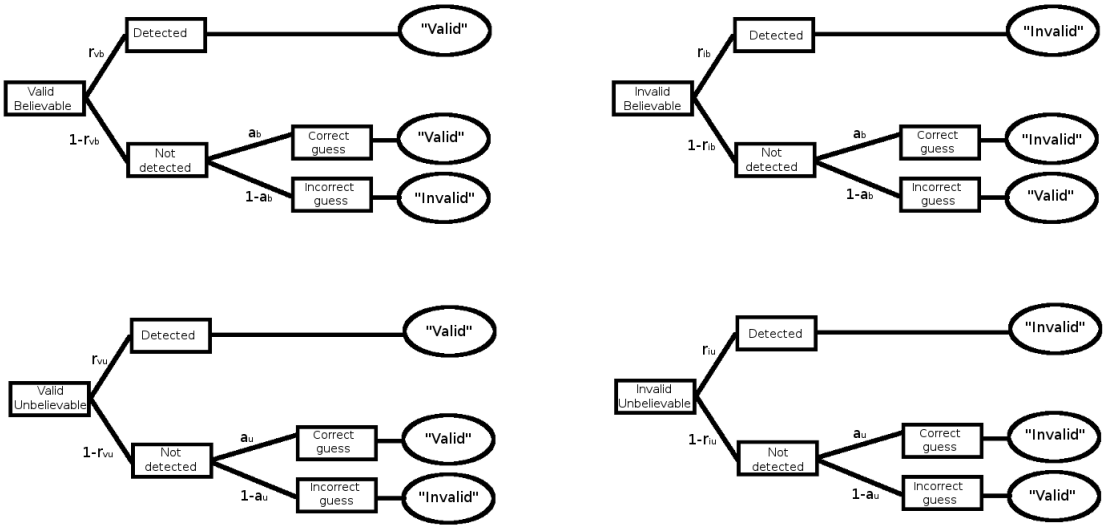


Figure 2.1: Diagrammatic representation of Klauer et al's MPT model of belief bias

Theoretical predictions about belief bias can be tested by examining which

parameter estimates are statistically different or equal; for example, in the figure, if r_{vu} is equal to r_{vb}, this indicates a lack of difference in the process for both types of valid problems. Using this methodology, Klauer et al. (2000) found evidence for belief bias affecting both response bias and playing a part earlier in the reasoning process. The idea of belief bias affecting the response stage is not novel; Oakhill & Johnson-Laird (1989) have suggested that belief bias is mainly a reasoning bias, although do allow for the possibility of "conclusion filtering", which takes place in the response stage. However, the MPT model of Klauer et al allows for specific statistical parameters to measure this response bias, and assess the two components of belief bias separately.

The MPT model was later criticised by Dube et al (2010), who argued that an SDT model is far more appropriate than an MPT model for the analysis of belief bias data. The mathematical basis of SDT models are discussed in greater depth in Chapter 3, but in brief, it characterises responses as lying somewhere on one of two continuous probability distributions (one representing valid problems, and the other representing invalid problems), lying on an X axis representing perceived argument strength. A visual representation of this can be found in Figure 2.2.

The distribution representing invalid problems, in most reasoners, lies to the left of the valid problem distribution, as such problems will have lower strength due to their lack of validity. A single cutoff point, the response criterion, shown by the bold black line, is indicative of response bias, and varies from person to person. If the argument strength is to the left of this threshold, a response of "invalid" will be given, whereas perceived argument strength greater than the criterion leads to the reasoner responding with "valid". As the probability of a certain response lies on a curve, changes in response bias will increment the number of hits (responses of "valid" to valid problems) and false alarms (responses of "valid" to invalid problems) by different amounts, and thus changes in response bias will still be erroneously conflated with changes in accuracy if this is not accounted for.

To further demonstrate the bases for this claim, Dube et al. (2010) employed

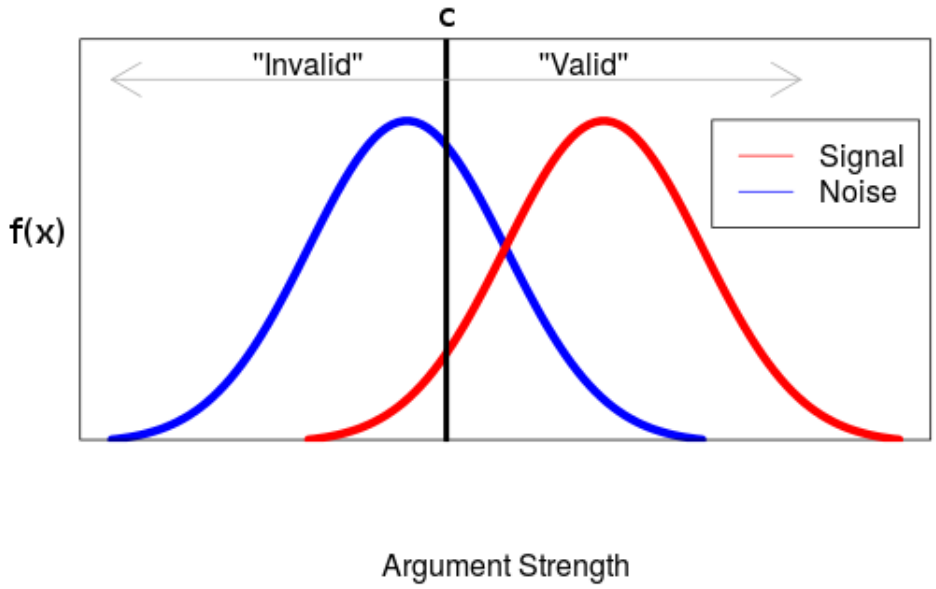


Figure 2.2: Plot showing SDT model of belief bias

a base-rate method, more commonly used in recognition memory research, in which participants in different conditions were either told that there was a very high or very low number of valid syllogisms in the materials that they would be presented with. In both conditions, despite instructions indicating otherwise, half of the syllogisms were valid and the other half invalid. Thus, this allowed Dube et al manipulate response bias whilst accuracy remains constant, as there is no theoretical reason that there should be any differences between the two conditions in being able to identify valid or invalid syllogisms as part of the reasoning process. When the resulting data was analysed using traditional approaches (i.e. logic, belief, and interaction indices), the base-rate manipulations appeared to affect accuracy, whereas the SDT model showed that in fact only response bias was altered. Plotting the predicted values of these statistical models against the actual data, and analysing the relevant goodness-of-fit statistics showed that the SDT model fit the data better than both the traditional models and the MPT model.

Further support for the increased accuracy of the SDT approach is provided

by Heit and Rotello (2014) who simulated data which was generated to reflect experimental manipulations in which participants' performance varied in terms of response bias but not reasoning accuracy. They found that the higher number of participants, and the higher number of trials per participant, the more likely it was that traditional analyses would erroneously indicated an accuracy difference between conditions, whereas SDT analyses clearly distinguished between the different components of belief bias. A graphical approach, the plotting of Receiver Operating Characteristic (ROC) curves were also used to explore this data. ROC curves plot changes in response bias whilst accuracy is held constant, and in the belief bias paradigm, are typically generated using confidence ratings. These are generated by asking participants to give a rating of how sure they are that they have given a correct response. The combination of binary response and confidence rating is then transformed into a single scale, with "totally sure, valid" at one end and "totally sure, invalid" at another end. The cumulative proportions are then plotted; an example can be seen in Figure 2.3.

More details can be found in Chapter 3, but the key point argued by Dube et al (2010) is that MPT and traditional approaches assume a linear ROC, whereas the SDT approach specifies a curved ROC, which reflects response bias affecting hits and false alarms in a non-linear fashion. When experimental data was plotted as ROCs, evidence overwhelmingly suggested that belief bias data produces curved ROCs and thus necessitated the use of SDT modelling (Dube et al., 2010; Dube, Rotello, & Heit, 2011; Trippas et al., 2013; Handley & Trippas, 2015). This may not be solely applicable to syllogistic reasoning; Heit & Rotello (2014) re-analysed data from earlier research (i.e. Rips, 2001; Markovits & Handley, 2005). Although these studies focussed on conditional reasoning rather than categorical syllogisms, a curved ROC was still found, showing the more general applicability of SDT analyses to investigating reasoning. In addition, the SDT approach has been used to criticise the way in which neurological data is used to support the theory that individuals implicitly detect a conflict between logic and belief. Rotello, Heit, and

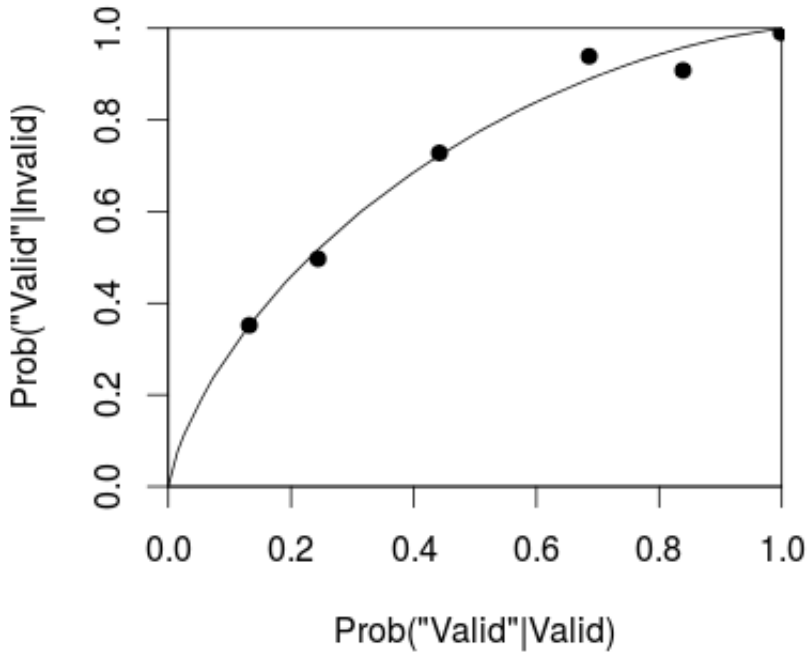


Figure 2.3: Example of a ROC Curve

Way (2014) argue that the use of traditional indices used to assess accuracy is deeply problematic, and many studies which claim support for these on the basis of activation of different regions of the brain whilst completing deductive reasoning problems are flawed. Whilst such studies argue that these patterns of activation give support to a conflict detection account, Rotello et al (2014) argue that this supposed link with conflict detection is, in fact, just activation of different regions because of differences in response bias.

The SDT approach has been criticised, however, on a number of grounds, by Klauer and Kellen (2011), their main arguments being that confidence rating data artificially produces curved ROCs, and that there were a number of methodological flaws in the procedure used by Dube et al (2010). Firstly, Klauer & Kellen (2011) claim that the use of confidence ratings ROCs artificially produce non-linear ROCs which would not be found with binary choice data. In response, Dube et al (2011)

argue that this claim is based upon poorly substantiated claims of Broder and Schutz, and conducted further experiments in which they found a curved ROC, despite using a binary choice method with no confidence ratings. Although Klauer & Kellen (2011) refit the SDT and MPT models to a large and varied dataset and found greater support for their MPT model, Dube et al (2011) argued that any improvement in model fit was due to the use of an overly narrow rating scale, and simulations on a wider scale provided greater support for the SDT model.

The application of the alternative modelling techniques led Dube et al (2010) to find that the apparent interaction between belief and logic disappears when an SDT approach is taken to analyse data. The lack of interaction effect renders many previous theories of belief bias incompatible with this SDT account. Therefore, Dube et al offer a number of possible explanations for how belief bias affects deductive reasoning, all based around a single process theory which suggests that perceived argument strength lies at the heart of the belief bias effect. Dube et al's *criterion shift* account suggests that believability affects the required level of strength of argument to determine whether a reasoner responds with "valid" or "invalid". Believable problems lead to the response criterion becoming more liberal, and thus, less evidence in favour of the particular conclusion is required for a reasoner to say that it is valid. Unbelievable problems, relative to believable problems, result in a more conservative response criterion. An alternative, the *distribution shift* account, posits that believability leads to a change in location of the argument strength distribution. Instead of two distributions representing valid and invalid arguments, there are four, with one for each problem type. The distance between the distributions for the two types of believable problems is equal to the distance between the distributions for the two types of unbelievable problems. However, Dube et al argue that the criterion shift account is more easily interpretable in line with their base-rate manipulation findings.

Although the SDT model suggested by Dube et al (2010) is a single-process theory and suggests that belief bias is purely response stage bias, more recent

research by Trippas et al (2013) manage to reintegrate SDT analyses with reasoning-based accounts of belief bias. The effects of time constraints, cognitive ability, and syllogism complexity were examined, given that dual-process accounts predict that all these factors should mediate the way in which believability affects deductive reasoning. Conversely, if belief bias is merely a response bias, as suggested by Dube et al, there should be no interactions between believability and any of these factors. Indeed, although Dube et al found that conclusion believability affected response bias but not accuracy, Trippas et al (2013) found that believability did affect accuracy unless participants were tested in circumstances under which sophisticated reasoning was unlikely to occur (i.e. with simple syllogisms, with a 10 second response time limit, and participants with a lower cognitive ability). Trippas and colleagues concluded that belief bias affects reasoning both at the processing stage as well as the response stage, although processing effects are only exhibited if complex processing can occur.

Further evidence for belief bias having components in both reasoning and response bias comes from Trippas, Verde, and Handley (2014), who found that when participants were presented with syllogisms side-by-side and asked to choose which was valid, when believability was manipulated between-participants, there was no effect of belief bias. The design naturally eliminated response bias by giving simultaneous presentations, and they argue that the lack of effect on reasoning stems from the fact that when all conclusions are either believable or unbelievable, participants can no longer rely on believability as a heuristic for responding. Later manipulations with believability as a within-participants factor did uncover reasoning accuracy differences due to belief bias.

The use of the SDT approach to investigate performance is neither novel to belief bias, nor mutually exclusive with previous theories of belief bias. Hayes, Heit, and Rotello (2014) argue that it is a common set of processes that are responsible for cognition in areas typically investigated entirely separately. They use the example of similarities in findings in memory, categorisation and reasoning research, and argue

that a more unified approach is important. A similar view is espoused by Khemlani & Johnson-Laird (2012), who discuss how a dual-process theoretical approach is important to tie together different ways of looking at syllogistic reasoning in order to acquire a more unified approach, combining ideas about both algorithmic ("how") and computational ("what") aspects of reasoning. Due to methodological and analytical differences, the reconciliation of theoretical models based upon SDT and non-SDT approaches may be a complicated thing to achieve.

However, Handley et al. (2011) suggest a modified elaboration of De Neys (2012) logical intuitions model, and their parallel competitive model suggests that conflict between logic and belief is bidirectional. Although much research investigates how beliefs can interfere with logical responding, there is also evidence to suggest that logic may interfere with belief-based responses (e.g. Handley et al., 2011; Pennycook, Trippas, Handley, & Thompson, 2014). In such cases, where participants are asked to provide a response based upon believability, the interference of logic on belief for conflict problems reduced the accuracy of these belief-based responses compared to non-conflict problems. Handley et al. (2011) argue that Type 1 processing is comprised both of processes related to structure, and of those related to knowledge or beliefs. Once Type 2 processing is engaged, any conflict between the two processes may be identified, after which point, one response is inhibited based upon the demands of the task. For simple problems, a structural Type 1 response is available rapidly, which accounts for the interference of logic on tasks in which a response based upon beliefs is required. However, for complex problems, a knowledge-based Type 1 response is more readily accessible, and so this must be inhibited in order to produce a logically normative response.

Although there is growing support for theories of belief bias which focus around conflict resolution, there is no clear consensus on which particular theory best describes the data. As discussed in Chapter 1, further insight into the deductive reasoning processes can be ascertained by examining how attempts to influence strategy change affects task performance. In the next section, I will consider both

general approaches to debiasing and ones specific to deductive reasoning and belief bias.

2.5 Debiasing

Cognitive biases are persistent, and reducing or eliminating them is a difficult task. Debiasing techniques that have been successful in one domain may not necessarily be effective when applied to other biases or tasks. The importance of the debiasing technique being specifically targeted at overriding the particular cause of bias has been important as being a key factor in the success of such an intervention (Slovic & Fischhoff, 1977). Given the lack of consensus regarding the specific mechanisms behind the source of belief bias, it is not immediately clear as to the approach which will garner the most success. Conversely, the discovery of a successful debiasing technique may lead to new insights into the belief bias effect. In addition, the majority of prior studies examining debiasing approaches have used statistical analysis techniques which risk conflating changes in reasoning bias and changes in response bias. This thesis will make a novel contribution to the field by examining debiasing interventions and analysing the resulting data using an SDT approach to adequately distinguish between the two types of belief bias.

A number of approaches to debiasing thinking errors are discussed by Arkes (1991). Arkes identifies various categories of errors, including association-based errors, such as belief bias. Association-based errors are thought to be caused by the influence of semantic associations in an individual's knowledge or memory interfering with more logical processes. In this section I will discuss general methods of debiasing thinking errors, along with their application to belief bias. Debiasing methods can be roughly divided into 3 categories: external factors which are independent from the given task, for example, financial incentives; prior instruction interventions, such as written descriptions of optimal strategies, which are given before the task; and online processing interventions, such as the provision of feedback, which attempt to modify how the individual engages with the task.

2.5.1 External Factors

Although providing individuals with incentives for better performance can be effective in overriding other types of bias, this strategy has been found to be ineffective for association-based errors. Arkes argues that individuals give high ratings of confidence in their own answers even when they are responding inaccurately, as they will proceed to create justifications for whatever answers they give. Even when given monetary incentives, confidence ratings remain high, despite incorrect answers (e.g. Fischhoff et al, 1977). This tendency has also been demonstrated in syllogistic reasoning research, such as Evans et al (1983), in which participants who were asked to explain their choice of answer out loud made reference to irrelevant information from the task in order to justify their erroneous responses. In some cases, for example, the base-rate task, incentives have been found to have a worsening effect on performance (Arkes, Dawes, & Christensen, 1986; Hogarth, Gibbs, McKenzie, & Marquis, 1991).

Repeated practice on tasks may lead to improvements in performance as participants become better acquainted with the demands of the task. However, where bias is present, it could also be argued that without any other intervention, the bias may become more deeply ingrained in the absence of corrective intervention. Johnson-Laird & Steedman (1978, Experiment 2) found that performance on a conclusion generation task taken two weeks apart led to improved performance. However, there was no belief bias manipulation in this experiment, and so we cannot make any inferences as to the effect of practice on reducing the reliance on this heuristic. Ball (2013) found that both feedback and practice led to improved performance, however, SDT analyses were not conducted, and it is unclear as to whether the response-bias or reasoning-bias components were affected by these manipulations.

2.5.2 Prior Instruction Interventions

Informing participants about the particular cognitive bias in question, and asking them to try to resist succumbing to it when making logical judgements is appealing because of its simplicity. However, this approach has been found to be ineffective (e.g. Fischhoff, 1975). This may be explained by the fact that association-based errors occur because of processes which are beyond the grasp of conscious awareness and these automatic behaviours cannot be easily altered (Arkes, 1991; Larrick, 2004). This argument has been developed further by Evans (1989) who argues that even when individuals are capable of engaging in effective analytic reasoning, such biases divert their attention to irrelevant aspects of the task in hand, and so providing an analytic solution to a heuristic-influenced problem will be unsuccessful. However, Macpherson and Stanovich (2007) found that when participants were both informed about a particular cognitive bias, and told to use the "consider the opposite" strategy, there was some increase in performance.

Nevertheless, descriptive materials which giving people in-depth instructions about logical necessity and the precise logical meaning of the different qualifiers in syllogisms has been claimed to reduce, although not completely eliminate, the belief bias effect (Evans et al., 1994; Newstead et al., 1992). This method of debiasing, however, required very detailed instructions, and simply explaining logical necessity but not further details led to no change in the levels of belief bias exhibited by participants (Evans et al, 1994, Experiment 2). These results have been disputed by Heit & Rotello (2014), who replicated this experiment, and found that fitting an SDT model to the data revealed that instructions had an effect on response bias, but not on reasoning bias. A more subtle variation on the previously mentioned specific instructions, which instead involves instructing individuals to perform a behaviour connected with debiasing has been found to be effective in some cases. Encouraging participants to imagine that their alternatives to their choice were true can reduce the effects of hindsight bias (Arkes, Faust, Guihnette, & Hart, 1988), can reduce overconfidence in the accuracy of responding (Koriat,

Lichtenstein, & Fischhoff, 1980; Hoch, 1985), and is more effective than simply instructing individuals to reason without bias (Lord, Lepper, & Preston, 1984). Arkes argues that this strategy is effective as association-based errors occur because a biased response is cued, and so cuing the alternative response as well leads to a lower tendency to simply choose the more easily accessible biased choice. Simply cuing debiasing behaviours is also effective; Tversky and Kahneman (1974) found that participants who were shown a story in which a character displayed certain decision making strategies were more likely use to the same technique themselves in a later task than participants who had not read the story. Explicitly telling participants to use a disconfirming strategy has been shown to be effective in debiasing performance on the Wason (1960) 2-4-6 task (Gorman & Gorman, 1984), but advising participants to test multiple hypotheses has little effect on belief bias (Evans et al, 1994).

2.5.3 Online Processing Interventions

Belief bias is a contextual bias; that is, it manifests predominantly not when a problem is considered alone, but in the context of other problems, some of which have believable conclusions, and others with unbelievable conclusions. When manipulated within-participants, the belief bias effect is often diminished (e.g. Evans & Pollard, 1990) or eliminated (Trippas, Verde, & Handley, 2014). This trend is also found when participants are asked to choose the valid syllogism from an unbelievable and a believable syllogism presented side by side (Trippas et al, 2014, Experiments 1 and 2).

Milkman, Chugh, and Bazerman (2009) discuss an approach to debiasing in decision-making research, which involves providing individuals with an alternative heuristic to the one which is the source of their bias, with the replacement being one which results in more accurate responding. However, given the strength of the belief bias effect, this seems unlikely to prove useful. Alternatively, as belief bias may be caused by an over-reliance on Type 1 responses, providing reasoners with a

strategy which explicitly engages a Type 2 response may produce improvements in performance. This approach has been found in other areas of cognition to lead to novice task performance becoming more accurate than that of experts (Dawes, 1971). The use of simultaneous presentation to encourage a shift from Type 1 to Type 2 responding discussed above has also been found in other domains (e.g. Bazerman, Loewenstein, & White, 1992). However, in everyday life, it is rare that logical arguments are encountered in such a specific context, and so alternative ways of altering processing are still important to examine.

The use of disfluent texts has been argued by some to be effective in this regard; Thompson, Prowse-Turner, et al. (2013) found that participants of higher cognitive ability were more accurate on the Cognitive Reflection Task (CRT; Frederick, 2005) when the text of the task was presented in a font which was difficult to read. They concluded that this was because it promoted extra processing. However, Trippas, Handley, and Verde (2014) found that, in the case of syllogistic reasoning, disfluency actually led to lower accuracy for higher ability participants, concluding that on complex tasks, disfluent fonts leads to additional processing which would otherwise be used to deduce a correct answer. Meyer et al. (2015) reported the results of research in which 16 separate experiments attempted to replicate the effect of disfluent fonts on improving performance. They found no effects of disfluency at all, even when cognitive ability was accounted for, and concluded the only consistent change in performance was an increase in response time.

Feedback has been used in other domains such as memory and decision making to improve task performance. When rapid feedback is available on a large number of decisions for which ratings of confidence have been given, individuals show a stronger link between accuracy and confidence in that domain (Wagenaar & Keren, 1986). Feedback and confidence ratings have also been used as a debiasing technique in a classroom setting, with participants only showing an increase in accuracy after being given both immediate feedback and having to give confidence ratings (Renner & Renner, 2001). Indeed, it has been argued that response confidence acts

as a metacognitive cue for reasoners in determining whether they give an answer consistent with Type 1 responding or Type 2 responding (Thompson, Prowse-Turner, et al., 2013). Therefore, asking reasoners to give a rating of confidence in a given answer, followed by evaluative feedback may lead to better calibrated confidence and higher accuracy.

Specific training can be an effective debiasing technique, as rather than attempting to modify behaviours, it simply gives people "those tools needed to arrive at correct answers" (Arkes, 1991). The benefits of undertaking a course in statistics, or training in statistical principles in the lab, have been shown to be transferable to a later task involving statistical inference (Fong, Krantz, & Nisbett, 1986). As discussed earlier, inhibition may play an important role in the suppression of intuitive response on conflict problems. Houdé et al. (2000) attempted to debias participants completing a matching bias task. Participants were given a pre- and post-test, with an inhibition training task in-between. It was found that participants showed lower levels of matching bias on the post-test, and increased activation in areas of the brain thought to be associated with inhibitory control. One problem with taking this as evidence of the success of inhibition training, however, is that Houd et al did not include a control condition in their experiment. Moutier, Angeard, and Houdé (2002) compared the effects of inhibition training and general logic training on performance on the matching bias task. It was found that inhibition training improved performance, but logic training had no effect.

Later research by Moutier and Houdé (2003) examined debiasing of the conjunction fallacy. The conjunction fallacy is often demonstrated by participants' response to the Linda problem (Tversky & Kahneman, 1983):

> "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.
>
> Which is more probable?

Linda is a bank teller.

Linda is a bank teller and is active in the feminist movement."

When given such a task, individuals tend to choose the second option, fitting in with the extra information that they have been given. However, logically, the first option is more likely given that membership of a single category cannot be probabilistically more likely than membership of that category in addition to another one. Moutier & Houde (2003) found that inhibition training was the key to enhanced performance on this task. When participants were given training which shared many underlying features with the main task, little improvement in performance was shown compared to a control group who were given no such training, even when the underlying logic of the task was explicitly discussed, and it was checked that the participants understood it fully. However, an additional condition combined this logic training with an inhibition component, in which the experimenter explained why the participant's intuitive response may be misleading, and had the participant repeat the explanation back to them until it was clear that they had a complete understanding of it. It was found that in this condition participants showed significantly less bias on a post-test than the control or logic-trained group, indicating that the demonstration of principles of inhibition were key to debiasing responses on this task.

Logic training alone is thought to fail, as people struggle to apply abstract the concepts they have been trained in to the specific tasks they are required to complete. Specific task training alone is also ineffective; to get any benefit, individuals must be simultaneously trained in both the specific task they will be tested on, and the underlying logical principles (Bransford, Sherwood, Vye, & Rieser, 1986; Cheng, Holyoak, Nisbett, & Oliver, 1986).

Task-specific training has been shown to improve performance in syllogistic reasoning tasks; Prowse-Turner and Thompson (2009) found that a short training session in which participants were taught how to construct diagrams to represent syllogisms, and then training on just 5 practice syllogisms subsequently improved

performance on a later task in which 16 syllogisms were evaluated. A more detailed approach was taken compared to the typical binary response design, which showed participants became better at categorising syllogisms on the basis of logical necessity. However, this study did not use belief-oriented materials, and it is unclear whether this technique would be sufficient to override an effect of such magnitude. In addition, Prowse-Turner & Thompson added further nuance to the issue of confidence rating data; item-by-item confidence was no better calibrated with training. However, a single overall estimate of performance was found to be significantly more accurate for the group who did receive training.

2.6 Conclusion

Debiasing reasoning is a complex task and little research has been done on whether the belief bias effect can be reduced or eliminated. Evidence from related areas, such as decision-making, affords us insight into methods which have worked here. There is some evidence that the belief bias effect can be attenuated (Ball, 2013; Macpherson & Stanovich, 2007; Newstead et al., 1992); however, the recent application of SDT methods to analysing belief bias data has highlighted the importance of distinguishing between reasoning and response bias when evaluating such methods (Heit & Rotello, 2014). In the next chapter, I go on to discuss the methodological bases of the analyses presented in this thesis, and then in Chapter 4 report the results from an experiment in which participants were given feedback to their responses on a syllogistic evaluation task. Chapter 5 examines the use of differing sets of instructions and aims to investigate whether individual differences have a part to play in the efficacy of debiasing methods. A final experiment reported in Chapter 6 continues the theme of individual differences, and examines whether such differences affect the ability of individuals to improve performance and reduce bias via the combination of training and feedback.

Chapter 3

Methodology

This thesis contains the results and analyses of three experimental studies. This chapter will discuss the statistical methodology used to analyse the results of these experiments. The analyses for all three experiments involves generating accuracy and response bias score using the SDT model. The effects of various factors on these scores will be analysed using mixed effects models. The same procedure will also be applied to analyse the traditional endorsement rates measures, the results of which will be compared to the SDT models. In this chapter, I will briefly discuss traditional approaches to analysing belief bias, and then move on to examining SDT, beginning by outlining the theoretical underpinnings of this model. I have already discussed the arguments for and against the use of SDT in the previous chapter; here, I will present a more in-depth discussion of the conceptual and mathematical principles of SDT. Following this, I will go on to discuss the use of mixed-effects models, and the criteria which can be used to select the fixed and random effects for the various models. Due to differences between each of the three experiments, discussion of methodological considerations such as sampling design and specific experimental design will be located in the methodology section of those specific chapters.

3.1 Traditional Approaches

Earlier approaches to analysing belief bias data involved aggregating multiple responses from a single participant in order to generate a number of indices which were then analysed as dependent variables. Typically, three different indices are calculated: the logic index, belief index, and interaction index. They are calculated thus:

$$LI = \Big((\text{“}Valid\text{”}|VB) + (\text{“}Valid\text{”}|VU)\Big) - \Big((\text{“}Valid\text{”}|IB) + (\text{“}Valid\text{”}|IU)\Big)$$
$$BI = \Big((\text{“}Valid\text{”}|VB) + (\text{“}Valid\text{”}|IB)\Big) - \Big((\text{“}Valid\text{”}|VU) + (\text{“}Valid\text{”}|IU)\Big)$$
$$II = \Big((\text{“}Valid\text{”}|VU) + (\text{“}Valid\text{”}|IB)\Big) - \Big((\text{“}Valid\text{”}|VB) + (\text{“}Valid\text{”}|IU)\Big)$$

The logic index is meant to be a measure of how much the participants' responses correspond with normative logic; belief index a measure of the influence of belief on their answers, and interaction index a measure of how much this effect of belief differs for valid and invalid conclusions. These indices are then typically analysed using analysis of variance (ANOVA). For reasons discussed in Chapter 2, the use of these indices has fallen out of favour, in many cases replaced by an SDT approach. I will now go on to discuss the origins and mathematical principles behind SDT analysis.

3.2 Signal Detection Theory

Recently, there has been a growing trend to investigate belief bias using a signal detection paradigm. It was originally a mathematical concept used to describe engineering problems, but has also been applied to topics in cognitive psychology. SDT has its roots in information theory, a concept developed by Shannon (1948), and is concerned with how information can be transmitted over a noisy channel. The term 'noise' refers to randomness and extraneous influences which can affect the interpretation of the signal.

3.2.1 Response Types

SDT differs from traditional approaches to examining belief bias in that it gives a more general description of the reasoning process and task responses, and puts less focus on the underlying mental architecture. It can be employed in binary choice tasks where participants make binary choice responses in regards to the presence or absence of a target stimulus. Participants' responses are divided into four categories; hits (responding "valid" to valid problems), false alarms ("valid"|invalid), correct rejections ("invalid"|invalid) and misses ("invalid"|valid). In the belief-bias paradigm, trials in which invalid problems are presented would be classified as *noise* trials, and trials in which valid problems are presented are *signal-plus-noise* trials, or simply *signal* trials. This is summarised in Table 3.1.

	"Valid"	"Invalid"
Signal	Hit	Miss
Noise	False alarm	Correct rejection

Table 3.1: SDT response classifications

3.2.2 SDT conceptualisation as distributions

Another way of representing an individual's responses is as two continuous probability distributions on a single axis, as shown in Figure 3.1. The x axis can be thought of as indicating perceived signal strength and the y axis is the probability of the response. In the traditional example of tone perception, the x axis would represent loudness, and in belief bias research it represents argument strength (Dube, Rotello & Heit, 2010). One of the two distributions represents the noise distribution, and the other represents the signal-plus-noise distribution. An individual's sensitivity (i.e. ability to discriminate between signal-plus-noise and just noise; accuracy) is represented by the distance between the centre points of the two distributions. It is quantified using the statistic d'.

The probability of an individual classifying a signal of a particular strength as either noise or signal-plus-noise depends on the height of the relevant distribution

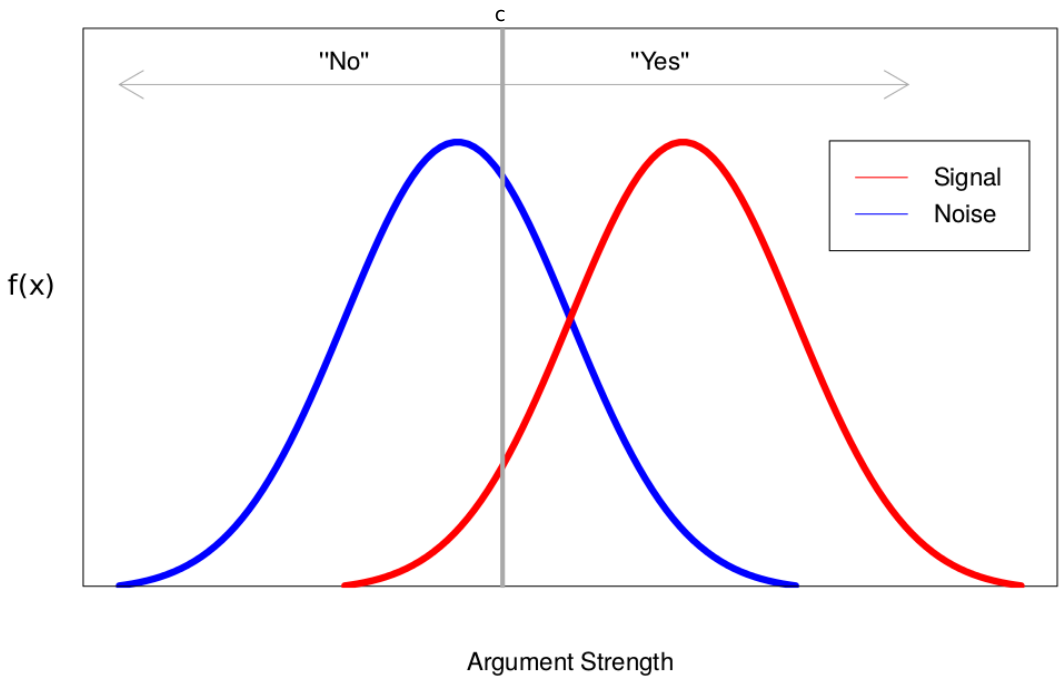


Figure 3.1: Plot showing example signal and noise distributions

at that particular point on the x axis. The greater the overlap between the two distributions, the greater the uncertainty of whether the signal has been perceived or not. This uncertainty is resolved by the use of a response criterion. This is a location on the x axis at which any stimuli that fall to the left of this point (i.e. have a weaker perceived signal strength), will be reported by the individual as being just noise, whereas to the right, individuals will report that they have detected the signal. In Figure 1, the grey vertical line demarks the response criterion.

The number of correct rejections made by the individual is determined by the total area under the noise distribution that falls to the left of the response criterion point, and false alarms by the area under the same distribution that falls to the right of the response criterion. Similarly, the equivalent applies to misses and hits depending on the location of the response criterion relative to the signal-plus-noise distribution. This is shown in Figure 3.2, which shows data from an individual, but with different attributes highlighted in each image. The first image shows the valid

distribution. For all problems that are actually valid, if the argument strength falls to the right of the individual's response criterion, they give a response of "valid", a hit, as highlighted in green; for arguments with a strength below (i.e. to the left) of that criterion, the individual will response with "invalid", a miss, highlighted in grey. The second image shows the "invalid" distribution, and here, arguments with strength lower than the criterion results in a correct rejection, highlighted in blue; those with a higher strength results in a false alarm, highlighted in orange.

Response criterion

A liberal response criterion would lead to high levels of hits and false alarms, and low levels of correct rejections and misses, reflecting a tendency to report that a signal has been detected even if the individual is uncertain. Conversely, a conservative response criterion would lead to more correct rejections and misses, but less hits and false alarms, due to an answer of "no" being unlikely in the case of uncertainty. The response criterion in Figure 3.1 is a fairly liberal response criterion, with high levels of hits, moderate levels of false alarms, low levels of misses, and moderate levels of correct rejections. An individual's response criterion is independent of their accuracy, as demonstrated in research that has included manipulations such as incentivising the use of more liberal or more conservative response criteria. In other words, their proficiency at the task remains constant whilst variation in response criterion being reflected in altered rates of hits, misses, false alarms and correct rejections.

3.2.3 Comparison to traditional approach

Traditional analyses of belief bias experiments calculate accuracy by simply deducting the number of false alarms from the number of hits (also known as a H-F index). As such analyses do not separate response criteria from accuracy, they may lead to erroneous conclusions that individuals with a more extreme response criterion are apparently showing poorer performance on a given task.

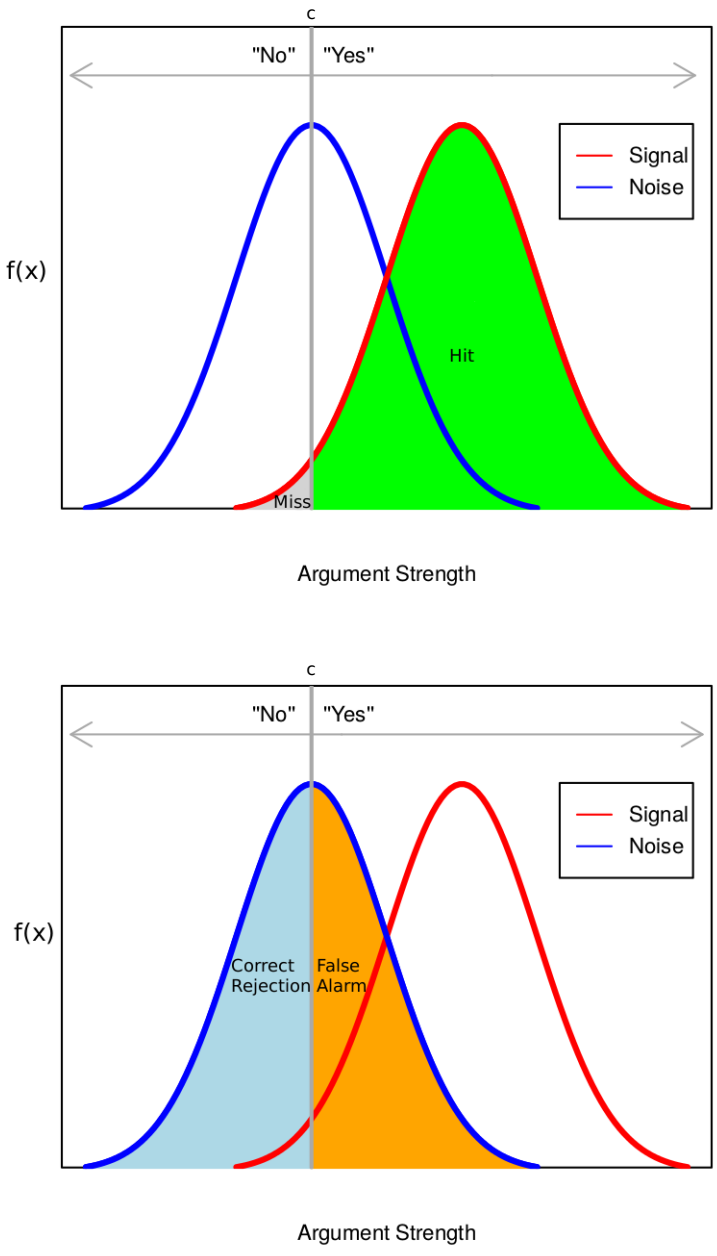


Figure 3.2: Plots demonstrating a) hits vs. misses and b) correct rejections vs. false alarms

This is shown in more detail in Figures 3.3, 3.4, and 3.5. These figures show responses from 3 different participants who have identical accuracy, as shown in the equal distance between the valid and invalid distributions, but vastly differing response criteria, depicted by the placement of the criterion line. Figure 3.3

shows the responses from Participant A, who has a very liberal response criteria, and so requires very little evidence to give a response of "valid". Their hit rate is approximately 99%, but false alarm rate is around 70%. Figure 3.4 shows Participant B, who requires more evidence to deem a conclusion as valid, and so their criterion is closer to the centre, and they have a hit rate of around 96% and a false alarm rate of around 50%. Participant C, as shown in Figure 3.5, requires a lot of evidence to say that a conclusion is valid, and so has a conservative response criterion. Their hit rate is around 40%, but false alarm rate is only 5%. Traditional analyses would show these participants varying significantly in terms of accuracy, with accuracy scores of 29%, 46% and 35% respectively. However, as can be seen in the SDT diagrams, their accuracy is the same, but they do differ in the levels of response bias shown.

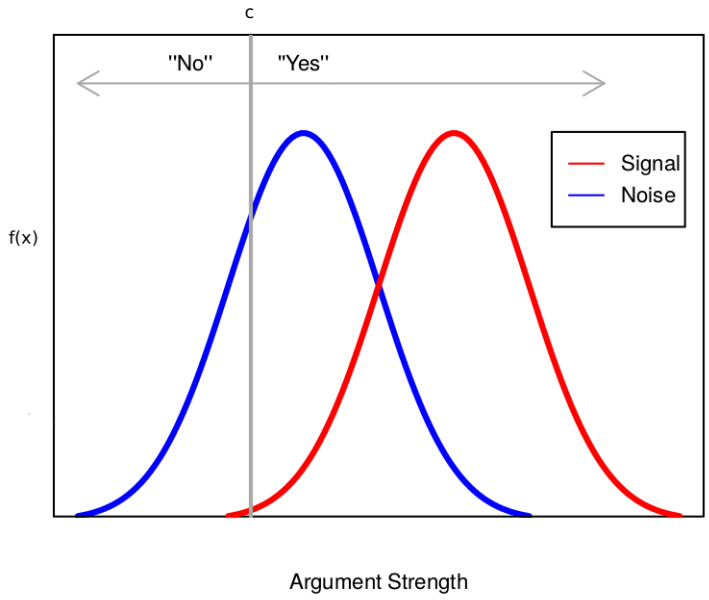


Figure 3.3: Plot showing example signal and noise distributions where the response criterion sets hits to 99% and false alarms to 70%

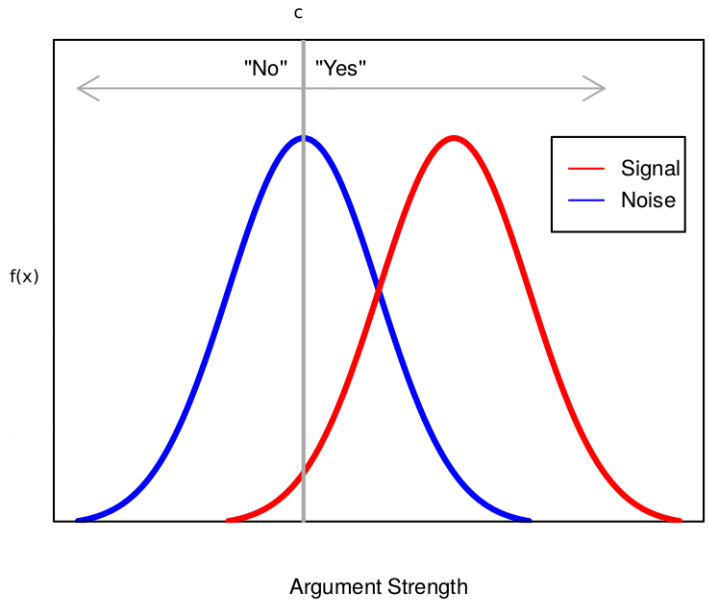


Figure 3.4: Plot showing example signal and noise distributions where the response criterion sets hits to 96% and false alarms to 50%

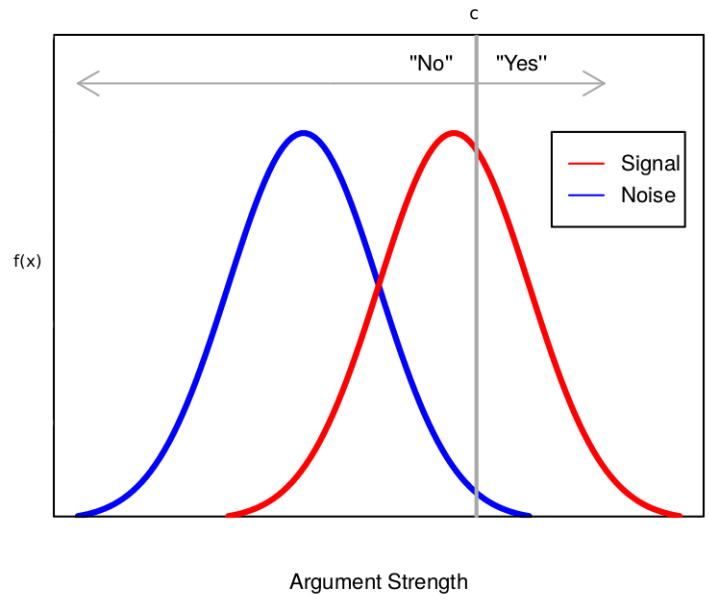


Figure 3.5: Plot showing example signal and noise distributions where the response criterion sets hits to 40% and false alarms to 5%

3.2.4 ROC Curves

An SDT approach therefore allows deeper insight into task performance. Graphical approaches, such as ROC (Receiver Operating Characteristic) curves allow us to

plot an individual's accuracy whilst taking into account their response criterion; and example can be seen in Figure 3.6. ROC curves are constructed by plotting the proportion of false alarms on the x axis and proportion of hits on the y axis. If multiple individuals' scores fit on the same curve at different points, as shown in Figure 3.6, then they have the same accuracy but different response criteria. The more bowed towards the top left the ROC curve, the higher the sensitivity. Chance-level performance results in a ROC curve which falls along the central diagonal. In the next section, I will explain the calculation of SDT measures, followed by a discussion of how these can be extended to data in which participants provide ratings of subjective confidence.

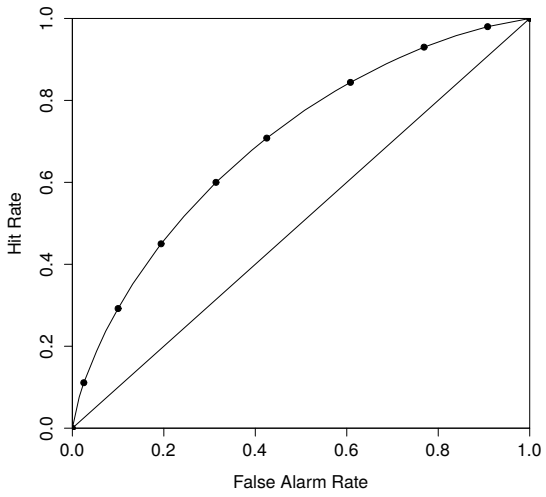


Figure 3.6: Plot of example ROC curve

SDT Statistics

Sensitivity, or accuracy, is measured using d' (d prime), the discrimination index and is calculated thus:

$$d' = z(H) - z(F) \tag{3.1}$$

That is, we work out the (one-tailed) z value that corresponds to the proportion of false alarms, and deduct it from the z value for the proportion of hits. If we plot $z(H)$ against $z(F)$, we get what is known as a zROC.

Figure 3.7 shows a ROC curve and its equivalent zROC, that is, the values from the ROC curve transformed into z scores and plotted on a graph. A line can then be fitted to these value to examine a number of things about the model. If the points fit the straight line well, it shows that the model is a good fit. If the line is at a 45 degree angle, this would indicate the suitability of an equal variance model, otherwise an unequal variance model is necessary.

The slope of a zROC, s, represents the ratios of standard distributions of the signal-plus-noise and noise distributions. One of the assumptions of the d' statistic is that the variances of the two distributions are equal, and so values of s other than 1 imply that this assumption has been violated. Dube, Rotello & Heit (2010) argue that s is often found to be values other than 1 in reasoning experiments, and so an unequal-variance adjusted version of d' should be used.

The unequal-variance SDT model takes into account the differences between the standard deviations of the two distributions, and d_a is used in this case:

$$d_a = \sqrt{\frac{2}{1+s^2}} d'_2 \tag{3.2}$$

where s is the SD of the signal-plus-noise distribution (the SD of the noise distribution is fixed at 1) and d_2' is the sensitivity of the signal-plus-noise distribution.

An alternative to d_a is A_z, which is an estimate of the area under the ROC curve.

$$A_z = \phi\left(\frac{d_a}{\sqrt{2}}\right)) \tag{3.3}$$

One advantage of using A_z is that its scale ranges from 0 to 1, with 1 representing complete accuracy and 0.5 representing chance-level performance.

The response criterion, denoted simply as c, measures levels of response bias. Negative values of c represent a liberal response bias, whereas positive values of c represent a conservative response bias. c is calculated thus:

$$c = -0.5[z(H) + z(F)] \tag{3.4}$$

In the case of unequal-variance SDT models, as with the sensitivity measure, the criterion measure is also adjusted.

$$c_a = \frac{-\sqrt{2s}}{\sqrt{1+s^2}(1+s)}[z(H) + z(F)] \tag{3.5}$$

3.2.5 Empty cell adjustments

When calculating the above measures, an adjustment must be made for proportions of hits or false alarms of 0 or 1, as the z scores of these would be $-\infty$ and ∞ respectively, making the rest of the solution mathematically intractable. A common solution to this problem is to add or deduct a frequency of 1 observation or a smaller fraction of an observation to values of 0 or 1. Wickens (2002) highlights the fact that these possible adjustment strategies are all equally valid. Following the recommendation of Snodgrass & Corwin (1988) that whatever correction is made, it should be applied to all cells and not just those containing values of 0 and 1, the following adjustment will be made:

$$H(adj) = \frac{((H * N) + 0.5)}{(N + 1)} \tag{3.6}$$

$$F(adj) = \frac{((F * N) + 0.5)}{(N + 1)} \tag{3.7}$$

where H and F represent the proportion of hits out of total signal stimuli and proportion of false alarms out of total noise stimuli respectively, and N is the number of items. The R code used to calculate the equations discussed above applied to the research presented in this thesis can be found in Appendix B.

Wickens (2002) highlights the fact that analyses on aggregate sensitivity and criterion placement scores will yield different results to analyses which instead calculate these scores for each individual participant. It is argued that both

approaches are legitimate, and therefore the aims of the research should inform the decision as to which approach is used. However, Klauer & Kellen (2011) argue that aggregate SDT models hide differences between individuals, rendering results inaccurate. This view is supported by Trippas et al (2015), who argue that when individual SDT models are compared to aggregate models, the individual models tend to provide a better fit to the data. Trippas and colleagues conclude that this is due to individual differences in reasoning strategies leading to analyses of aggregate data being inappropriate, due to the extra variance introduced because of these differences. This recommendation is reiterated by Cohen et al (2008), who argue that aggregate analyses are useful when there are few data points per participant, but individual analyses are preferential when this is not the case.

3.2.6 Confidence Rating ROCs

One of the central presumptions of SDT is that the ROC is curved. A curved ROC indicates that the SDT approach is necessary, whereas a linear ROC would indicate that an alternative approach, such as an MPT model may be more appropriate. One way to test that a ROC is curved is to find a way of altering response bias, whilst keeping accuracy constant. This has been demonstrated by Dube, Rotello & Heit, 2010, who manipulated across conditions the expectations of participants of the proportion of trials that will be signal-plus-noise trials, whilst leaving these rates actually unchanged, and found evidence for curved ROCs, thus supporting the use of the SDT model.

An alternative approach to this is to plot confidence-rating ROC curves. These are generated by instructing participants to make both a binary response regarding signal presence or absence, along with an expression of how confident they are in their response. Then, the detection rate for each individual level of confidence is calculated. For example, a binary choice experiment may require participants to give a response of “yes” or “no”, along with a rating from 1-3 to indicate how confident they are in their response, with a rating a 1 meaning “completely sure”

and a rating of 3 indicating that their response is merely a guess. These responses are then transformed to a scale from 1 to 6, with a rating of 1 corresponding a response of "yes, completely sure" and a rating of 6 corresponding to "no, completely sure". The z score for each rating is then calculated by working out the z score for the cumulative proportion of ratings. So for a rating of '1', the proportion is the number of '1' rating score out of the total rating scores; for a rating of '2', the proportion is the number of '1' and '2' ratings out of the total, and so on. A line is then fit between the z scores for the signal and noise distributions, and individual sensitivity and response bias statistic shown in Equation 3.3 and Equation 3.5 can be calculated, with the intercept and coefficient of the line corresponding to s and d'_2.

These calculated values correspond to the multiple criterion for the confidence ratings, as can be seen in Figure 3.8. Here, there are 6 possible responses, and so there are five different response criterion. This is expressed mathematically as:

$$
\begin{aligned}
P(\text{“1”}|Valid) &= \phi\left(\frac{\mu_v - c_1}{\sigma_v}\right) \\
P(\text{“2”}|Valid) &= \phi\left(\frac{\mu_v - c_2}{\sigma_v}\right) - \phi\left(\frac{\mu_v - c_1}{\sigma_v}\right) \\
P(\text{“3”}|Valid) &= \phi\left(\frac{\mu_v - c_3}{\sigma_v}\right) - \phi\left(\frac{\mu_v - c_2}{\sigma_v}\right) \\
P(\text{“4”}|Valid) &= \phi\left(\frac{\mu_v - c_4}{\sigma_v}\right) - \phi\left(\frac{\mu_v - c_3}{\sigma_v}\right) \\
P(\text{“5”}|Valid) &= \phi\left(\frac{\mu_v - c_5}{\sigma_v}\right) - \phi\left(\frac{\mu_v - c_4}{\sigma_v}\right) \\
P(\text{“6”}|Valid) &= \phi\left(\frac{c_5 - \mu_v}{\sigma_v}\right)
\end{aligned} \tag{3.8}
$$

$$
\begin{aligned}
P(\text{“1”}|Invalid) &= \phi\left(-c_1\right) \\
P(\text{“2”}|Invalid) &= \phi\left(-c_2\right) - \phi\left(-c_1\right) \\
P(\text{“3”}|Invalid) &= \phi\left(-c_3\right) - \phi\left(-c_2\right) \\
P(\text{“4”}|Invalid) &= \phi\left(-c_4\right) - \phi\left(-c_3\right) \\
P(\text{“5”}|Invalid) &= \phi\left(-c_5\right) - \phi\left(-c_4\right) \\
P(\text{“6”}|Invalid) &= \phi\left(c_5\right)
\end{aligned}
$$

Note that the equations are actually identical for the invalid and valid distributions, but as the calculation involved setting μ_i to 0 and σ_i to 1, these terms become redundant.

One possible problem with using confidence rating ROCs arises when participants do not use the full range of the rating scale, rating their confidence as equal on all or most trials. In this case, even with the empty cell adjustment discussed above, the confidence rating ROC will be a poor fit, and not accurately reflect the data. Here, the model should be treated as an equal variance model, and d prime calculated.

A further criticism of ROC and SDT methods is that, due to the need for a

number of confidence ratings across multiple participants, they require relatively large amount of data.

3.3 Mixed Effects Models

The analyses presented in this thesis utilise mixed effects regression models. Mixed effects model are also known as hierarchical regression models, multilevel models, or linear mixed models, and are an extension of fixed effects models. They allow for the inclusion of random effects and can be preferable to fixed effect models as they account for variability caused by taking multiple measurements from the same participants. In the case of the experiments presented in this thesis, they allow to examine the effects of numerous interventions on reasoning, whilst, to some extent, taking into account individual differences.

Mixed effects models can be specified as Level 1 and Level 2 models to account for the grouping nature of the data.

The Level 1 model for a random intercept model can be specified as:

$Y_{ij} = \beta_{0j} + \beta_{1j}x_{1ij} + ... + \beta_{nj}x_{nij} + \epsilon_{ij}$

Where Y_{ij} is response i for individual j, β_{0j} is the individual-specific intercept and explained further in the level 2 model. The slope coefficients $\beta_{nj}x_{nij}$ are the parameters for the other covariates. This model differs from fixed-effects models, as each individual has their own error term, ϵ_{ij}.

The level two model can be expressed as:

$\beta_{0j} = \gamma 00 + U_{0j}$

$\beta_{1j} = \gamma 10$

The parameter β_{0j} is the same as is in the level 1 model. It is compromised of $\gamma 00$, the grand mean, also known as the level 2 intercept, which is the mean for all individual data points. The other term, U_{0j}, is the error term for level 2, effectively, the individual deviation from the grand mean. The term $\gamma 10$ is the value for covariate β_{1j} and in the case of a random intercept model has no random

component itself.

The above formulation only includes random intercepts; however, if we wished to include random slopes, for example, on the covariate β_{1j}, the level 1 model would remain the same, but level 2 model would be expressed as:

$\beta_{0j} = \gamma 00 + U_{0j}$

$\beta_{1j} = \gamma 10 + U_{1j}$

where U_{1j} is individual deviation from fixed slope effect $\gamma 10$.

One advantage of fitting a mixed effects model to the data is that it allows us to fit both overall and individual intercepts and slopes to the data. Fixed-effects can be used to assess the effects of factors such as different experimental conditions, and are interpretable in the same way as they are in general linear models, generalised linear models, and ANOVA models. Random effects allow us to account for other variance and error due to differences between individuals. The inclusion of random intercepts in a model allow us to assess the impact of, for example, an experimental condition, but takes into account individual variance when assessing how much variance in the data that factor accounts for.

The inclusion of random slopes allow us to assess whether the effects of covariates on dependent variables differ between individuals. For example, in Experiment 1, we can examine whether the changes over time are the same for all participants or differ from person to person. This is done by including a random slope for the variable 'session'.

Random slopes and intercepts may be correlated or uncorrelated. If uncorrelated, there is variance between scores and variance between susceptibility to interventions, but no link between the two. However, if these random effects are correlated, it could mean, for example, that participants with a higher initial score show more improvement over time.

Another advantage of mixed effects modelling over the traditional approach of repeated measures analysis of variances (RM-ANOVA) is that it allows a greater variety of outcome variables. Whilst RM-ANOVA only permits continuous variables

as dependent variables, mixed effects modelling allows discrete, categorical and ordinal data, amongst others. In addition, mixed-effect models allow continuous variables as predictors, which RM-ANOVA does not. Although such variables can be transformed into categorical variables (e.g. 'high ability' vs. 'low ability' grouping), this leads to a potential loss of information with fixed effects which would have been significant if entered as continuous often showing up as non-significant when transformed to categorical variables (Cohen, 1983).

As data is analysed in the long format rather than wide format, mixed effects modelling can accommodate missing data, in cases where participants drop out part way through the study.

As mentioned earlier, traditional analyses of belief bias data involves calculating logic, belief, and interaction indices for each participant. These aggregations transform the binary data into a format which can then be used in an ANOVA. However, this aggregation leads to some loss of information, and as mixed effects models can accommodate binary responses, we are able to fit a model to the individual responses. In addition, we are now able to better explore the effects of materials, i.e. the specific syllogisms and their content.

3.3.1 Assumptions

Mixed effects model operate under a number of assumptions, which I will discuss below. One assumption of mixed effects models is that of linearity - the relationship between variables is linear in nature. Furthermore, it is assumed that the error terms are normally distributed and also that there is homogeneity of variance - population variances are equal. In addition, there is an assumptions of independence - scores are independent of one another. In this case, multiple scores are taken from the same participant, however, this is accounted for in the models, and would only be problematic if the score from one participant was dependent on the score from another participant. Graphical and statistical methods will be used to ensure that these assumptions have not been violated.

3.3.2 Multicollinearity

One potential problem when modelling large numbers of variables is that of multicollinearity; when independent variables are correlated with one another, typically with r>0.8. To account for this, the correlations between the individual differences measures will be calculated and reported, and all such variables will be centred around zero by deducting the mean from each individual score. This reduces issues caused by multicollinearity, and also makes interactions between variables simpler to interpret.

3.3.3 Model Comparison

The mixed effects models presented in this thesis were fitted using the lme4 package (Bates, Maechler, & Bolker, 2011) in the R Environment for Statistical Computing (R Development Core Team, 2009). The p-values for the fixed effects were calculated using likelihood-ratio tests, which compare the deviance of the model with and without the effect being tested against a chi-square distribution with K df, where K is the number of extra parameters in the more complex model being proposed. The comparison of fixed effects was done on models which specified in the syntax that REML=FALSE. This is based upon a recommendation by Pinheiro & Bates (2000) for models which vary in terms of fixed effects but not random effects. The default setting of REML=TRUE modifies the calculations by making restricted maximum likelihood calculations, and is only applicable for cases of model comparison where models vary in terms of random effects, the variance of which needs to be accounted for.

Although the selection of fixed effects can be done using the LRT, this test has been argued to be inappropriate for the selection of random effects, as it has been shown to be biased in some cases, and produces p values which are too conservative, leading to the rejection of random effects which should be included in the model (Stram & Lee, 1994). An alternative method of selection random effects is to use a bootstrap method.

This method calculates restricted likelihood-ratio tests, and calculates the p values for the random effects by simulating the data for both the reduced and full model in order to calculate a p value derived from the deviances of both the models. This method, termed the 'slow bootstrap' can be computationally intensive and time-consuming, and a quicker version, the 'fast bootstrap' designed by Crainiceanu & Ruppert (2004) has been implemeted in R software under the RLRsim package (Scheipl, Greven & Kuckenhoff, 2008). However, this version of the bootstrap does not allow for correlated random effects, and so these will be tested for using the slow bootstrap. The code used for the slow bootstrap is adapted from code found in Long (2012) and can be found in Appendix B.

All models presented in this thesis use the following procedure to select the best fitting model. A step-up approach is used, which begins with a null model and then iteratively adds main effects and their interactions, with only the statistically significant terms being retained in the model. A theory-driven approach is taken in specifying interaction effects in order to avoid spurious results that may occur if a data-driven, often termed 'data dredging' approach is taken. As recommended by Long (2012), the selection of fixed effects and random effects are conducted separately. This is of most importance in Experiment 1 (Chapter 4), in which data is taken not only from multiple participants, but also on multiple occasions from the same participants. In this case, the recommendations of Wallace & Green (2002) were followed - the fixed effects were first evaluated, using models which contained a random intercept, but no random slope. Once the final fixed effects model had been chosen, the inclusion of random slopes was tested for significance. If this was statistically significant, the fixed effects were tested once again to assess whether the inclusion of random slopes had an impact upon their significance, and fixed effects terms were removed if necessary.

In Experiments 2 and 3 (Chapter 5 and 6), there is no longitudinal nature to the data, and all participants give numerous measurements on a single occasion. Thus, the inclusion of random intercepts for each of the participants is automatic,

as this reflects the repeated-measures nature of the data.

3.3.4 Selection of Fixed Effects

In the case of categorical response variables, for which a logistic mixed effects regression model is fitted, Wald p-values will be reported.

However, p-values for continuous response variables are calculated using the t-distribution which requires a value for the degrees of freedom. There is ambiguity as to what constitutes an appropriate degrees of freedom in a mixed effects (Pinheiro & Bates, 2000). Thus, for continuous response variables, a likelihood ratio test will be used, and the reported p value will be that returned when comparing the model with the effect in question and the model without that effect.

The likelihood ratio test can be expressed as:

$\chi^2 = deviance_{reduced} - deviance_{full}$

and compares the log-likelihoods of the two models being assessed. This is assumed to approximate the chi-square distribution and so a p value can be generated, with degrees of freedom (df) being calculated from the df for the alternative model minus the df for the null model.

It should be noted that the LR test can only be used when comparing nested models, and so the models constructed here will be developed using a stepwise approach in order to fulfil this requirement.

3.4 Summary

It is only recently that the importance of considering both reasoning and response bias in analysing belief bias data has been fully acknowledged by researchers in this field. Research has shown that traditional analyses are misleading as they conflate response bias with reasoning bias. The analyses presented in this thesis will also examine the results of traditional analyses, and discuss the degree to which conclusions drawn from these results differ from those drawn from the SDT

analysis.

In addition to accuracy scores, response times and confidence ratings will also be considered. It has been argued that response accuracy does not provide sufficient information to acquire an in-depth view of the processes underlying belief bias. Stupple & Ball (2014) advocate what they term a "multi-method approach", and emphasise the importance of triangulation, that is, considering a number of different methods. Given that the focus of this thesis is debiasing reasoning, such an approach is necessary in ascertaining the specific nature of any shift in strategy applied by participants.

Response times will be analysed to attempt to determine whether a change in these can be brought about by debiasing interventions. This is with the caveat that a change or lack of change in response times is not directly indicative of strategy alteration; if indeed it is the type of processing and not the amount of processing that affects performance, as predicted by selective processing theory, this may not be reflected in the response times. Nevertheless, these measures are important in terms of triangulation of available data, and still may give some indication of processing demands, and changes in response times between sessions, for example, may give additional insight into the effect a given intervention is having upon reasoning. Given the importance of individual differences in determining how problems are mentally represented, a change in response time could be taken to indicate a difference in strategy, which may not be apparent in accuracy scores, if, for example, participants are employing additional effort but have insufficient cognitive capacity to arrive at the normatively correct answer.

Confidence ratings also play a crucial part in these analyses; some theories of belief bias argue that it is metacognitive cues, such as the confidence with which a response can be given, that governs whether a Type 1 or Type 2 response is given (e.g. Thompson et al, 2013). Thus, this measure will be analysed both in terms of the raw confidence ratings, and how such ratings corresponds with accuracy, in order to evaluate whether debiasing interventions are effective in better calibrating

response confidence.

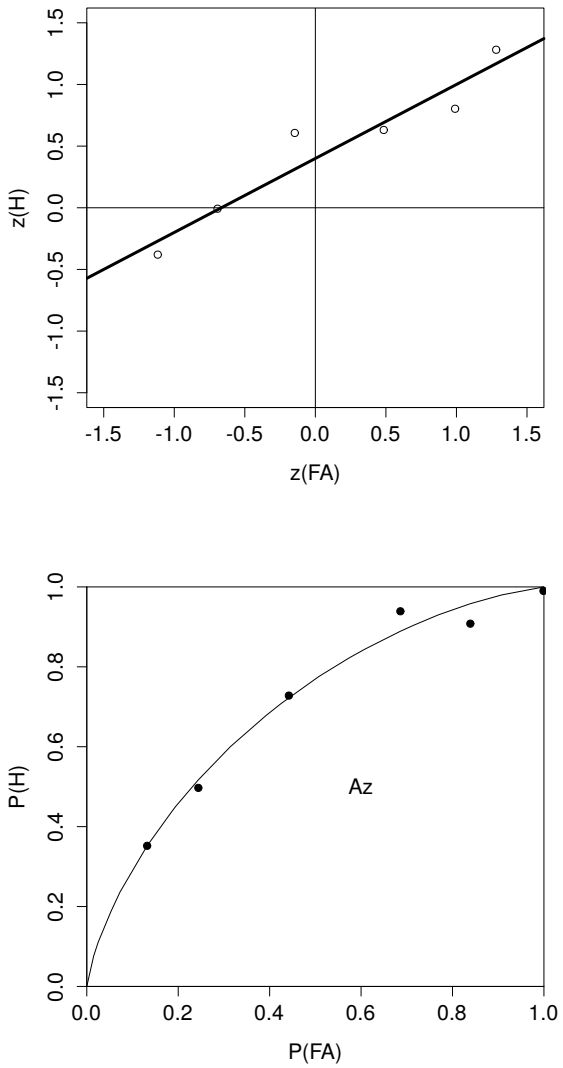
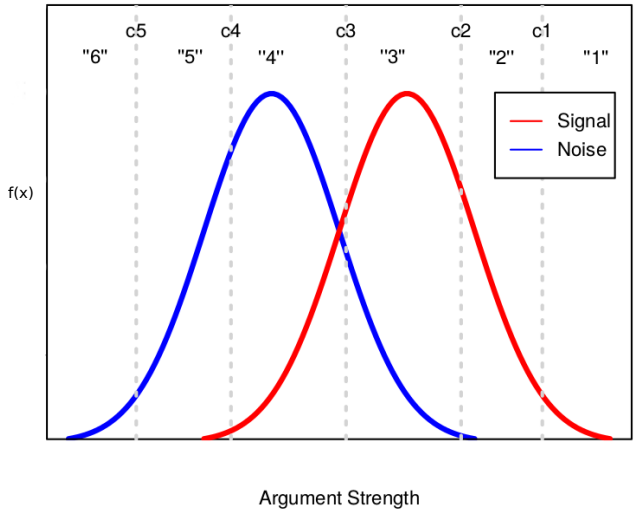


Figure 3.7: Plot showing example of a ROC and an equivalent zROC

Figure 3.8: Plot showing example confidence rating ROC; c1...c5 are the various response criteria.



Chapter 4

Experiment 1

In Chapter 2, I discussed potential ways of debiasing reasoning, and the importance of using SDT analyses in order to correctly distinguish between changes in reasoning and changes in response bias. This chapter presents an experiment in which immediate feedback is used in an attempt to reduce the extent of belief bias. Feedback is one of the simpler potential debiasing methods as it can be automatically generated and thus requires little time or effort to administer. It has been found to be effective in improving cognitive processes in other domains, such as conditional reasoning (Cheng et al., 1986) and analogical reasoning (Cheshire, Ball, & Lewis, 2005). It has been claimed that this is because it causes reasoners to engage in self-reflection, which is key to their strategy development (Ball, Hoyle, & Towse, 2010). Feedback is also thought to lead to higher levels of motivation and more effortful strategies used by learners, even in studies spanning multiple sessions during trials where participants are aware that they will later be given feedback on their answers, but feedback has not yet been given (Vollmeyer & Rheinberg, 2005). It should be noted that feedback does not always lead to improvements in performance; Kluger and Denisi (1996) conducted a meta-analysis of 607 studies which involved a variety of different feedback interventions. They highlight how feedback interventions can produce mixed results depending on the specific nature, relevance and timing of the feedback, and that in some cases, feedback interventions

can lead to deficits in performance. Carroll and Kay (1988) found that for complex tasks involving learning to use a novel computer program, the provision of feedback can interfere with the learning of the task itself, and Jacoby et al (1984) found that more able participants tend to ignore overly simple feedback when completing a task involving discovering underlying rules in a decision-making problem. Kluger and Denisi (1996) argue that one of the main ways in which feedback affects performance is in shifting some focus from the task to the individual; however, too much of this can lead to decreased performance. For example, Mikulincer, Glaubman, Ben-Artzi, and Grossman (1991) found that although feedback led to improved performance on a simple memory task, it led to a decrease in performance on a complex memory task which placed greater demands on cognitive resources. They also highlight how evidence suggests that feedback can increase motivation in task performance, but may be disadvantageous when the removal of such feedback leads to a decrease in motivation (e.g. Komaki, Heinzmann, & Lawson, 1980). Roberts and Newton (2003) distinguish between partial feedback (i.e. correct/incorrect) and full feedback (presenting both an indicator of accuracy along with the correct answer), and highlight how partial feedback tends to have very little effect on improving performance on certain tasks, but full feedback can be much more effective. Empirical evidence supporting this comes from Newton and Roberts (2000), who administered the compass direction task, a task in which participants are asked to work out where a person would end up relative to their initial location after a number of directions such as "they take one step north" and "they take one step west". Full feedback led to a higher chance of participants developing a new strategy than partial feedback. Previous research has investigated the use of feedback in reducing belief bias. Ball (2013) found that evaluative feedback, that is, simple "correct"/"incorrect" feedback was useful in reducing the level of bias shown by participants, and even those who did not receive any feedback showed some increase in performance over time. This is contrary to the predictions of Roberts & Newton, as participants in the study by Ball (2013) were not shown

the original problem alongside the corrective feedback. Ball found that feedback had an immediate effect in reducing belief bias, and it was concluded that this was because the provision of feedback led to a rapid strategy change and decreased reliance on belief to guide reasoning. However, no measures of confidence were taken, and traditional logic, belief and interaction indices were examined rather than SDT measures of accuracy, and so it is unclear whether this change was rooted in accuracy or response bias. Ball (2013) used a microgenetic design (Siegler, 1995), which is more commonly used in developmental studies, and involves repeated testing of the same participants over a short period of time. Siegler and Chen (1998) argue that it is useful as the intensive testing allows insight into changes in strategy over time, and thus allows for a more in-depth exploration than simply comparing separate conditions during a single session. Therefore, a similar design will be used here in order to allow an examination of strategy change over time. Although it has been demonstrated that feedback can improve performance in other reasoning tasks, such as analogical reasoning (Cheshire et al, 2005), there has been little research assessing the impact of feedback on belief bias, with Ball (2013) claiming novelty for the use of this manipulation. The present experiment also is a novel approach, with there being no previous research assessing the effects of feedback on the belief bias effect with an SDT analysis.

Predictions

Theories of belief bias suggested by Dube et al (2010) argue that belief bias manifests as predominantly a response bias. However, given that the syllogisms presented in this study are complex syllogisms, which elicit the reasoning stage effect of belief bias (Trippas et al, 2013), it is predicted that belief bias will initially have an effect on both reasoning and response bias for participants in both the feedback and no feedback conditions. Given that the rapid feedback should alert participants to the fact that their reliance on this heuristic leads to inaccurate responding, it is predicted that feedback will lead to a decrease in response bias.

Stanovich (2009) discusses the importance of both the reasoner's tendency to engage in complex reasoning, and their capacity to do so. Given that even when cognitive ability is accounted for, differences in reasoning accuracy are still predicted by the cognitive style of the individual and how likely they are to engage in more reflecting thinking, the provision of feedback may override many reasoners' natural tendency towards cognitive miserliness, and lead to improved performance, reflected in higher accuracy and longer response times. Previous research has shown that participants tend to show longer response latencies for conflict problems, that is, those in which a response made on the basis of logic would be different to a response given upon the basis of believability (Stupple, Ball, & Ellis, 2012). Valid-unbelievable and invalid-believable problems are both conflict problems, although the effect of longer response latencies is more pronounced on the latter (Stupple et al., 2011). This increase in response time has been attributed to the recognition of the conflict, and increase in the type of processing that is necessary to validate a potential response. It is unclear whether feedback will lead to an increase or decrease in response times. On the one hand, feedback could increase response times, as participants become more adept at recognising such problems and employing more effort to come to a correct answer. Alternatively, feedback may lead to decreased response times, as participants learn to solve such problems by simply engaging in logical reasoning and ignoring belief based cues. However, given that such cues are automatically generated and still need to be suppressed, it seems more likely that any increase in accuracy will be coupled with longer response times. The importance of metacognitive cues in determining whether an analytic or a heuristic response is given is discussed by Thompson, Prowse-Turner, et al. (2013), who found that the more confident reasoners are in their initial heuristic response, the less likely they are to give a result consistent with further, analytic, processing. Thus, feedback may affect reasoning by alerting reasoners to the fact that their initial confidence is not an effective predictor of their response accuracy, especially for conflict problems. This leads to the prediction that for the feedback group,

there will be an initial decrease in confidence, as participants realise that their reliance on believability leads to incorrect responding. However, if their accuracy increases, it is predicted that this will be accompanied by an increase in confidence as a more appropriate strategy is applied, and participants become aware of their improved performance as verified by the feedback that they are receiving.

4.1 Method

Participants

The participants were 48 undergraduate students from Lancaster University who took part in exchange for £18.50. No information on age or gender balance was collected. None had formal training or study in logic or the psychology of reasoning.

Design

Feedback was manipulated between-participants (feedback vs. no feedback) with half of the participants randomly allocated to each condition. Conclusion validity (valid vs. invalid), conclusion believability (believable vs. unbelievable) and session (1, 2, 3, 4, and 5) were within-participants variables.

Materials

Five sets of categorical syllogisms were constructed, which were identical to those used by Ball (2013), plus an additional set. Each set contained 64 syllogisms 16 different subsets of content appeared once each as valid-believable, valid-unbelievable, invalid-believable and an invalid-unbelievable problem. Within each set, half of the premises were in EI mood, and the other half in IE mood, with all conclusions in the O mood. These particular moods were used to prevent confounds relating to syllogistic structure (see Chapter 2), and in order to be comparable to the results of similar studies which used these moods. Only figures one (ABBC) and two (BACB)

were used, for the same reason. Each set contained equal numbers of syllogisms in the different combination of premise mood, figure and conclusion order, to prevent these variables acting as confounds.

The terms (i.e. the specific words which are substituted for 'A', 'B' and 'C' in the above description) in each set of syllogisms were unique to that particular set. Each set was divided into four different subsets so that all content appeared in a different validity and believability combination in each subset. For illustrative purposes, a table containing a single set of syllogisms divided into four subsets can be found in Appendix A.

For the first four sessions, the ordering of sets and subsets presented to each participant was randomised. The subset presented during the final session was the same across all participants in order to ensure that any effects arising at the final stage of testing were not materials-specific.

The materials were rated for conclusion believability 20 independent raters. Once again, these raters were students at Lancaster University. Each conclusion was presented as a sentence on its own on a computer screen and raters were asked to provide an assessment of believability by clicking a radio button. Believability was rated on a scale from -3 (totally unbelievable) to +3 (totally believable) in increments of 1. Once a rating had been given for a sentence, the next sentence was displayed on screen. Two conclusions which were rated between -1 and +1 were replaced with new conclusions which had ratings outside these boundaries. The mean believability for believable problems was 2.20 (SD=0.56) and for unbelievable problems was -2 (SD=0.29).

In addition to the categorical syllogisms, a set of conditional syllogisms was constructed in order to assess whether any changes in reasoning on the categorical syllogisms as a result of feedback would transfer to another task. The conditional syllogisms were similar to the double conditionals constructed by Santamara et al (1998) and were controlled and randomised in the same way as the categorical syllogisms. These were similar in structure (i.e. mood and figure) to the categorical

syllogisms, but contained conditional statements; for example:

> If David has a cup of tea, then David's room gets messy.
> If David tidies his room, then David has a cup of tea.
> Therefore, if David's room gets messy, then David tidies his room.

Procedure

Testing took place across the course of 5 sessions, which took place on alternate days across the course of two weeks. Participants were given the following instructions:

> This is an experiment to test peoples reasoning ability. In this experiment, you will be given sixteen problems in total. For each problem you will be shown two premises and a conclusion. You must presume that the two premises are definitely true, and judge whether the conclusion logically follows on from them or not.
>
> Here is an example problem:
>
> All humans are mortal
> All Greeks are human
> Therefore, all Greeks are mortal.
>
> You will be asked to click either yes" or "no" to indicate whether you think the conclusion follows on from the two sentences. In this case, the answer would be "yes".
>
> You will also be asked, on a scale from 0 to 10, how sure you are that you have answered correctly. 0 indicates that you don't know at all and your response was a guess, and 10 means that you are totally sure that your answer was correct.

> There will be a pause for ten seconds after you have made a response for the answer and how sure you are. The statement will remain on screen, along with your response.
>
> If you are in the feedback condition, during sessions 2, 3, and 4, the word "correct" or "incorrect" will also be displayed on screen.
>
> After 10 seconds has elapsed, the next problem will be shown. Please ensure your phone is on silent, and if you have any questions about the experiment or any details of the task, feel free to ask them now."

The experiment was conducted using software written in Java. Upon confirming that they understood the instructions and were ready to begin, participants clicked a button marked 'start' on the computer screen. The first syllogism then appeared in the centre of the screen, with the question "Does the conclusion necessarily follow from the first two statements" underneath and a pair of buttons labelled "yes" and "no" further down. Beneath this the question "how sure are you that you have answered correctly?" was displayed, with an accompanying set of radio buttons from 0 to 10 with which participants indicated their confidence in their response, with 0 indicating a guess and 10 indicating total confidence in their response.

Once a validity choice and confidence rating had been given, the chosen buttons were highlighted, and the screen paused for ten seconds, before displaying the next syllogism. During this pause, the syllogism and the participant's responses remained on screen. For those in the feedback condition, during sessions two to four only, the word "correct" or "incorrect" was also displayed at the bottom of the screen during the pause between trials. After ten seconds had elapsed, the screen went blank for a second, and then the process was repeated with the next syllogism.

During session five, after completing the final categorical syllogistic reasoning task, participants also completed the conditional syllogistic reasoning task. No feedback was given to any participant on this task.

4.2 Results

Endorsement Rates

Figure 4.1 shows the mean proportion of conclusions of each type endorsed in each condition. A binary logistic mixed-effects model was fitted to the data. The

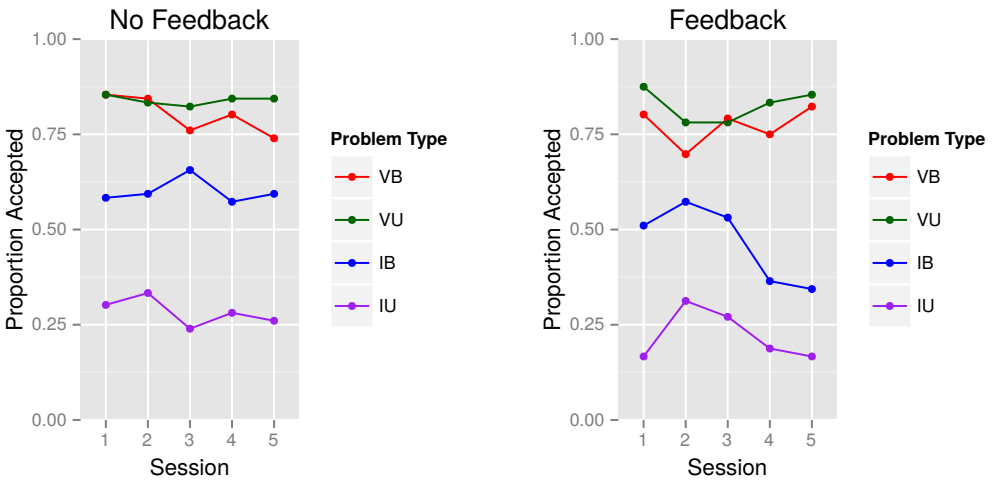


Figure 4.1: Plot of mean endorsement rates by session number, problem type, and feedback status

response variable was conclusion endorsement (whether the participant responded with "valid" or not) and the explanatory variables were feedback condition, validity, believability, and session.

Validity was significant; more valid conclusions were accepted than invalid

Table 4.1: Parameter Values for Main Task Endorsement Rate Model

Variable	β	SE	z	Wald-p
Session	-0.032	0.050	-0.649	0.516
Feedback	-0.070	0.276	-0.253	0.800
Validity (valid)	2.909	0.280	10.401	< 0.001
Believability (believable)	1.309	0.102	12.792	< 0.001
Validity * Believability	-1.619	0.157	-10.305	< 0.001
Session * Feedback	-0.137	0.0715	-1.911	0.056
Valid * Session	-0.068	0.079	-0.863	0.388
Valid * Feedback	-0.514	0.367	-1.399	0.162
Valid * Session * Feedback	0.270	0.110	2.447	0.014

conclusions, $\chi^2(1) = 754.44$, $p<0.001$. Believable conclusions were endorsed more than unbelievable conclusions, $\chi^2(1) = 57.05$, $p<0.001$.

There was a significant interaction between validity and believability, $\chi^2(1) = 106.82$, $p<0.001$. For valid problems, a conclusion was less likely to be accepted if it was believable, $\chi^2(1) = 7.133$, $p=0.007$. However, for invalid problems, believable conclusions led to higher endorsement rates, $\chi^2(1) = 187.44$, $p<0.001$. No other two-way interactions were observed.

A three-way interaction was found between session, feedback, and validity, $\chi^2(4) = 11.17$, $p=0.02$. For valid problems, there was no interaction between session and feedback, $\chi^2(1) = 2.63$, $p=0.10$. However, for invalid problems, the session by feedback interaction was marginally significant, $\chi^2(1) = 3.52$, $p=0.06$. For invalid problems, session did not have a significant effect in the no feedback condition, $\chi^2(1) = 0.42$, $p=0.52$. However, for invalid problems in the feedback condition, less conclusions were endorsed as time passed, $\chi^2(1) =10.46$, $p=0.001$.

The fit of the model was not improved by the inclusion of random slopes, $\chi^2(1) = 0.00$, $p=1.00$.

SDT Measures

ROC curves for the feedback and no feedback conditions for all five sessions can be found in Figure 4.2 and Figure 4.3 respectively. Both of these figures show curved ROCs for most problem types during most sessions. The location of the ROC curves is higher for unbelievable when compared with believable problems, indicating the possibility of higher accuracy on these problems.

The SDT measures were calculated using the method discussed in Chapter 3. The resulting accuracy and response criterion measures were then analysed using a mixed effects model. The response variables were accuracy and response criterion, and explanatory variables were feedback, session, and believability.

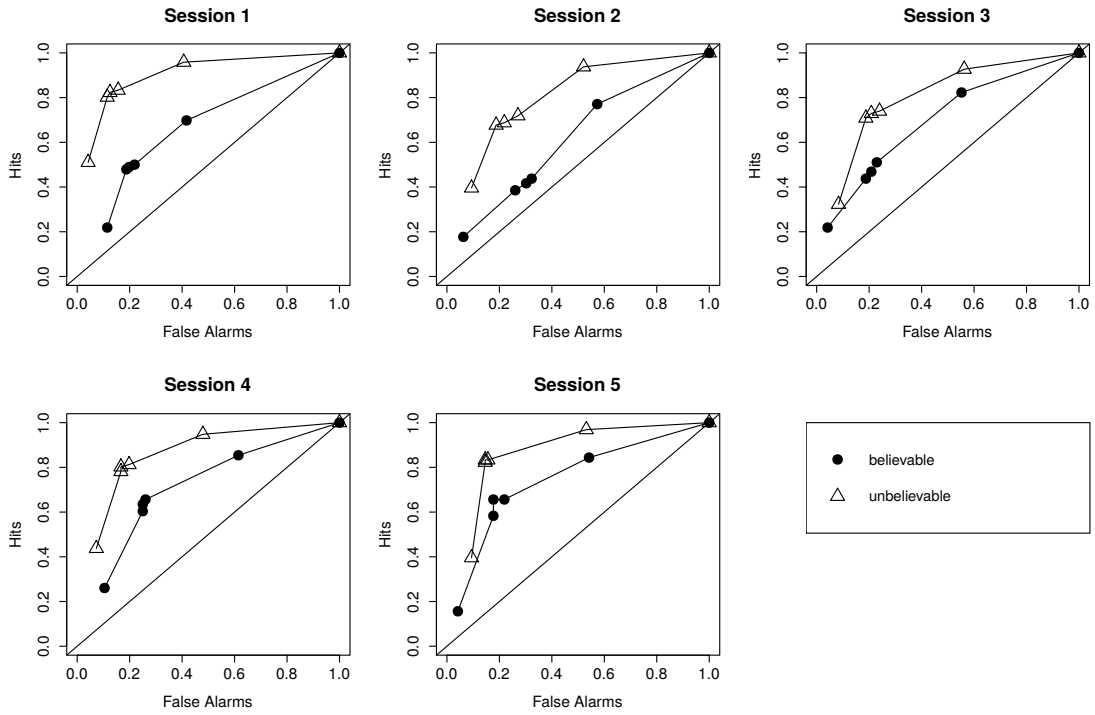


Figure 4.2: ROC curves by session and believability for feedback condition

Accuracy

Figure 4.4 shows the accuracy scores. Participants responded more accurately to unbelievable than believable conclusions, $\chi^2(1) = 39.48$, p<0.001. No other main effects or their interactions were statistically significant. Model parameters can be seen in Table 4.2.

Response Criterion

Figure 4.5 shows the response criteria. There was a main effect of believability, $\chi^2(1) = 35.85$, p<0.001, with participants having a more liberal response criterion for believable problems.

For the response criteria, there was a marginally significant interaction between

Table 4.2: Parameter Values for Main Task Accuracy Model

Variable	β	SE	t	LRT-p
Believability (believable)	-0.083	0.013	-6.42	0.001

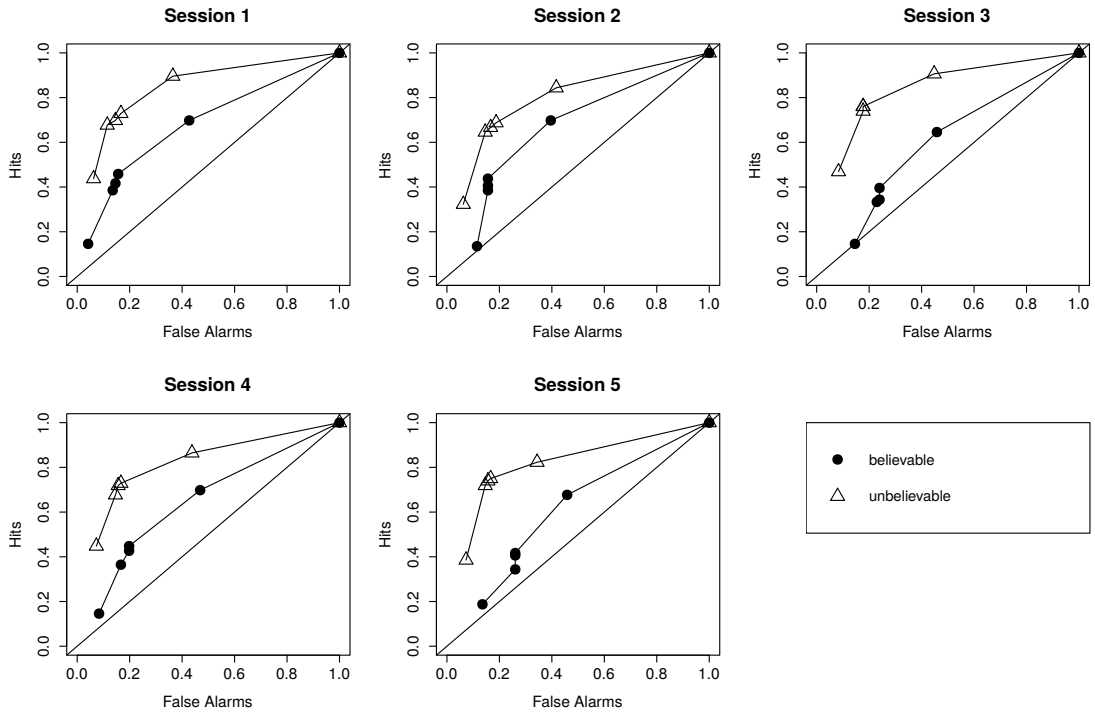


Figure 4.3: ROC curves by session and believability for no feedback condition

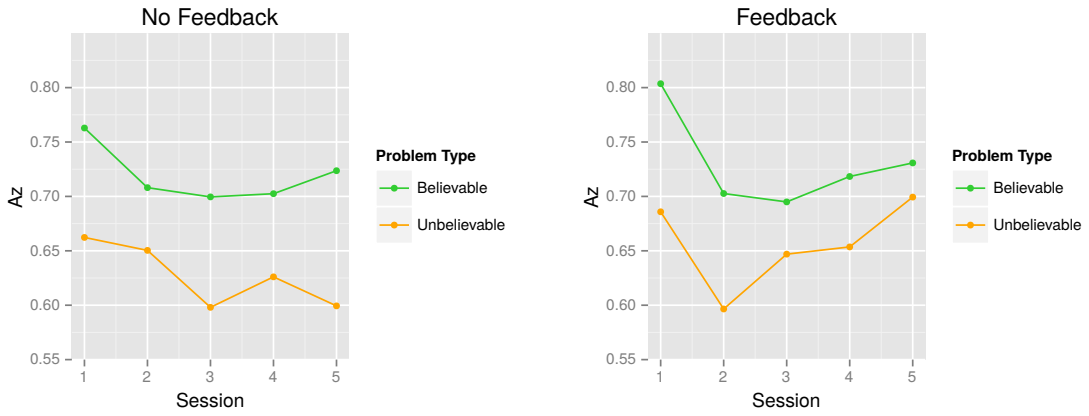


Figure 4.4: Mean accuracy scores by believability and feedback group status

feedback and belief, $\chi^2(1) = 3.43$, p=0.06. For the believable problems, participants in the feedback group had a more conservative response criterion, $\chi^2(1) = 4.17$, p=0.04. However, for the unbelievable problems, there was no difference between the two groups, $\chi^2(1) = 0.42$, p=0.52. Model parameters can be found in Table 4.3.

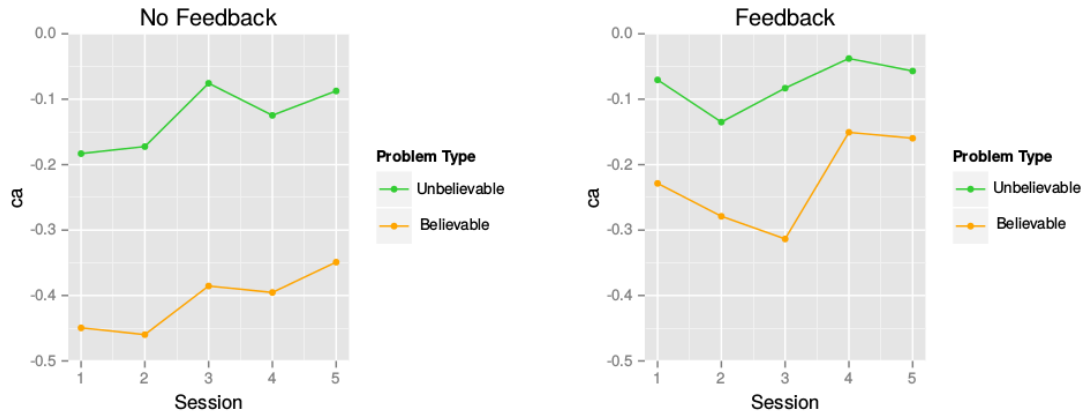


Figure 4.5: Mean response bias scores by believability and feedback condition

Response Times

Figure 4.6 shows the response times. There was a main effect of session, $\chi^2(1) = 53.08$, $p < 0.001$, with response times decreasing across the five sessions. Validity was significant, $\chi^2(1) = 97.90$, $p < 0.001$, with participants responding faster to valid than invalid problems. Believability had an effect, $\chi^2(1) = 39.77$, $p < 0.001$, with participants taking longer to respond to believable problems. Feedback was not significant, $\chi^2(1) = 0.99$, p=0.3191. No two-way or higher order interactions between the main effects were significant.

Correlated random slopes improved the model, $\chi^2(2) = 88.56$, $p < 0.001$, indicating that there was a significant amount of variation between participants' response times, and those who had an initially longer response time showed a greater increase in response time across the five sessions. Parameter values can be found in Table 4.4.

There were no effects of accuracy or response criterion on response time.

Table 4.3: Parameter Values for Main Task SDT Response Bias Model

Variable	β	SE	t	LRT-p
Believability (believable)	-0.279	0.050	-5.639	< 0.001
Feedback (present)	0.052	0.077	0.067	0.089
Believability * Feedback	0.130	0.070	1.850	0.064

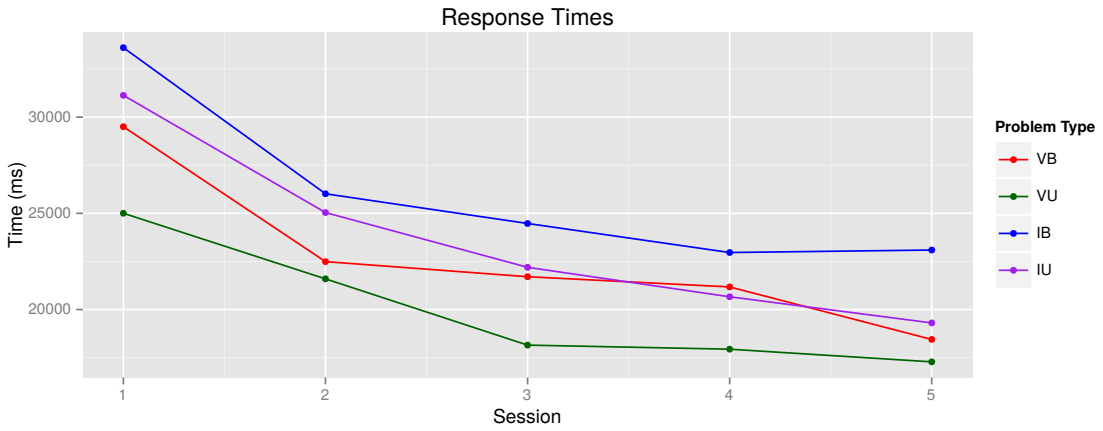


Figure 4.6: Response times by problem type

Confidence

Figure 4.7 shows the confidence ratings across time for both groups. There were no main effects of feedback or session on confidence. Participants were more confident in the responses they gave to valid problems, $\chi^2(1) = 154.61$, p<0.01, and less confident in responses given to believable problems, $\chi^2(1) = 8.53$, p<0.01.

There was a two-way interaction between validity and believability. For valid problems, believability had no effect on confidence rating. However, for invalid problems, believability led to lower confidence ratings, $\chi^2(1) = 29.10$, p<0.01.

Validity and feedback, also interacted. Splitting the data by validity revealed no significant effects; that is, for both the valid and invalid problems, the feedback group were no more confident than the no feedback group. However, when split by feedback, participants who were given feedback were more confident of their responses to valid than invalid problems, $\chi^2(1) = 49.33$, p<0.01. The same pattern was observed for the no feedback group, $\chi^2(1) = 110.78$, p<0.01; however, the

Table 4.4: Parameter Values for Main Task Response Times Model

Variable	β	SE	t	LRT-p
Session	-0.102	0.011	-9.75	< 0.001
Validity (valid)	-0.143	0.014	-10.01	< 0.001
Believability (believable)	0.092	0.014	6.41	< 0.001

differences between the two problem types were much greater.

There was a significant interaction between session and feedback; participants who were not given feedback became more confident of their responses across time, $\chi^2(1) = 4.67$, p=0.03, whereas participants who were given feedback became less confident of their answers across time, $\chi^2(1) = 16.85$, p<0.01.

There was a three way interaction between validity, session, and feedback. For valid problems, the interaction between session and feedback was not significant. However, for invalid problems, the interaction between session and feedback was significant, $\chi^2(3) = 24.63$, p<0.01. Breaking this down by feedback group, on the invalid problems, confidence ratings increased across time for the no feedback group, $\chi^2(1) = 8.27$, p<0.01, t=2.88. However, the feedback group became less confident on invalid problems across time, $\chi^2(1) = 16.93$, p<0.01. Parameter values can be found in Table 4.5.

Final Session

The results from the final session were analysed separately in order to see whether they differ from the above results which explore change over time.

Table 4.5: Parameter Values for Main Task Confidence Model

Variable	β	SE	t	LRT-p
Validity (valid)	0.916	0.173	5.310	< 0.001
Believability (believable)	-0.397	0.0704	-5.634	< 0.001
Session	0.111	0.035	3.150	0.196
Feedback (present)	0.557	0.411	1.355	0.360
Session * Feedback	-0.255	0.050	-5.113	< 0.001
Validity * Feedback	-0.840	0.234	-3.594	0.018
Validity * Believability	0.353	0.010	3.545	< 0.001
Validity * Session	-0.109	0.050	-2.196	0.385
Validity * Session * Feedback	0.190	0.070	2.706	0.024

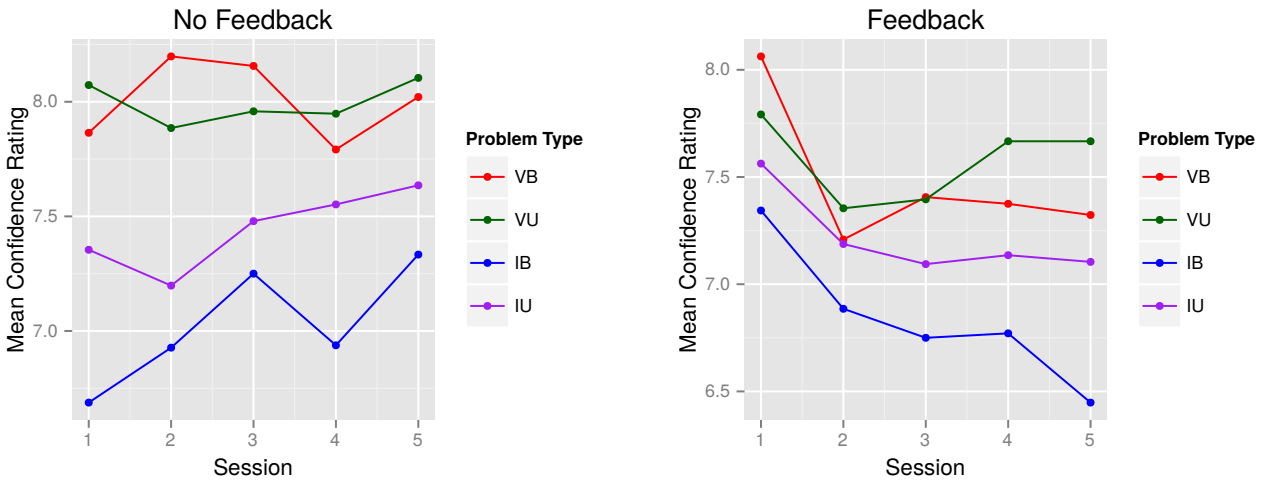


Figure 4.7: Confidence ratings by problem type and feedback condition

Endorsement Rates

Validity was significant, $z = 8.412, p < 0.001$, with more valid than invalid problems being accepted. Believability was also significant, $z = 5.409, p < 0.001$, with more believable than unbelievable problems accepted. Feedback was also significant, $z = -3.209, p = 0.001$, with the feedback group endorsing less conclusions.

Validity and believability interacted, $z = -4.855, p < 0.001$; for valid problems, there was no effect of believability, $z = -1.773, p = 0.076$, whereas for invalid problems believable conclusions were accepted more often than unbelievable conclusions, $z = 1.667, p < 0.001$. Validity and feedback also interacted, $z = 3.302, p < 0.001$; for valid problems, there was no effect of feedback, $z = 0.918, p = 0.358$, whereas for invalid problems, feedback led to lower rates of acceptance, $z = -2.198, p = 0.028$.

SDT - Accuracy

Only believability significantly predicted accuracy; $\chi^2(1) = 7.431, p = 0.006$, with participants responding more accurately to unbelievable problems.

SDT - Response Bias

Believability also predicted response bias, $\chi^2(1) = 4.433, p = 0.035$, with a more liberal bias on believable problems.

Response Times

Participants who were given feedback during sessions 2 to 4 showed longer response times to problems during session 5, $\chi^2(1) = 4.929, p = 0.026$. Participants took longer to respond to invalid problems, $\chi^2(1) = 23.913, p < 0.001$. Believable problems were answered more slowly than unbelievable problems, $\chi^2(1) = 7.632, p = 0.006$.

Confidence Ratings

Participants were more confident on their answers to valid problems, $z = 6.501, p < 0.001$. Believable problems also led to lower confidence, $z = -3.578, p < 0.001$.

Conditional Reasoning Task

Endorsement Rates

The results from the final task containing the conditional reasoning problems were examined using linear mixed models. Valid problems were more likely to be accepted than invalid problems, $\chi^2(1) = 28.28$, p<0.001. Believable problems were more likely to be accepted than unbelievable problems, $\chi^2(1) = 11.67$ p<0.001. There was no effect of feedback, $\chi^2(1) = 1.67$, p=0.20. Validity and believability did not interact, $\chi^2(1) = 0.76$, p=0.38. No two-way or higher order interactions were significant.

SDT Measures

Mixed effects models were fitted to the accuracy and response criterion measures, using feedback and believability as predictors. For the accuracy score, neither

factors were significant ($ps > 0.05$).

For the response criterion, participants had a more liberal response criterion for believable than unbelievable conclusions, $\chi^2(1) = 11.71$, p=0.001. Feedback was not significant, and did not interact with believability.

Response Times

There was no difference between participants in terms of response times for the two groups, and no effects of validity or believability.

There was a significant effect of accuracy on response time, with more accurate responses leading to longer response times, $\chi^2(1) = 4.75$, p=0.03. There were no effects of feedback, believability or response criterion, and none of the effects interacted.

Confidence Ratings

There were no effects of validity, believability, or feedback on confidence ratings, and none of these factors interacted.

4.3 Discussion

This experiment aimed to investigate whether feedback is effective in reducing belief bias. Feedback is not an effective intervention for reducing or eliminating the reasoning component of the belief bias effect, although it does affect response bias, as shown by the significant interaction between feedback and believability on participants' response criteria, but no effect on their reasoning accuracy. However, the analysis of the final session alone shows that levels of response bias were similar for both groups; whether this was due to the reduced statistical power of only analysing responses from a single session, or due to a lack in difference between the two groups, it is clear that feedback did not debias reasoning. The present research adds to the growing body of evidence which suggests that using traditional analyses

such as logic, belief, and interaction indices can lead to misleading results, conflating reasoning and response and the SDT approach taken here is crucial in separating these effects. Relying upon the analysis of raw endorsement rates may have led to the conclusion that the decrease in acceptance of invalid problems across time for the feedback group reflected a genuine improvement in reasoning. However, the SDT analyses clearly show that giving participants feedback had no effect on their reasoning accuracy, but did alter their response criteria, making them less likely to endorse believable conclusions, but without reflecting any change in their ability to distinguish between valid and invalid stimuli. The ROC curves also support the use of the SDT model, as it can be seen that they are curved rather than linear in shape, an indicator of the need for an SDT approach (Dube et al, 2010). The SDT analyses showed a significant effect of belief on both accuracy and response bias. These findings are at odds with a number of theories of belief bias which argue that the belief bias effect is either a reasoning bias (e.g. selective scrutiny Evans et al, 1983; misinterpreted necessity Evans et al, 1983; mental models Oakhill et al, 1989; metacognitive uncertainty Quayle and Ball, 2000) or response bias (e.g. Dube et al, 2010), but not both. These findings are consistent with the model proposed by Trippas et al (2013) which argues that belief bias manifests both as a reasoning and response bias, with the former effect only occurring when complex reasoning takes place. Feedback led to a more conservative response bias. In concordance with Ball (2013), the effect of feedback was immediate, implying that the participants' knowledge that they would be receiving feedback on their answers prompted a strategy change. However, unlike the findings of Ball (2013), who concluded that feedback improved reasoning, here the lack of change in accuracy shows that the strategy change was suboptimal. The response time data also support this view; one might expect that a change in reasoning strategy should be accompanied by a change in response time to reflect the change in processing; however, this was not the case, with there being no significant main effect of feedback nor any interactions with any other variables. The only effect of feedback on response time was during

the final session; however during this session no feedback was provided for either group, so the longer responding may have been simply the consequence of using a slightly different format. A possible limitation of this study is that the feedback provided was not sufficiently detailed. Participants were shown the word "correct" or "incorrect" on screen, along with the original problem and the answer that they had selected. Although the feedback presented here is closer to what Roberts and Newton (2003) term "full feedback" than "partial feedback", as the accuracy indicator is presented alongside the original problem. Although the correct answer is not presented, given that the problem is binary choice, extrapolation of the correct answer is simple. However, this aspect of presentation may still be important in inducing effective strategy change. The standard belief bias effect of superior performance on unbelievable problems was somewhat supported in both the SDT and endorsement rate analyses; participants endorsed more invalid problems if they were believable, and showed lower SDT accuracy scored for believable problems. Surprisingly, however, it was also found that valid conclusions were less likely to be endorsed if they were believable. Although this contradicts previous findings; typically valid-believable and valid-unbelievable problems show similar levels of endorsement (e.g. Evans et al, 1983; Ball et al, 2006; Klauer et al, 2000; Newstead et al, 1992); it could be a further indication that participants were trying to change their reasoning strategy to be less reliant on heuristics, but were unable to do this in a sophisticated manner. This is supported by the disappearance of this effect in the analysis of session 5 alone; believability only had an effect here on invalid problems. An alternative interpretation is simply that the problematic nature of endorsement rates leads to inconsistencies in the validity by believability interaction; Heit and Rotello (2014) comment that it isn't uncommon for there to be variability in the interaction effect or lack of effect altogether on occasion when using traditional measures. There were no effects of feedback on accuracy or response times for the conditional syllogistic reasoning task. Given the lack of effects of feedback in the categorical reasoning tasks, this is unsurprising. There

may have also been intrinsic differences between the two tasks; there were no effects of belief bias as a reasoning bias on the conditional syllogisms, but it did show up as a response bias. Unpublished doctoral work by Solcz (2011) supports this idea, showing evidence that the way in which reasoners evaluate premises and conclusions is different for categorical and conditional syllogisms. The results of the conditional reasoning task show further evidence of inconsistencies in the belief bias effect as measured by traditional indices; no interaction between validity and believability was found in the endorsement rate analysis; however, the SDT analysis shows problem believability affecting response bias.

These results may seem to disagree somewhat with the claims of Thompson et al (2013), who argue that a reasoner's confidence in their response acts as a cue to whether Type 2 processing takes place. Here, the decrease in confidence for the feedback group did not lead to an increase in accuracy. However, the results do not fully dispute this theory, as this link was found by Thompson et al to occur when an answer and initial confidence rating were given immediately after the presentation of an argument, presumably before complex processing could take place, whereas here the confidence rating was given after an answer has been settled on. Nevertheless, this does raise the question of whether the lack of improvement in performance of participants in the feedback condition was due to a lack of override of a Type 1 response, or if the override was successful, but a lack of complete knowledge of the principles of logical necessity resulted in their continued erroneous responses, with participants taking the presence of confirming model for invalid believable problems as evidence of their validity. In other words, participants presumed that the presence of a confirming model meant that the problem was valid, rather than both the presence of a confirming model and absence of disconfirming model. This view is consistent with claims by Stanovich and Stanovich (2010) that it is ineffective to suppress a heuristic response when there is no alternative to replace it. If there is a lack of knowledge of the rules or strategies needed to successfully give a Type 2 response, a Type 1 response may be given instead; this

is referred to by Stanovich as a "mindware gap" (Stanovich, 2009), and this will be explored in more detail in the next chapter. Although feedback did not improve reasoning across the five sessions in this experiment, it may play a useful purpose in encouraging participants to engage in reflective behaviour. This is evidenced by the divergence in confidence ratings for the two groups on invalid problems; whilst participants who did not receive feedback became more confidence of their answers, those who did receive feedback became less confident. Although no improvement in reasoning was conferred from the provision of feedback, across a longer time scale, the metacognitive change may lead individuals to further seek validation for their reasoning. Another potential explanation for the lack of result of feedback is individual differences between reasoners. There has been a growing awareness of the need to examine individual differences in belief bias. In a rule-based learning task, Kelley and McLaughlin (2012) found that participants with a higher cognitive ability performed better with less intensive feedback, with this being reversed for participants with a lower cognitive ability, and highlighted the importance of an intervention being suitable for the individual learner. These results are broadly in agreement with research by Heit and Rotello (2014), who conducted research involving a different potential debiasing intervention which involved giving participants altered instructions, which had been found in previous research to alter the extent of the belief bias effect (Newstead et al, 1992). However, when Heit and Rotello (2014) applied an SDT analyses to the results of their replication of this experiment, they found that the altered instructions led to a change in response bias but not reasoning accuracy. Given that the degree to which believability affects reasoning can be moderated using manipulations of factors such as time constraints and problem complexity (Trippas et al, 2013), it seems interesting to examine which other interventions can be used to moderate the effect of belief on reasoning accuracy and not just a shift in response bias. Individual differences, for example, cognitive abilities, also have an effect on to what degree the belief bias effect manifests itself (Trippas et al, 2013) and examining how cognitive ability

moderates susceptibility to these interventions is also an important question for future research. In the next chapter, I will present an experiment which extends the approach taken by Heit and Rotello (2014). The aforementioned study included a manipulation which involved giving participants instructions with differing content, one standard set and one which emphasised the importance of logical necessity. These instructions led to a change in response bias, but not reasoning bias. In Experiment 2, participants will be given identical instructions to these, but by also including a range of individual difference measures, I will be able to explore whether participants of differing cognitive abilities and styles show differences in their responses to debiasing interventions.

Chapter 5

Experiment 2

The results of Experiment 1 show that even when reasoners are provided with feedback to indicate that their reasoning strategy is suboptimal, without a superior strategy to replace it with, there is no distinct change in accuracy. Instead, only response bias is affected. A possible explanation for the results of Experiment 1 is that participants may not have fully understood the application of logical necessity to syllogistic reasoning; that is, if a syllogism is merely possible but not necessitated by the premises, it is invalid. Although Heit and Rotello (2014) found that such a manipulation did not improve reasoning performance, no consideration was made of individual differences. Accounting for such differences may be crucial in identifying successful debiasing strategies given that it has been found that participants with higher cognitive ability show belief bias as both a reasoning and response bias, whereas those of lower cognitive ability only exhibit signs of response bias (Trippas et al, 2013), and it has been claimed that ignoring individual differences obscures differences in response patterns (e.g. Stupple et al, 2011; Trippas et al, 2014). Experiment 2 aims to investigate whether providing participants with instructions that emphasise the importance of logical necessity will lead to improved performance, and if any changes found are contingent on participants' cognitive ability or cognitive style (for example, the tendency to engage in open-minded thinking). Previous research has found that reasoning

ability can be predicted by measures of cognitive ability (e.g. Trippas et al, 2013) and cognitive style (e.g. Baron, 2008), and so I will discuss below how the efficacy of debiasing interventions may also be affected by these differences. Firstly, I will discuss accounts of individual differences in reasoning, before moving on to outline various measures of these differences. I will then examine the literature around using instructions to alter strategy use before outlining arguments surrounding the use of online experiments.

Theories of Individual Differences

Stanovich (2009) outlines a tripartite model of thinking which explains individual differences in deductive reasoning task performance; it is comprised of the autonomous mind, the reflective mind, and the algorithmic mind. The autonomous mind is responsible for the kind of automatic responses that many dual process theories associate with Type 1 responses. Type 2 processing is governed by the reflective mind and the algorithmic mind. The algorithmic mind is associated with differences in working memory capacity and the ability to hold multiple representations of a problem in mind at any one time, and is related to individual differences in cognitive ability. The reflective mind governs the tendency to examine a variety of evidence before drawing conclusions, and these tendencies can be distinguished by measures of cognitive style. Thus, if an individual of higher cognitive ability has a low tendency to consider alternative representations, this higher ability will confer them few advantages and so are likely to still provide biased responses (Stanovich & West, 1997). This is reflected in research which has shown that these thinking dispositions or styles are still predictive of task performance on heuristics and bias tasks even when differences in cognitive ability have been controlled for (Toplak & Stanovich, 2003).

Stanovich (2009) constructed a taxonomy of thinking errors and sources of bias with five main categories of thinking errors: override failures, defaulting to autonomous responses, mindware gaps, contaminated mindware, and serial

associative cognition with focal bias. Override failures are due to reasoners engaging in both Type 1 and Type 2 processing, but being unable to substitute their automatic (Type 1) response with a more analytic (Type 2) one. Override failures differ from defaulting to an autonomous response, as the latter involves no Type 2 processing at all and only Type 1 processing is engaged to give a response. The term "mindware" refers to the individual's knowledge of rules or strategies, such as the laws of probability or logical necessity. Whilst override failure accounts presume that the relevant mindware is available but not utilised, mindware gap accounts suggest that an override is attempted but the relevant rules are not available. Furthermore, mindware contamination describes situations in which the reasoner has given a response based upon an incorrect or illogical rule. The final category, serial associative cognition with focal bias, refers to situations when a reasoner does not employ sufficient effort in exploring multiple representations of a problem to ascertain the normatively correct answer. Although Stanovich (2009) argues that the belief bias is down to cognitive miserliness, in the form of override failure, other accounts place its effect as the result of other thinking errors. For example, the modified version of selective processing theory suggested by Stupple et al (2011) which argues that reasoners engage in satisficing searches, fits better into the serial associative cognition category, and indeed an earlier account by Stanovich and West (2008) makes similar inferences. In addition, an override failure account may be ignoring the more subtle details of the phenomenon. Thompson and Johnson (2014) found that higher ability reasoners showed better performance compared to lower ability reasoners, and focussed more on Type 2 processing, measured by time take to rethink an answer and probability of changing this answer, specifically on conflict problems. However, for this higher ability group, the extra time rethinking an answer did not lead to an increase in normative responding, despite the extra Type 2 processing. Thompson and Johnson (2014) suggest that this may be due to differences in reasoning accounted for by cognitive capacity emerging only at the point when an initial response is given, as the higher ability group may find it

easier to apply the relevant logical rules intuitively and immediately (cf. De Neys, 2012). However, this is at odds with the findings of Trippas et al (2013), who found that when solving syllogisms under time pressure, higher ability reasoners actually performed worse on unbelievable problems than believable problems, a reversal of the typical pattern found when no time limit exists. The differences between these findings could be due to task complexity; Thompson and Johnson (2014) note differences between findings on syllogistic reasoning tasks when compared to other, similar, heuristics and biases tasks such as base-rate neglect, and point to the increased complexity of the former. There is also debate as to whether both cognitive style and ability play a part in determining reasoning ability, or whether the high correlation between the two means that the significant effects of one can account for the positive effects of the other (Trippas, Pennycook, Verde, & Handley, 2015). The only research to date which has attempted to answer this question using SDT analyses to examine the effects of belief bias is that of Trippas, Pennycook, et al. (2015) who found that having an analytic cognitive style accounted for the effects of cognitive ability on belief bias for both response bias and reasoning bias. They argue that this shows support for the quantity over the quality argument differences in performance can be attributed to individuals with a more analytic cognitive style engaged in a greater amount of reasoning, rather than being better reasoners. The lack of consistency in accounts of the role of individual differences suggests that more research is needed to better describe the underlying causes of belief bias, and it is clear that individual differences must be accounted for in any such account. As mentioned earlier, a number of different scales and tests exist which allow us to distinguish between individuals on the basis of cognitive ability and cognitive style. Such tasks enable us, for example, to establish whether an individual tends to rely on their initial response, or if they prefer to engage in more deliberate and effortful thinking, and thus are linked with Type 1 and Type 2 responding. Previous research has shown an association between performance on these tasks and syllogistic reasoning performance (e.g. Trippas et al., 2013; Trippas,

Pennycook, et al., 2015). Below, I will discuss a number of these measures.

Measures of Individual Differences

The Rational-Experiential Inventory (REI; Epstein, Pacini, Denes-Raj, & Heier, 1996) aims to measure individual differences in rational and heuristic thinking tendencies. It is a self-reported scale comprising a number of items from the Need For Cognition scale (NFC; Cacioppo & Petty, 1982), which is designed to measure analytic thinking, and extra questions from a Faith in Intuition (FI) scale, designed to capture differences in reliance on heuristics. Epstein et al. (1996) argue that this scale is correlated with independent measures of heuristic thinking. A revised version containing 40 items (REI-40; Pacini & Epstein, 1999) was developed, which has been reported to correspond to scores on the ratio bias task. The ratio bias task is believed to invoke conflict between a Type 1 and a Type 2 response (Mevel et al., 2014), and Pacini and Epstein (1999) argue that high reliance on experiential (Type 1) processing, inadequate rational (Type 2) processing, or both, leads to higher levels of bias. Prowse-Turner and Thompson (2009) found that high REI scores were related to higher levels of response confidence, despite no relationship between REI score and accuracy. The Actively Open-Minded Thinking scale (AOT; Stanovich & West, 1997) is another self-reported scale designed to measure thinking dispositions, or the tendency to be reflective (Stanovich & West, 1998) and contains 41 items. It predicts differences in response patterns on tasks involving the evaluation of arguments, even when ability has been controlled for (Stanovich & West, 1997), and is thought to indicate the likelihood of reasoners to consider opposite possibilities, with higher scorers also having better calibrated confidence (Baron, 2008). Trippas, Verde, and Handley (2015) discuss the importance of considering using behavioural rather than self-reported scales when measuring cognitive abilty. An alternative measure of ability is the Cognitive Reflection Test (CRT; Frederick, 2005) which differs from the REI and AOT as it is not a self-reported scale. Instead, the CRT contains just three items:

1. A bat and a ball cost $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?.

The questions which comprise the test are all designed to lead to an easily accessible quick heuristic response which must be suppressed in order for a correct answer to be given. For example, the heuristically cued response for question 1 is "10 cents", whereas the correct answer is "5 cents". Frederick (2005) reports links between CRT scores and delayed gratification. Frederick administered a battery of tests to performance, and found that CRT scores were correlated with measures of cognitive ability, they were highly correlated with self-rated measures of cognitive style which are believed to measure similar constructs, for example, the NFC. This indicates the possible unreliability of such scales in distinguishing between different types of respondents. The AH4 Group Test of General Intelligence (Heim, 1970) has been used in belief bias research to distinguish between participants on the basis of cognitive ability (e.g. Newstead et al, 2004; Trippas et al 2013). It is a 65 item test which contains questions related to verbal and numerical reasoning, and is positively correlated with performance on deductive reasoning tasks. Some of its contents include items which involve simple calculations which require multiple components to be remembered simultaneously, number series, and questions containing complex conditional statements. Although many studies have linked scores on these tests and scales to reasoning performance, it should be noted that the vast majority of studies examining the link between these individual measures and syllogistic reasoning

ability use the traditional indices, and so may conflate individual differences in reasoning ability with individual differences in response bias. Given the ways in which individual differences can have strong and varied effects on individuals' reasoning, it is only logical to assume that individual differences will mediate the effect of any debiasing interventions. The experiment presented in this chapter includes a number of these measures to allow us to specifically investigate where variation in susceptibility lies. I will now go on to discuss the use of instructional manipulations in debiasing reasoning.

The Use of Instructions to Reduce Belief Bias

Instructions on reasoning principles have been found to be successful in debiasing the sunk-cost fallacy (Larrick, Morgan, & Nisbett, 1990). The sunk-cost fallacy is a phenomenon whereby individuals tend to base decisions about future investments of time, money or effort too heavily on previous investments rather than choosing the most beneficial option on the basis of future rewards. Larrick et al. (1990) found that giving participants training using more normative rules by which to make decisions led to participants being less biased by sunk costs. Macpherson and Stanovich (2007) found that instructions encouraging participants to rely less on prior beliefs improved accuracy on syllogisms for which there was a conflict between logic and belief. Cognitive ability moderated the effectiveness of the instructions, with participants of lower cognitive ability gaining the most benefit from the instructions stressing the need to ignore prior knowledge. However, the analysis used in this study did not separate reasoning and response bias. Relating more specifically to syllogistic reasoning tasks, Prowse-Turner and Thompson (2009) found that reasoners had a tendency to misclassify conclusions that were merely possible as necessary, but those who had been trained to represent syllogisms using Venn diagrams made fewer such errors. Using Stanovich' taxonomy of thinking errors, the predominant theories of belief bias could be said to place the locus of this effect in override failure or serial associative cognition. Either way, the error

can be seen as based upon overreliance on Type 1 responses. However, participants of high cognitive ability and possessing a reflective cognitive style should be able to complete syllogistic reasoning tasks to normative standards, but consistently do not. Therefore, the remaining component to improve reasoning standards should be mindware. Thus, I predict that providing participants with instructions which explicitly outline the importance of logical necessity should lead to participants of higher cognitive ability (as measured by AH4 score) and those with a reflective or open-minded cognitive style or tendency to inhibit Type 1 responses (as measured by REI, AOT, and CRT scores) to improve performance, whereas those of lower ability or less reflective style will show little improvement, as they will still be unable to generate an accurate representation.

Heit and Rotello (2014) is the only current study which has used an SDT approach to analyse data collected after instructional manipulations. Participants were given either standard or augmented instructions, based on prior research by Evans et al (1994) and Newstead et al (1992). The instructions were as follows. The italicised sections were only included in the augmented instructions condition:

> This experiment is designed to find how people solve logical problems. Your task is to decide whether each conclusion follows logically from the information given in that problem. The premisesthe information givenappear above the line and the conclusion appears below the line.
>
> You must assume that all the information you are given is true; this is very important. If, and only if, you judge that a given conclusion logically follows from the information given you should answer "Valid." If you think that the given conclusion does not necessarily follow from the information given you should answer "Not Valid." Also, you will be asked how confident you are in this judgement.
>
> *Please note that according to the rules of deductive reasoning, you can only endorse a conclusion if it definitely follows from the information given. A conclusion that is merely possible, but not necessitated by the*

> *premises is not acceptable. Thus, if you judge that the information is insufficient and you are not absolutely sure that the conclusion follows you must reject it and answer "Not Valid."*
>
> Please take your time and be certain that you have the logically correct answer.
>
> *REMEMBER, IF AND ONLY IF YOU JUDGE THAT A GIVEN CONCLUSION LOGICALLY FOLLOWS FROM THE INFORMATION GIVEN YOU SHOULD ANSWER "Valid," OTHERWISE "Not Valid."*

It was found that the augmented instructions affected response bias, but did not lead to an improvement in accuracy. However, no individual difference measures were taken, and given the variation found across participants in other studies, this may account for the lack of change. Other approaches to debiasing have suggested that manipulations such as disfluency can have an effect on task performance that is contingent on differences between participants although the results of these studies vary and seem somewhat task-dependent. Thompson, Turner, et al. (2013) found that disfluency increased accuracy on the CRT for participants of high cognitive ability, whereas Trippas, Handley, and Verde (2014) found that the use of a disfluent font in syllogistic reasoning problems led to a detriment in performance for more able reasoners, due to the increased demands on cognitive resources. Macpherson and Stanovich (2007) found a correlation between correct answers on conflict syllogisms and both cognitive style, and cognitive ability. In addition to this, they found that they were able to debias participants using specifically designed instructions. However, these effects varied according to cognitive ability; whilst the instructional manipulation had little effect on participants in the higher three quartiles, participants in the lower quartile showed a marked improvement when given more detailed instructions. However, as these results were analysed using traditional indices, it still remains to be seen whether these differences were caused by changes in reasoning bias or response bias. Previous studies have shown that

belief bias is associated with reasoning bias in reasoners of higher cognitive ability (Trippas et al, 2013), but only with response bias in participants of lower cognitive ability. This difference in reasoning accuracy dependent on problem believability is thought to be the consequence of a different search strategy for unbelievable problems. Thus, the use of instructions emphasising logical necessity may alter the search strategy participants use for unbelievable problems and decrease the belief bias shown by participants who show this bias as a reasoning bias. Although previous studies have shown null effects of instructional manipulations (e.g. Evans, Handley, & Harper, 2001; Heit & Rotello, 2014), they did not distinguish between participants on the basis of cognitive ability, and given the differences in the way belief bias affects participants of different abilities, it is yet to be seen whether this distinction is important. Given that research has shown that reasoners are adept at identifying conflict between logic and belief even when responding inaccurately, as shown by the lower confidence ratings and longer response times for conflict problems, the augmented instructions are expected to increase their reasoning accuracy by providing extra motivation to analyse problems in more depth to ensure the conditions for necessity are met. Indeed, as cognitive style has been shown to predict reasoning accuracy even when cognitive ability has been controlled for, the augmented instructions which encourage further processing should lead participants who do not typically tend to engage in additional processing to show an increase in these tendencies.

Online Studies Using Amazon Mechanical Turk

The study presented below consists of data collected in the lab, and data collected online using Amazon Mechanical Turk. Amazon Mechanical Turk, hereafter referred to as MTurk, is a website which is typically used by individuals or organisations ("requesters") to pay anonymous individuals ("workers") to complete tasks of varying lengths and complexities. Details of requests, termed Human Intelligence Tasks (HITs) are displayed to workers, who choose which HITs to accept, based

upon their preferences regarding task type and compensation rate. In more recent years, there has been a growing trend towards the use of MTurk for data collection in psychology (e.g. Alter, Oppenheimer, & Zemla, 2010; Eriksson, Simpson, et al., 2010). One advantage of using MTurk is the possibility of increased sample diversity and thus overcoming the oft-cited criticism of psychology studies testing fairly homogenous groups, typically psychology undergraduate students (e.g. Sears, 1986). Buhrmester, Kwang, and Gosling (2011) compared data collected using MTurk to what they described as a "standard internet sample", and found similar proportions of respondents in terms of gender, but greater diversity in terms of age and nationality. One drawback in the use of online testing is that there is no way of determining whether participants are cheating, and no real way of having any control over their surrounding environment. Participants in online studies have a higher tendency to use a cognitively simpler strategy or just click quickly on options without properly considering them (Krosnick, 1991; Sargis, Skitka, & McKeever, 2013). However, such responses can be checked for by examining response times and excluding participants who appear to be simply clicking through the task. Strategies to prevent such behaviour involve including "attention checking questions" or "instructional manipulation checks", such as an item which asks participants to choose their third item out a list in order to prove they are still paying attention. In fact, studies comparing MTurk and lab participants have shown that, often, MTurk participants actually are more likely to give a correct answer on such questions. Peer, Vosgerau, and Acquisti (2013) found that these methods, whilst reliable, were rendered unnecessary by restricting HITs to workers with high approval ratings on the MTurk website from their previously completed tasks. Restricting workers to include only ones with high approval ratings may lead to a higher proportion of workers who have previously completed other psychology studies, and so may be able to guess the premise of the study and thus show an altered pattern of effects. However, this problem is likely to be much more prevalent with lab based samples, given the high reliance on psychology students. Empirical research shows

that this is not likely to be the case; Sprouse (2011) compared data from 352 participants who took part in a judgement task, half online and the other half in the lab. Other than a higher drop-out rate for online participants, Sprouse found that the two samples were almost identical on a variety of statistical measures. A further argument against the quality of MTurk data is that, although participants may read instructions thoroughly, their attention may waver during the task. This claim is supported by evidence from Chandler and Kapelner (2013), and Clifford and Jerit (2014), who present data from self-reported measures which show that some workers admit to using mobile phones or completing other tasks at the same time. That said, this does not seem to be representative of all workers, given their similarity on various cognitive measures, and results from where attention checking questions were included partway through a study. Although it may be of ethical concern that the rates of compensation are typically much lower than that of lab-based studies, Sargis et al. (2013) highlight how participation is of a voluntary nature, and many workers do not consider financial recompense to be a primary motivator for participating (Buhrmester et al., 2011). It may be tempting to suppose that the level of compensation offered to participants would affect the quality of data collected; this has been found to not be the case with both survey data (Buhrmester et al., 2011) and non-survey data (Paolacci, Chandler, & Stern, 2010). Indeed, participants are typically well-motivated, ranking that they "enjoy doing interesting tasks" and want "to kill time" as the most common motivators for participating, with participating "to make money" being a much lower priority (Buhrmester et al., 2011). Further criticisms of online methodology have been highlighted and addressed by Gosling, Vazire, Srivastava, and John (2004), who conducted a large-scale comparative study with an N of 361703 online participants and compared these with data from 510 existing studies. The overall conclusion that they drew was that many concerns with the use of online testing are unfounded, and those that held true were easily accounted for by altering small aspects of the study design. Finally, (Paolacci et al., 2010) found that participants given a

number of heuristics and biases tasks performed similarly to lab-based participants, and were equally as likely to follow instructions accurately. They advocate the use of MTurk for collecting such data as long as care is paid to the potential pitfalls discussed above. In the online experiments presented in this thesis, other potential ethical issues have been identified and accounted for. Participant data files remained anonymous, and workers verified that they had completed the experiment via a randomly generated code, displayed to them once they had completed the experiment, which they then input on the MTurk website for validation. This allowed the experimenter to verify that the task had been completed, without compromising participant anonymity. The data files were stored on a server in a folder which was inaccessible without a password, and data files were regularly downloaded and then deleted from the server.

Hypotheses

This study aimed to investigate a number of properties of reasoning. The first aim was to compare the use of SDT and traditional analyses in analysing data on individual differences in reasoning. Secondly, I aimed to evaluate the use of online testing for complex reasoning tasks, and compare with data collected in a lab. Thirdly, to investigate whether the use of altered instructions can aid reasoning. Finally, to evaluate whether an individual differences approach leads to more in-depth insight into any changes due to differing instructions. The inclusion of multiple individual difference measures should allow for comparison between them, as well as evaluation of whether it is cognitive style, ability, or both, which contribute towards the differences in reasoning ability and belief bias. It is predicted that the inclusion of individual differences measures will lead to an effect of instruction on belief bias. It is unclear which direction this difference is likely to be in; one the one hand, the extra information contained in the augmented instructions could provide a useful cue for participants with a more analytic cognitive style or higher cognitive ability to engage in increased efforts or better reasoning. On the other hand, the

augmented instructions may lead to an improvement for participants with a lower cognitive ability or less analytic cognitive style, as the extra information may inform them that in order to complete the task an increased level of effort beyond what they might have otherwise put in is required for successful task completion.

5.1 Method

Participants

There were 96 participants, half of whom were students at Lancaster University, and the other half were recruited online via the Amazon Mechanical Turk website. Lab-based participants were paid £3.50 each, and online participants were paid $5 (US dollars).

Design

Instruction set (augmented vs. standard instructions) was manipulated between-participants, and conclusion validity (valid vs. invalid) and conclusion believability (believable vs. unbelievable) were manipulated within-participants.

Materials

Participants were presented with 16 syllogisms. The syllogisms used were from one of the sets from Experiment 1, and thus have identical characteristics, such as EIO or IEO mood, Figures 1 and 2, and an equal number of AC and CA conclusions. These particular moods and figures were chosen in order to prevent mood or figure producing confounding effects. Four subsets were created, as in the previous experiment, so that the content words in each syllogism appeared in each subset as either valid-believable, valid-unbelievable, invalid-believable, or invalid-unbelievable, so that the content was not a confound. A subset of 16 syllogisms was randomly allocated to each participant.

As a measure of cognitive ability, participants completed Part 1 of the AH4 Group Test of General Intelligence (Heim, 1967), which has been used to assess verbal and numeric ability. Measures of cognitive style were the CRT (Frederick, 2005), which measures the tendency to resist an initial Type 1 response; the REI (Epstein, 1994) which measures the self-reported tendency to rely on logic or intuition, and the AOT (Stanovich & West, 1997), which measures the tendency for open-minded thinking.

Procedure

Both the lab-based and online participants viewed the experiment in a web browser Participants in the lab-based group were tested individually with the experimenter present. Participants were shown instructions according to the condition they had been randomly assigned to, which were identical to those used by Heit and Rotello (2014).

Once participants had finished reading the instructions, they then clicked a button to begin the experiment. The first syllogism appeared on screen along with buttons marked "valid" and "invalid", and radio buttons from 1-10 with the question "how sure are you that you have answered correctly?" above them. Once both a validity and confidence rating had been made, the screen paused for 5 seconds, and then the next syllogism was displayed.

To prevent the problem of differences in internet connection speeds leading to differing lengths of time between problems as the webpage loaded, the underlying code was designed so that all problems loaded in the background upon initial page access, and so any delays would only affect the speed at which the responses were saved.

After participants had completed the syllogisms, they went on to complete the CRT, REI, AOT, and then AH4. Participants were given the opportunity to take a break between tasks, but asked to not to pause during the course of individual tasks. The study took approximately 60 minutes in total to complete.

Only MTurk workers with a high approval ratings were accepted, as per the recommendations of Peer et al (2013). In addition to this, the login details to access the website which was hosting the experiment were included at the end of the instructions in order to screen out any participants who did not thoroughly read the instructions.

5.2 Results

Comparing the different samples

Before the data were analysed in-depth, a comparison of the online and lab-based datasets was conducted in order to explore any differences between the two groups and identify outliers.

Syllogisms

The response times for the syllogisms were examined first, with mixed effects models fitted to the data to account for the repeated-measures nature of this component of the experiment. There was no differences in response times between the two groups, $\chi^2(1)=0.21$, p=0.65, with similar response times for the online (M=25.01s) and lab-based samples (M=23.68s).

There was, however, a difference in ranges of times, with the lab-based participants responding within 5s to 130s, and the online participants responding between 1s to 384s. Due to the theoretically predicted difference in response styles, simply identifying outliers using the standard method of eliminating values based on their distance from the mean or interquartile range would not be appropriate here. Instead, participants were eliminated if 4 or more of their response were given within 5 seconds or less. An upper bound was considered in order to identify participants who were completing the task whilst doing something else. However, no participants consistently showed unusually long response times. In addition, data were removed from participants who responded with either "valid" to all

syllogisms, or "invalid" to all syllogisms. This method of identifying problematic responses led to the removal of data from 2 lab-based participants, and 9 online participants.

Cognitive style and ability measures

Linear models were fitted to the cognitive style and ability measures to assess differences between the online and lab-based participants. There were no differences in AH4 scores between the two groups, $t(81) = 1.31$, $p=0.19$, and no differences in AOT scores either, $t(81) = 1.11$, $p=0.27$. The REI is divided into four subscales. There were no differences between participants on the Rational Ability ($t(81)=1.04$, $p=0.30$), Rational Engagement ($t(81)=0.64$, $p=0.52$), Experiential Ability ($t(81)=-0.56$, $p=0.58$), or Experiential Engagement ($t(81)=-1.68$, $p=0.10$) scales.

As the scores on the CRT ranged from 0 to 3, it would be inappropriate to fit a linear model to such data, and instead it was transformed into proportion correct, and a beta model was instead fitted. A beta model is an extension of the generalised linear model which allows for response variables that range from 0 to 1. Due to problems fitting the model to data containing proportions of 0, such scores were transformed to values of 0.000000001, so the model could still be fitted, with the transformation having negligible effects on the results. There was a significant difference between the online and lab-based participants, with the online participants showing higher CRT scores ($M=2.35$) than the lab participants ($M=1.33$), $z=5.85$, $p<0.001$.

Correspondence between style and ability measures

Table 5.2 shows the correlations between the different cognitive ability and style measures. As expected, there were correlation between the AH4, an ability measure, and the majority of the cognitive style measures.

	AH4	AOT	REI(RA)	REI(RE)	REI(EA)	REI(EE)
AH4						
AOT	0.42***					
REI (RA)	0.03	-0.08				
REI (RE)	0.16	0.25*	0.64***			
REI (EA)	-0.33**	-0.40***	-0.29**	-0.28*		
REI (EE)	-0.38***	-0.26*	-0.30**	-0.19	0.81***	
CRT	0.31**	0.33**	-0.02	0.02	-0.28*	-0.18

Table 5.1: Correlations between measures *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

Endorsement Rates

Binary logistic mixed effects models were fitted to the endorsement rates in order to examine what inferences might be made from traditional, non-SDT methods of analysing the data. Initially models with just the main effects of validity, believability, instruction set and participant group were fit to the data, without the cognitive ability and style measures. Next, the effects of the cognitive ability and style measures were examined, along with interactions with any of the initial variables which were significant, in order to explore whether the inclusion of the ability and style measures modified any existing effects.

A binary logistic mixed effects model was fitted to the data, with conclusion endorsement as the response variable, and instruction set, participant group, validity and believability as explanatory variables.

Instruction set and participant group were not significant as main effects, and did not interact with any other variables. There was a main effect of validity, $\chi^2(1)$=152.60,p<0.001, with more valid problems being endorsed than invalid problems. There was also a main effect of believability, $\chi^2(1)$=41.63, p<0.001, with more believable problems than unbelievable problems accepted.

Validity and believability also interacted, $\chi^2(1)$=13.84, p<0.001, β=0.93. For valid problems, believability had no effect on problem acceptance, $\chi^2(1)$=2.78, p=0.10. However, invalid conclusions were more likely to be accepted if they were believable, $\chi^2(1)$=61.96, p<0.001.

Table 5.2: Parameter Values for Endorsement Rate Model

Variable	β	SE	z	Wald-p
Validity (valid)	1.027	0.177	5.800	< 0.001
Believability (believable)	1.243	0.165	7.542	< 0.001
Validity x believability	0.931	0.249	3.742	< 0.001

SDT Measures - Accuracy

Models using the SDT measures as response variables were examined next. A mixed effect model was fitted with accuracy as the response variable, and instruction set (standard vs. augmented), participant group (online vs. lab), problem believability (believable vs. unbelievable), CRT score, REI subscale scores, AOT score, and AH4 score as explanatory variables.

Due to the large number of explanatory variables, the inclusion of terms and their interactions in the model was based on the following approach. The main effects and interactions of the variables concerning the experimental effects (instruction set, participant group, and believability) were examined first. Next, the individual difference measures (CRT score, REI subscales score, AOT score and AH4 score) were examined as main effects and only tested as interactions where the experimental effects or their interactions were significant.

Experimental Effects None of the main effects significantly predicted accuracy. However, there was a marginally significant interaction between believability and participant group, $\chi^2(3)$=7.45, p=0.06. For the lab participants, believability did not affect their accuracy, $\chi^2(1)$=0.04, p=0.84. However, for the online participants, believability was significantly related to accuracy, $\chi^2(1)$=7.00, p=0.008, with participants in this group showing more accurate performance on unbelievable syllogisms.

Individual Difference Measures A higher AH4 score led to higher accuracy, $\chi^2(1) = 6.288, p = 0.01$. In addition, a higher score on the REI Experiential Subscale led to lower accuracy, $\chi^2(1) = 4.346, p = 0.037$. There was a three-

way interaction between participant group, believability and AOT score. For the online participants, the believability by AOT score interaction was not significant, $\chi^2(1) = 0.880, p = 0.348$. However this interaction was significant for the lab participants, $\chi^2(1) = 9.499, p = 0.002$, for whom a higher AOT score did not predict accuracy on believable problems, $t = -0.055, p = 0.956$, but was linked to higher accuracy on unbelievable problems, $t = 3.924, p < 0.001$.

SDT Measures - Response Bias

Believability was a significant predictor of response bias, $\chi^2(1)$=17.62, p<0.001, β=-0.25, SE=0.06, t=-4.29, with participants showing a more liberal response bias to believable problems. No other main effects or their interactions were significant.

None of the individual differences measures or their interactions with other variables affected response bias.

Figure 5.1 shows the ROC curves for the online and lab participants.

Response Times

A mixed effect model was fitted to the data with the log-transformed response times as the response variable. As with the SDT analyses, firstly the experimental effects (validity, believability, instruction set, and participant group) and their interactions were examined as potential explanatory variables, followed by the

Table 5.3: Parameter Values for SDT Accuracy Model

Variable	β	SE	t	LRT-p
Believability (believable)	0.932	0.314	2.968	0.083
Participant Group (online)	0.667	0.294	2.272	0.142
AH4 Score	0.007	0.002	3.025	0.001
REI Experiential Score	-0.022	0.010	-2.196	0.037
AOT Score	0.004	0.001	0.393	0.685
Believability x Participant Group	-0.121	0.056	-2.147	0.059
Believability x AOT score	-0.005	0.002	-2.962	0.301
Participant Group x AOT Score	-0.004	0.002	-2.250	0.307
Believability x Participant Group x AOT score	0.004	0.002	1.700	0.015

Table 5.4: Parameter Values for SDT Response Bias Model

Variable	β	SE	t	LRT-p
Believability (believable)	-0.258	0.060	-4.286	< 0.001

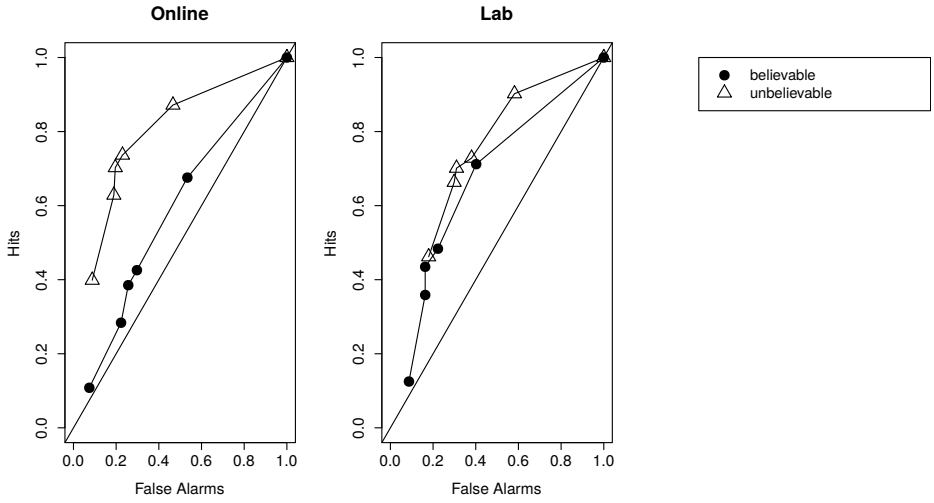


Figure 5.1: ROC curves for online and lab participants by believability

individual difference measures (CRT score, REI subscale scores, AOT score, and AH4 score) and any interactions with the experimental effects.

Experimental Effects Participants responded faster to valid than invalid problems problems, $\chi^2(1) = 10.461, p = 0.001$. Response times were slower for believable problems than they were for unbelievable problems, $\chi^2(1) = 13.705, p < 0.001$. Participants who completed the experiment online responded faster than those who completed it in the lab, $\chi^2(1) = 5.585, p = 0.018$.

Individual Difference Measures A higher REI Experiential subscale score was related to faster response times,$\chi^2(1) = 8.520, p = 0.003$. There was an interaction between CRT score and participant group, $\chi^2(2) = 15.869, p < 0.001$. For the online participants, a higher CRT score was related to a faster response time, $\chi^2(1) = 14.150, p < 0.001$. However for the lab participants, a higher CRT score was marginally related to slower response time, $\chi^2(1) = 3.415, p = 0.064$. AOT score and participant group also interacted, $\chi^2(2) = 8.769, p = 0.012$. For the

online participants, there was no relationship between AOT score and response time, $\chi^2(1) = 2.335, p = 0.127$. However, for the lab participants, AOT score was related to longer response times, $\chi^2(1) = 6.318, p = 0.012$. Believability interacted with AH4 score, $\chi^2(1) = 9.187, p = 0.010$. For unbelievable problems, AH4 score did not predict response times, t(654)=1.461, p=0.144. However, higher AH4 scores were linked to longer response times for believable problems, t(654)=3.719, $p < 0.001$.

Table 5.5: Parameter Values for Response Time Model

Variable	β	SE	t	LRT-p
Validity (valid)	-0.323	0.123	-2.620	0.001
Believability (believable)	0.486	0.236	2.056	< 0.001
Participant Group (online)	1.894	0.574	3.298	0.018
REI Experiential Score	-0.102	0.029	-3.514	0.003
CRT Score	0.041	0.047	0.869	0.064
AOT Score	0.005	0.002	2.110	0.623
AH4 Score	0.016	0.007	2.347	0.056
Participant Group x CRT Score	-0.277	0.085	-3.277	< 0.001
Participant Group x AOT Score	-0.009	0.003	-2.595	0.012
Believability x AH4 score	-0.011	0.004	-2.499	0.010
Validity x REI Experiential Score	0.037	0.019	1.954	0.028

Confidence Ratings

Finally, confidence ratings were examined as response variables, with validity, believability, group, and the individual difference measures as explanatory variables.

Experimental Effects Participants who were given the standard instructions were more confident than those who were given the augmented instructions, z=2.217, p=0.027. Responses given to unbelievable problems were less confident than those given to believable problems, z=-2.064, p=0.039. Validity and believability interacted, z=-3.613, p< 0.001. For valid problems, participants were equally confident, regardless of believability, $z = -0.284, p = 0.777$. However, for invalid problems, participants were less confident if the problem was believable, $z =$

$-3.813, p < 0.001$.

There was an interaction between validity, believability, and participant group. For the online participants, the validity by believability interaction was not significant, $z = -0.160, p = 0.873$. However, this interaction was significant for the lab participants, $z = -3.618, p < 0.001$. On valid problems, there was a marginally significant effect of believability, with participants responding more confidently to believable problems, $z = -1.958, p = 0.050$. However, for invalid problems, participants responded more confidently to unbelievable problems, $z = 3.320, p < 0.001$.

Individual Difference Measures Participants with a higher AH4 score were less confident in their responses, z=-2.874, p=0.004. AH4 score interacted with believability; on unbelievable problems, AH4 score did not predict confidence. However, for believable problems, a higher AH4 score led to lower confidence, $z = 2.018, p = 0.043$. AH4 score also interacted with validity. On valid problems, AH4 score did not predict confidence, $z = 0.146, p = 0.884$; however, for invalid problems, a higher AH4 score was related to lower confidence, $z = -1.989, p = 0.047$.

Table 5.6: Parameter Values for Confidence Model

Variable	β	SE	z	Wald-p
Validity (valid)	-1.375	1.050	-1.310	0.019
Believability (believable)	-2.169	1.051	-2.064	0.039
Instruction Set (standard)	0.570	0.257	2.217	0.027
Participant Group (online)	0.340	0.353	0.962	0.552
AH4 Score	-0.064	0.025	-2.512	0.004
Validity x Believability	-1.147	0.317	-3.618	< 0.001
Validity x Participant Group	-0.766	0.333	-2.303	0.021
Believability x Participant Group	-0.302	0.325	-0.930	0.352
Believability x AH4 Score	0.053	0.0191	2.776	0.005
Validity x AH4 Score	0.048	0.019	2.562	0.010
Validity x Believability x Participant Group	1.103	0.469	2.353	0.019

5.3 Discussion

The use of instructions emphasising the importance of the rules of logical necessity do not lead to an improvement in reasoning performance, even when individual differences between reasoners are accounted for. There were no effects of instructions on accuracy, response bias, or response time, indicating that the instructions did not alter reasoning strategy. There was an effect of instruction set on confidence ratings, with participants who had received the augmented instructions showing lower overall confidence, implying that the expanded description did not aid participants in reasoning but only served to confuse them. These results agree in part with those of Heit and Rotello (2014) who found that instructional manipulations did not lead to a change in reasoning bias (although Heit and Rotello found no reasoning bias at all). However, these results diverge in that, unlike Heit and Rotello (2014), no effect of instructions on response bias was found. These differences may have been due to differences between the samples in terms of general levels of response bias, or due to the fact that the marginal effect of response bias found by Heit and Rotello (2014) may have been a statistical artefact. None of the individual difference measures interacted with instruction set either, showing that the hypothesis that these would mediate the effect of individual differences was not supported. This experiment also aimed to compare the use of SDT and traditional indices as well as examining any differences between groups of participants. The analyses presented here show that SDT analyses are once again crucial in distinguishing between reasoning bias and response bias. A troubling finding is that the endorsement rate analysis was unable to distinguish between the online and lab participants in terms of belief bias. Whilst the endorsement rate analysis showed a general effect of belief bias for all participants, the SDT analysis revealed that both groups showed response bias, the online participants all showed reasoning bias, but only the lab participants with a higher AOT score showed any signs of reasoning bias. Reliance on endorsement rates would have led to treating both groups as identical, and been unable to detect the similarities and differences between these groups in terms of

reasoning and response bias. Psychological studies have been historically criticised for tending to rely on psychology students as participants. The present study clearly demonstrates that this can be highly problematic for making inferences from experimental studies. The lab participants, a group comprising mainly psychology students, were influenced by belief bias solely as a response bias. However, the online participants, presumably made up of a much more varied group in terms of age and educational background, showed belief bias as both a response bias and a reasoning bias, results which correspond with previous studies. A limitation of the present study is that no data were collected regarding the age or educational background of participants, and so these factors could not be accounted for in the analysis. Although it may be tempting to ascribe this difference in reasoning bias to these factors, or to cognitive ability or style differences between the two groups, the individual differences measures show that this is unlikely to be the case, with both groups showing identical performance on all but one measure (CRT). The similarity between the two samples in terms of the individual differences measures further supports the validity of using online testing in psychology studies of belief bias. There was no difference between the online and lab participants in their mean AOT, REI, and AH4 scores. A subsequent Kolmogorov-Smirnov test conducted to examine if there were any differences in terms of variability between the two groups showed no differences for AOT scores, AH4 scores, and REI Experientiality subscale score (ps all >0.09), but differences in the CRT score ($p<0.01$) and the REI Rationality score ($p=0.04$). Although the online participants scored significantly higher on the CRT, this may have been due to either cheating or familiarity; task familiarity is an increasingly common problem found with the CRT as discussed by Toplak, West, and Keith (2013). The link between CRT score and faster response times for the syllogisms for the online participants could support either of these explanations with participants with a higher CRT score showing faster responses being linked to either the tendency to engage in less effort or simply familiarity with these kinds of tasks. The differences in REI Rationality score may account

for the differences in amount of reasoning bias between the two groups. Greater variation in the tendency to rely on logic may have led to more visible differences due to a higher number of individuals with a tendency to rely more on logic and higher number of those less likely to rely upon logic. The quality of the online data was lower than that of the lab participants; 9 out of 48 responses collected online were discarded because of participants simply clicking "valid" for all syllogisms, or responding within a time insufficient for the problem to have even been read. However, such responses were easily and automatically identified, and this led to 2 of the lab participants' data being discarded a level of data checking which may not have taken place if only using data from lab participants. In addition, the lack of differences between the two groups in terms of response times or accuracy supports the argument in favour of the quality of the MTurk data, consistent with arguments outlined by Hauser and Schwarz (2015).

The lab participants only showed signs of belief bias as a response bias, whereas the online participants showed both reasoning bias and response bias. This may also be explained by levels of engagement; belief bias as a reasoning bias is found mainly in circumstances under which sophisticated reasoning is likely to take place (Trippas et al, 2013). Individuals who participate in experiments on Amazon MTurk have been found to be motivated to participate for reasons other than money, they are after all participating in experiments for low compensation rates and may thus have higher levels of interest in the study (Buhrmester et al., 2011), and so may have simply been more engaged in the study than the lab participants. Cognitive ability, as measured via the AH4 test was predictive of general accuracy. The AH4 test measures cognitive ability. In the context of syllogistic reasoning research, this is thought to relate to the number of models that an individual can hold in mind at any one time, and thus affects their ability to compare different models of the premises. Therefore, a general link with accuracy is unsurprising participants with a higher AH4 score were presumably better able to model a larger range of possibilities and so better discriminate between valid and invalid

syllogisms. Regardless, AH4 score did not affect the amount of reasoning bias shown, which would have been indicated by an AH4 by believability interaction. In other words, higher ability participants are better at reasoning generally, but equally as biased as lower ability participants. The same claim has been made by Trippas, Pennycook, et al. (2015) who also found that cognitive ability affected overall accuracy but not bias.

There was, however, a three-way interaction between believability, group and AOT score. Despite the lack of effect of belief on accuracy for the lab participants, there was still an effect on unbelievable problems of AOT, with those with a tendency towards more open-minded thinking being more accurate. In other words, for the lab participants, only those with a more open-minded thinking style showed an advantage for unbelievable problems. AH4 score did not interact with believability in the accuracy analysis. However, it did interact in the response time analysis. Participants with a higher AH4 score spent longer examining believable conclusions; that is, the effect of believability of increasing response times for believable problems is even stronger for participants of a higher cognitive ability. However, these longer response times did not lead to an increase in accuracy. This supports the view suggested by Trippas, Pennycook, et al. (2015) and Stanovich and West (2008) that an analytic cognitive style is a crucial component of susceptibility to belief bias; if one is able to engage in better reasoning, this advantage is wasted if one does not have the tendency to do so. The results presented here can also be useful in arbitrating between different individual difference measures. Newstead, Handley, Harley, Wright, and Farrelly (2004) found that self-report measure of cognitive style were not predictive of general accuracy, but were linked to levels of bias shown and claimed that this was due to the ability to generate multiple representations of a problem. The results presented here differ in that they do show a link between self-reported cognitive style and overall accuracy lower scores on the REI Experiential scale were related to higher accuracy. Given that Newstead et al. (2004) did not employ SDT analyses, this may account for the difference between

those results and results in the present study. Trippas, Pennycook, et al. (2015) discuss the fact that analytic cognitive style can be measured using both self-report scales (such as AOT and REI) or performance measures (such as CRT). Although Trippas, Pennycook, et al. (2015) advocate the use of performance measures, these results show the opposite; CRT score was not significant in any of the analyses after variables were entered in the order of most variance explained to least, and only became significant in the response time analysis, possibly due to task familiarity for the online group. Both of the self-report measures (AOT and REI) were significant in analyses, indicating that they both provide a useful contribution to explaining reasoning accuracy and belief bias despite measuring similar constructs. Differences between the predictive ability of AOT and REI have been found in studies in other domains, for example, in factors which affect pathological gambling (Maclaren, Fugelsang, Harrigan, & Dixon, 2012). To conclude, instructions were ineffective in debiasing participants. However, the results did show some interesting findings regarding the use of online testing; it appears that online participants show higher levels of engagement than lab participants. The feedback manipulation used in Experiment 1 and the instructional manipulation in Experiment 2 constituted fairly simple attempts at debiasing, which were unsuccessful. In Chapter 6, the final empirical chapter of this thesis, I present an experiment in which the use of training and feedback are combined to evaluate whether a more in-depth approach to debiasing will be successful. It is clear that individual differences play an important part in explaining belief bias, and so these measures will also be examined in the next experiment.

Chapter 6

Experiment 3

The previous experiment demonstrated that even when accounting for individual differences in cognitive style and ability, instructions emphasising the importance of logical necessity were not sufficient to reduce the effects of belief bias on syllogistic reasoning. A possible explanation for lack of debiasing effect for both the feedback manipulation discussed in Chapter 4, and the instructional manipulation discussed in Chapter 5, is that neither of these interventions clearly provided participants with an obvious strategy for overcoming their belief bias. In Experiment 1, feedback seemed to alert participants to the fact that their strategy was suboptimal, as shown by the change in response bias and longer response times. However, no accompanying change in accuracy was observed. Similar results were found in Experiment 2; the augmented instructions led to a decrease in confidence but no other effects. It seems that instructing individual to engage in additional effort is not sufficient to lead to a reduction in bias. If participants are only constructing partial representations, the belief bias effect may be affecting these initial models. Thus, for a debiasing intervention to be effective, it may be necessary to provide a method which reasoners can use to exhaustively represent all possibilities. Previous research by Prowse-Turner and Thompson (2009) has found that training participants in constructing diagrams to represent syllogisms has led to improved accuracy; however, this research did not use SDT analyses or belief-based materials so it

is unclear whether this could be used to achieve a debiasing effect. This chapter reports an experiment in which participants are explicitly provided with a way of graphically representing the syllogistic reasoning problems, as well as with feedback, with the aim that this may lead to an altered strategy and the reduction of belief bias as both a reasoning bias and as a response bias.

Training

Training has been shown to improve reasoning across a range of tasks and biases including Wason's selection task (Cheng et al., 1986), conjunction bias (Moutier & Houdé, 2003), matching bias (Moutier & Angeard, 2010), base-rate problems (Fong et al., 1986) and, more specifically, syllogistic reasoning (Leighton, 2006; Prowse-Turner & Thompson, 2009). However, none of these studies directly address the question of whether training in alternative reasoning strategies can lead to a reduction in belief bias as they neither manipulate believability or use SDT analyses. Luo et al (2014) conducted a study in which fMRI was used to examine the effects of logical training on neural activation whilst participants completed conditional reasoning tasks including a believability manipulation. They found that this intervention led to a change in neural activity, and so concluded that this was due to changes related to cognitive control, that is, participants were suppressing belief-generated responses in favour of logic-derived answers. However, the model fitted to this data was not able to take response bias into account as it used traditional indices rather than SDT measures, and so it is unclear whether the changes reported here were the results of a change in reasoning or simply a change in response bias. It has been argued that decontextualisation is important for reasoning (e.g. Baron, 1995; Toplak & Stanovich, 2003; Kuhn & Udell, 2007) and Macpherson and Stanovich (2007) found that instructions which told participants to reason in a more abstract manner led to small accuracy improvement in a syllogistic reasoning task. Being able to separate beliefs from the reasoning process may be more easily achieved through the use of diagrams; if participants use the diagrams

to map out all the possible models, they should be uninfluenced by their prior beliefs. Regardless of the way in which belief bias affects reasoning, exhaustively representing all possibilities should lead to normative responding, as predicted in the extended selective processing model of Stupple et al. (2011). It has also been suggested that there might be different sub-categories of reasoners. For example, Ford (1995) asked participants to give answers to a number of syllogisms and explain how they came to their conclusion. Although it should be noted that this is a conclusion generation rather than evaluation task, and so different strategies may have been used, Ford found that reasoners could be roughly divided into two types. Firstly, there were spatial reasoners who drew diagrams to map out links between the terms in the premises. Secondly were the verbal reasoners who typically discussed the problems in a more algebraic way. Although these differences may point towards further individual differences beyond cognitive ability and style being important to consider, the findings of Prowse-Turner and Thompson (2009) that participants were able to improve their reasoning with the use of diagrams suggests that simply equipping reasoners with an alternative strategy to their existing one may be sufficient to improve reasoning for most individuals.

The content of any training manipulation is also important to consider. An understanding of the concept of logical necessity is crucial for reasoners to be able to correctly evaluate the validity of categorical syllogisms (Evans, Jonathan, Handley, Harper, & Johnson-Laird, 1999; Prowse-Turner & Thompson, 2009; Torrens, Thompson, & Cramer, 1999).Although there is a great deal of evidence to suggest that the misinterpreted necessity account of belief bias (Evans et al., 1983) does not accurately encapsulate this phenomenon; for example Klauer et al (2000) highlight how misinterpreted necessity can explain the further difficulties individuals find with indeterminately invalid syllogisms, but cannot explain how people still find some determinately invalid syllogisms difficult to evaluate. Despite this, there is some evidence to suggest that it may well still play some role in participants' performance on syllogistic reasoning tasks. For example, Prowse-Turner and Thompson (2009)

used a methodology in which, rather than rating syllogisms as "valid" or "invalid", participants had to indicate whether they thought they were "necessary", "possible", or "impossible". These categorisations were then compared to the logical state of the syllogism, which were divided into "necessary", "possible strong", "possible weak", and "impossible". The terms "necessary" and "impossible" refer to determinately valid and determinately invalid problems, respectively. "Possible strong" and "possible weak" are derived from results in a study by Evans et al (1999), with all such syllogisms containing conclusions which are possible but not necessitated by the premises (i.e. indeterminately invalid syllogisms). The strong/weak distinction is based on the number of participants who rated that conclusion as "possible" in Evans et al. (1999). Prowse-Turner and Thompson (2009) trained participants by showing them how syllogisms could be represented as Venn diagrams, and giving a verbal explanation of how logical necessity in the context of an individual problem. Participants in the training condition were trained on four problems in total as a group, given feedback and training on individual problems, and allowed to ask questions at any point. All participants went on to complete a further 16 syllogisms individually. Prowse-Turner and colleagues found that, with training, participants became much more competent at correctly categorising syllogisms, and their confidence in their own responses was better calibrated than those who were given no training.

As mentioned earlier, research by Ford (1995) shows that there is variation between strategies which people employ to reason about syllogisms, and that it is possible to roughly classify reasoners as either spatial or verbal, depending on strategy use. Thus, a similar methodology to that of Thompson et al (2009), which involved a combination of both verbal and visual descriptions will be used here in order to account for the fact that some participants are more or less likely to engage in verbal or visual strategies. Combining training with logical instructions is not only important because of the range of strategies used by reasoners. Heijltjes, Gog, and Paas (2014) provided participants with training on a number of tasks, including

a syllogistic reasoning and a base-rate task and assessed the impact of instructions, training, and practice on performance in a later reasoning task (involving passages of prose text). Heijltjes et al. (2014) found that it was the combination of both specific instructions with an opportunity to test out new strategies which led to a real improvement in participants' performance on tasks which required critical thinking. Therefore, the main component of the task in the present experiment will be prefaced by a section in which participants solve practice syllogisms. Providing feedback alongside training may provide a key advantage in improving performance, which training alone cannot deliver. Previous research has shown that whilst training alone did not improve performance on the Wason selection task (Cheng et al, 1986), Leighton (2006) found that giving participants feedback led to an increase in accuracy, presumably because the corrective feedback caused individuals to engage in greater reflection about the strategies which they were using. Thus, the effects of feedback during the practice phase of the experiment will be examined.

Predictions

It is predicted that training will improve overall accuracy, similarly to Prowse-Turner & Thompson (2009), and may reduce belief bias. The aforementioned study did not employ a believability manipulation, and so it is not immediately clear whether the beneficial effects of training are likely to extend to effectively reducing belief bias. However, given that the training procedure involves teaching participants ways of visually representing syllogisms, it seems feasible that those who fully engage with this technique may show a reduction in bias as it could result in participants exhaustively representing all possible conclusions.

It is predicted that participants who receive both training and feedback will show the most accurate performance on the syllogisms. The accuracy of those who receive training alone, will vary based upon how successful they are for the practice syllogisms.

Those who are given feedback with no training may show lower levels of

response bias, as was the case for participants in the feedback group in Experiment 1. However, it is possible that this effect may not be present, due to the small number of training syllogisms.

It is also hypothesised that the effects of training will be moderated by individual differences. The training will most benefit participants with pre-existing tendencies to rely on prior knowledge rather than logic, that is, those with higher scores on the REI Experiential subscale.

Due to the differences in the previous experiment between the lab and online participants, in this study, all data was collected from online participants. In order to get a greater insight into the specific characteristics of this sample and how they compare to a typical lab sample, a range of questions were asked concerning things such as age, gender, and highest level of education completed.

6.1 Method

Participants

In total, 205 participants were recruited online using the Amazon Mechanical Turk website. No participants had completed training in formal logic. Participants were aged between 18 and 62, with a mean age of 35.65 years. There were 104 (55%) female participants, and 86 (45%) male participants. Highest educational level completed was divided into 6 strata; 1% of participants did not graduate high school, 12% were high school graduates, 31% had completed some college or university, 11% had an associate's degree or foundation degree, 38% had a bachelor's degree, 7% had a postgraduate qualification other than a doctorate, and 0% had a doctorate.

The minimum approval rating was identical to that of the previous experiment. In addition, workers who had completed the previous experiment were blocked from participating in this one.

Materials

The syllogisms used were all three-model syllogisms with premises in moods AE, EA, EI, and IE. A similar classification system to that used by Prowse-Turner & Thompson (2009) was used to categorise the possible strong" and "possible weak" syllogisms. This was based upon data from Evans (1999) in which participants were asked to provide judgements of whether syllogisms were necessary, possible or impossible. The "possible strong" syllogisms used in the present study were those that followed a structure that was logically possible but not necessitated by the premises that had been deemed "possible" by 67

A total of 32 syllogisms were shown to each participants; 8 each of necessary, possible strong, possible weak, and impossible. For each logical validity category (i.e. necessary, possible strong, possible weak, impossible), the premise moods, conclusion orders, and conclusion moods were balanced so that they were not confounded with believability. Necessary and possible strong conclusions were in E or O moods, and possible weak and impossible conclusions were in A and I moods; again, these were split evenly so as to not act as a confound.

As in the previous experiment, the AOT and REI were used as measure of cognitive style, and AH4 as a measure of cognitive ability.

Procedure

The following instructions were shown to participants:

> This experiment is designed to find how people solve logical problems. Your task is to decide whether each conclusion follows logically from the information given in that problem. The premises - the information given - appear above the line and the conclusion appears below the line. You must assume that all the information you are given in the premises is true; this is very important. If, and only if, you judge that a given conclusion logically follows from the information given you

should answer 'Valid'. If you think that the given conclusion does not necessarily follow from the information given you should answer 'Invalid'. Also, you will be asked how confident you are in this judgement. A higher number should be used to rate higher confidence.

In addition, you will be asked to judge whether the conclusion is 'necessary', 'possible', or 'impossible'. If you judge that a given conclusion follows logically from the information given, you should choose 'necessary'. If you believe that the following conclusion is merely possible but not necessitated by the premises given, you should choose 'possible'. If you think that the conclusion given does not follow logically from the information given, you should choose 'impossible'.

Once you have given a response, your answer will be highlighted in bold and for the practice problems only, you will be given feedback on your answers. When you are ready click "next" to progress to the next screen.

Please take your time and be certain that you have the logically correct answer.

All participants completed a single practice syllogism. Those in the 'training' condition then completed a further four syllogisms for which they provided an answer, and were shown a comprehensive description containing Venn diagrams showing why their response was correct or incorrect (see Appendix D for more details). Those in the 'no training' condition completed these syllogisms as further practice syllogisms, with no explanation of possible reasoning strategies. In addition, feedback was given to half of the participants in each condition. For those who were given feedback, the word "correct" or "incorrect" was displayed on screen, with the correct response highlighted.

All of these initial syllogisms contained belief neutral conclusions. All participants then proceeded to answer the main set of 32 syllogisms, for which no feedback or additional explanation was provided.

Each syllogism was displayed individually on screen and participants clicked on one of two buttons to indicate whether they thought the problem was valid or invalid. Underneath these buttons were a series of radio buttons where participants selected how confident they were in their answer on a scale from 1 to 3. Finally, participants clicked on another button to indicate whether they thought the conclusion was 'necessary', 'possible', or 'impossible'.

There was a five second pause between each syllogism, during which time the screen went blank. Following the syllogistic reasoning part of the experiment, participants were given the opportunity to take a break. They then completed the REI, AOT and AH4.

6.2 Results

Data Cleaning

As in the previous experiment, the data were screened for unusually long or short response times. Response times ranged from 1 second to 3482 seconds. Once again, all data from a participant were removed if more than a quarter of responses from a given participant were made within 4 seconds or less. This led to 10 participants' responses being removed. An upper limit was again, considered, but no participants showed consistently exceptionally long response times to multiple problems; all of the 5 participants with a response time greater than 180 seconds only had an exceptionally long response time for a single syllogism. This number of participants to be removed was lower than expected; in the previous experiment 19% of participants' data had to be excluded, where in the present experiment this constituted only 5% of data. However, system logs showed that 234 individuals began the experiment but only 200 completed it. After the data were cleaned, a total of 190 participants remained.

Endorsement Rates

A binary logistic mixed-effects model was fitted to participants' 'valid'/'invalid' judgements as a response variable, with feedback, training, validity, and believability as explanatory variables. As this analysis was conducted primarily to compare any evidence for belief bias found here with that found in the SDT analysis, individual differences variables were not included.

There were no main effects of feedback or training. However, participants' chance of responding 'valid' was contingent on the syllogism's logical validity, $\chi^2(3)$=2320.60, $p < 0.001$. Necessary problems were most likely to be accepted as valid (M=0.76), more so than possible strong(M=0.58), followed by possible weak (M=0.16), followed by impossible problems (M=0.08). The paired comparison differences between all four types of problems were statistically significant ($ps <$ 0.001). There was a main effect of believability, $\chi^2(1)$=144.9, $p < 0.001$, with more participants accepting believable than unbelievable problems as valid.

There was an interaction between training and validity. For the possible strong and impossible problems, participants in either condition were equally likely to accept these problems as valid. For necessary problems, participants who had been trained were less likely to accept these problems as valid, $\chi^2(1) = 10.16, p = 0.001$. For the possible weak problems, trained participants were more likely to accept these as valid, $\chi^2(1) = 5.68, p = 0.02$.

An interaction also existed between validity and belief, $\chi^2(3) = 8.66, p = 0.03$. For believable problems, the effect of validity was the same as that for the main effect of validity ($ps < 0.001$). Similarly, for unbelievable problems, necessary problems were more likely to be accepted as "valid" than possible strong, in turn more likely than possible weak ($ps < 0.001$). However, possible weak and impossible problems were equally likely to be deemed to be valid, $p = 0.055$.

There were no other interactions between variables in this analysis.

SDT Analyses

SDT analyses presuppose a pair of distributions lying upon a single axis. These pair of distributions represent signal and noise distributions, or in the case of syllogistic reasoning research, the distributions for valid and invalid stimuli. Although it may be tempting to treat the categories of 'necessary', 'possible', and 'impossible' problems as three distributions on the same axis, differing in terms of argument strength, this raises further complications for confidence rating analyses; although valid/invalid problems and confidence ratings from 1 to 3 can simply be converted to a 6 point scale, it seems unclear how one would do this with the present stimuli. A single scale approach would be inappropriate here, as it does not seem reasonable theoretically to presume that a high confidence response of "impossible" lies next to a low confidence response of "possible" on such a scale. Thus, in order to acquire results which are theoretically interpretable, the SDT analyses are divided into 3 subsections; 1 for each of the pairings of necessary/possible strong stimuli, necessary/possible weak stimuli, and necessary/impossible stimuli. These pairings are possible due to the high numbers of stimuli per participant.

Necessary vs. Possible Strong

Table 6.1: Parameter Values for Necessary vs. Possible Strong - Accuracy Model

Variable	β	SE	t	LRT-p
AH4	0.010	0.002	4.542	< 0.001
REI Rationality Subscale	0.012	0.005	2.227	0.038
Training (present)	-0.037	0.019	-1.922	0.055
Believability (believable)	0.309	0.147	2.098	0.060
AH4 x Believability	-0.006	0.003	-2.315	0.022

Accuracy There was a marginally significant main effect of training on accuracy, $\chi^2(1) = 3.74, p = 0.053$, although not in the direction expected; participants who had received training were less accurate than those who had not. Believability was also marginally significant, $\chi^2(1) = 3.53, p = 0.060$, with lower accuracy on believable compared to unbelievable problems.

In terms of individual difference measures, AH4 score was significant, $\chi^2(1) = 16.80, p < 0.001$, with higher AH4 score linked to higher accuracy. The REI Rational subscale was significant, $\chi^2(1) = 4.281, p = 0.039$, with higher scores linked to higher accuracy.

There was a two-way interaction between believability and AH4 score, $\chi^2(1) = 5.283, p = 0.022$. For believable problems, there was no relationship between AH4 score and accuracy. However, participants with higher AH4 scores were more accurate than those with lower AH4 score on unbelievable problems, $t(188) = 4.746, p < 0.001$.

Table 6.2: Parameter Values for Necessary vs. Possible Strong - Response Bias Model

Variable	β	SE	t	LRT-p
Believability (believable)	-0.414	0.047	-8.752	< 0.001
Training (present)	0.113	0.058	1.945	0.052

Response Bias Participants who were trained showed a trend towards a more conservative response bias, $\chi^2(1) = 3.787, p = 0.052$, being less willing to respond with 'valid' to most syllogisms. Participants showed a more liberal response bias for believable problems, $\chi^2(1) = 64.65, p < 0.001$.

Necessary vs. Possible Weak

Table 6.3: Parameter Values for Necessary vs. Possible Weak - Accuracy Model

Variable	β	SE	t	LRT-p
AH4	0.008	0.002	5.025	< 0.001
Training (present)	-0.056	0.018	-3.141	0.002
REI Rationality Subscale	0.010	0.005	1.864	0.064

Accuracy Trained participants were less able than untrained participants to accurately distinguish between necessary and possible weak problems, $\chi^2(1) = 9.559, p = 0.002$. A higher AH4 score was linked to higher accuracy, $\chi^2(1) =$

$25.512, p < 0.001$. Higher scores on the REI Rational subscale were marginally significantly related with increased accuracy, $\chi^2(1) = 3.442, p = 0.064$.

Table 6.4: Parameter Values for Necessary vs. Possible Weak - Response Bias Model

Variable	β	SE	t	LRT-p
Believability (believable)	-1.944	0.385	-5.042	< 0.001
Training (present)	0.152	0.062	2.444	0.246
AH4	-0.009	0.005	-1.746	0.126
Believability * Training	-0.182	0.082	-2.224	0.027
Believability * AH4	0.031	0.006	4.521	< 0.001

Response Bias Participants had a more liberal response bias for believable problem, $\chi^2(1) = 45.044, p < 0.001$. Training interacted with believability; $\chi^2(2) = 7.186, p = 0.028$. On believable problems, training had no effect on response bias. However, for unbelievable problems, trained participants showed a more conservative response bias, $t(188) = 2.569, p = 0.011$. Believability and AH4 score also interacted. For unbelievable problems, there was only a marginally significant effect of AH4 score on response bias, $t(188) = -1.90, p = 0.059$, in the direction of a more liberal bias. However, for believable problems, higher AH4 score was associated with a more conservative response bias, $t(188) = 4.249, p < 0.001$.

Necessary vs. Impossible

Table 6.5: Parameter Values for Necessary vs. Impossible Accuracy Model

Variable	β	SE	t	LRT-p
Believability (believable)	0.115	0.041969	2.732	0.042
AH4	0.007	0.001540	4.621	< 0.001
REI Rational Subscale	0.017	0.006137	2.792	< 0.001
Training (present)	-0.037	0.018449	-2.013	0.045
Believability x REI Rational Subscale	-0.011	0.005869	-1.856	0.065

Accuracy Trained participants were less accurate at distinguishing between necessary and impossible problems than untrained participants, $\chi^2(1) = 4.15, p = 0.042$. Participants showed higher accuracy on believable, compared to unbelievable

problems, $\chi^2(1) = 14.595, p < 0.001$. A higher AH4 score was related to greater accuracy, $\chi^2(1) = 22.035, p < 0.001$. Once again, REI Rational subscale was significant, $\chi^2(1) = 4.019, p = 0.045$, with a higher REI R score linked to greater accuracy.

There was a marginally significant interaction between believability and REI Rational Subscale, $\chi^2(2) = 5.283, p = 0.65$, with no effect of REI R score on accuracy for believable problems, but a marginal effect on unbelievable problems, $t(188) = 1.901, p = 0.059$, with higher REI R scores associated with higher accuracy on these problems.

Table 6.6: Parameter Values for Necessary vs. Impossible Response Bias Model

Variable	β	SE	t	LRT-p
Believability (believable)	-1.689	0.376	-4.488	< 0.001
Training (present)	0.109	0.041	2.699	0.016
AH4	-0.005	0.005	-1.127	0.017
Believability x AH4	0.027	0.007	4.023	< 0.001

Response Bias Trained participants showed a more conservative response bias than untrained participants, $\chi^2(1) = 5.838, p = 0.016$. Participants showed a more liberal response bias to believable compared to unbelievable problems, $\chi^2(1) = 19.371, p < 0.001$. Participants with a higher AH4 scores showed a more liberal response bias. Believability and AH4 score also interacted, $\chi^2(1) = 16.053, p < 0.001$. For unbelievable problems, AH4 score was not related to response bias. However, for believable problems, a higher AH4 score was predictive of a more conservative response bias, $t(188) = 4.459, p < 0.001$.

Misclassification

As responses consisted of both binary validity responses, and ternary possibility responses, it seemed relevant to examine to what degree participants really understood the principle of logical necessity. A new binary variable "misclassified" was created, with a value of 0 for all responses for which the binary validity judgement

and possibility judgement were consistent with logical necessity, and a value of 1 for all responses that were not (i.e. if a participant deems a problem to be both valid and either possible or impossible, or both invalid and necessary).

Out of all responses, 31% were found to be misclassified.

A mixed effects model was fitted to the misclassification variable to identify the causes of this problem. Participants were more likely to have misclassified believable than unbelievable problems, $z = -5.93, p < 0.001$. Validity was also significant, although surprisingly reflecting a similar pattern to likelihood of conclusion acceptance, with necessary>possible strong>possible weak>impossible (ps all < 0.001).

Participants with a higher AH4 score were less likely to misclassify syllogisms, $z = -5.71, p < 0.001$, as were those with a higher REI rational engagement score, z=-2.21, p=0.027.

Due to the extent of misclassification found, a further SDT analysis was conducted on the binary responses, with corrections made to responses. The ternary responses ("necessary", "possible", and "impossible") were taken as being representative of participants' true judgements, and the binary ("valid"/"invalid") responses were corrected in order to be logically consistent with these. So, for example, a response of 'possible' and 'valid' would be adjusted to 'invalid'.

Necessary vs. Possible Strong - Adjusted

Table 6.7: Parameter Values for Necessary vs. Possible Strong Accuracy Model - Adjusted

Variable	β	SE	t	LRT-p
Believability (believable)	-0.043	0.016	-2.670	0.008
AH4	0.005	0.001	2.970	0.002
REI Rationality Subscale	0.011	0.006	1.899	0.059

Accuracy Participants were more accurate at distinguishing between necessary and possible strong conclusions when conclusions were unbelievable, $\chi^2(1) = 6.999, p = 0.008$. Those with a higher AH4 score were more accurate, $\chi^2(1) =$

$9.295, p = 0.002$. There was a marginal effect of REI Rational subscale, $\chi^2(1) = 3.571, p = 0.059$, with higher accuracy linked to higher scores on this scale. All of these variables were significant in the unadjusted analyses presented earlier, but the effects of training and interaction between AH4 score and believability were no longer significant.

Table 6.8: Parameter Values for Necessary vs. Possible Strong Response Bias Model - Adjusted

Variable	β	SE	t	LRT-p
Training (present)	0.138	0.065	2.127	0.024
Believability (believable)	-0.241	0.043	-5.495	< 0.001
AH4	-0.016	0.005	-2.979	0.003

Response Bias There was a significant effect of believability, $\chi^2(1) = 28.023, p < 0.001$, with more liberal response criterion for believable problems. Training was also significant, $\chi^2(1) = 5.115, p = 0.024$, with trained participants responding more conservatively. There was also a significant effect of AH4 score, $\chi^2(1) = 8.671, p = 0.003$, with participants with a higher AH4 score responding more liberally. The added effect of AH4 score was the only variable which was different to those found in the unadjusted analysis.

Necessary vs. Possible Weak - Adjusted

Table 6.9: Parameter Values for Necessary vs. Possible Weak Accuracy Model - Adjusted

Variable	β	SE	t	LRT-p
AH4	0.008	0.001	4.702	< 0.001
REI Rational subscale	0.018	0.006	2.817	0.005

Accuracy Participants with higher AH4 scores could better distinguish between necessary and possible weak arguments, $\chi^2(1) = 21.916, p < 0.001$. Participants with a higher score on the REI Rationality subscale were also more accurate, $\chi^2(1) = 7.775, p = 0.005$. These variables were significant in the unadjusted analysis; however, training was no longer significant.

Table 6.10: Parameter Values for Necessary vs. Possible Weak Response Bias Model - Adjusted

Variable	β	SE	t	LRT-p
Believability (believable)	-0.901	0.363	-2.481	0.002
AH4	-0.015	0.005	-2.814	0.056
Believability x AH4	0.014	0.006	2.166	0.031

Response Bias Participants showed a more liberal response bias on believable problems, $\chi^2(1) = 9.036, p = 0.002$. There was a marginally significant effect of AH4 score, with participants with higher scores showing a more liberal bias, $\chi^2(1) = 3.639, p = 0.056$. Believability and AH4 score interacted, $\chi^2(1) = 4.635, p = 0.031$. For believable problems, AH4 score had no effect on response bias. However, for unbelievable problems, a higher AH4 score was related to a more liberal response bias. The adjusted analysis differs in that AH4 score is now significant, but the believability by training interaction disappeared.

Necessary vs. Impossible - Adjusted

Table 6.11: Parameter Values for Necessary vs. Impossible Accuracy Model - Adjusted

Variable	β	SE	t	LRT-p
AH4	0.009	0.002	4.754	< 0.001
REI Rational Subscale	0.017	0.006	2.612	< 0.001

Accuracy Participants with a higher AH4 score were better able to distinguish between necessary and impossible problems, $\chi^2(1) = 22.348, p < 0.001$. Those with a higher REI Rationality subscale score were more accurate, $\chi^2(1) = 29.053, p < 0.001$. This analysis differed from the unadjusted analysis in that believability, training and an interaction between believability and REI Rationality subscale were no longer significant.

Response Bias Trained participants had a more conservative response bias, $\chi^2(1) = 4.300, p = 0.038$. There was a more liberal response bias on believable problems, $\chi^2(1) = 16.463, p < 0.001$. There was an interaction between believability

Table 6.12: Parameter Values for Necessary vs. Impossible Response Bias Model - Adjusted

Variable	β	SE	t	LRT-p
Training (present)	0.106	0.053	1.992	0.038
Believability (believable)	-1.211	0.358	-3.383	< 0.001
AH4	-0.016	0.005	-2.978	0.125
Believability x AH4	0.018	0.006	2.949	0.004

and AH4 score, $\chi^2(1) = 10.861, p = 0.004$. For beleivable problems, AH4 score was unrelated to response bias, $t(188) = 0.351, p = 0.726$. However, for unbelievable problems, a higher AH4 score was related to a more liberal response bias, $t(188) = -3.289, p = 0.001$. This analysis differs to the unadjusted analysis due to the lack of a significant main effect of AH4 score in the present analysis.

ROC curves for the unadjusted and adjusted analyses can be found in Figures 6.1 and 6.2 respectively.

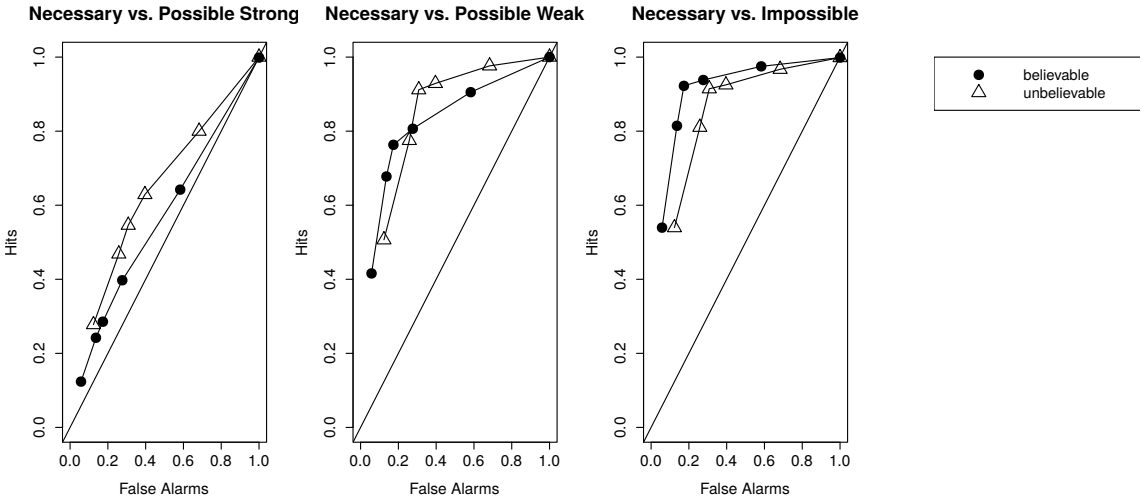


Figure 6.1: ROC curves for unadjusted analysis by problem type and believability

Response Times

Individual item response times were firstly analysed using a mixed effect model, with feedback, training, validity, and believability as explanatory variables. Response times were log transformed in order to satisfy the assumption of being normally

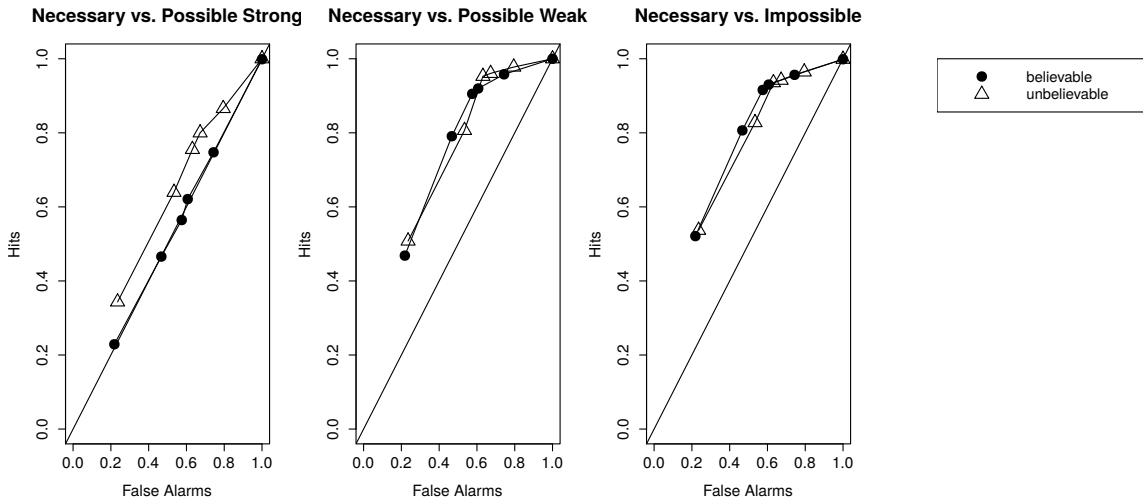


Figure 6.2: ROC curves for the adjusted analysis by problem type and believability

distributed.

Although data were cleaned earlier to remove participants who had simply clicked through the experiment, the data had to be cleaned again to remove outlier which would skew the results of the model. Therefore, responses over 180 seconds were removed. This led to the removal of 30 responses, all from different participants.

Response time was significantly predicted by validity, $\chi^2(3) = 25.055, p < 0.001$, with the ordering of response times: necessary = possible strong > possible weak = impossible.

Participants responded more slowly to believable problems, $\chi^2(1) = 5.892, p = 0.015$. Participants who had received training spent longer on each problem, $\chi^2(1) = 4.272, p = 0.039$, as did participants with a higher AH4 score, $\chi^2(1) = 28.734, p < 0.001$. REI Experiential subscale was also significant, $\chi^2(1) = 17.297, p < 0.001$, with participants who scored highly on this scale spending less time per syllogism.

There was an interaction between validity and believability, $\chi^2(3) = 11.781, p = 0.008$; for possible weak and impossible problems, believability did not predict response time ($ps > 0.05$). However, participants spent longer responding to believable conclusions if the problem was either necessary ($\chi^2(1) = 9.027, p = 0.003$)

Table 6.13: Parameter Values for Response Time Model

Variable	β	SE	t	LRT-p
Validity (necessary)	0.203	0.075	2.708	0.001
Validity (possible strong)	0.317	0.075	4.215	< 0.001
Validity (possible weak)	0.028	0.075	0.382	0.962
Training (present)	-0.507	0.261	-1.939	0.039
Believability (believable)	0.004	0.026	0.163	0.015
AH4 score	0.027	0.005	5.305	< 0.001
REI Experientiality	-0.049	0.019	-2.484	< 0.001
REI Rationality	-0.030	0.022	-1.340	0.684
Validity (necessary) x Believability	-0.085	0.036	-2.310	0.003
Validity (possible strong) x Believability	-0.075	0.036	-2.057	0.006
Validity (possible weak) x Believability	0.016	0.036	0.440	0.214
Validity (necessary) x REI Experientiality (low)	-0.014	0.011	-1.334	0.013
Validity (possible strong) x REI Experientiality (low)	-0.032	0.011	-2.935	0.023
Validity (possible weak) x REI Experientiality (low)	-0.004	0.011	-0.395	0.785
Training x REI Rationality	0.098	0.036	2.702	0.023

or possible strong ($\chi^2(1) = 7.484, p = 0.006$).

Validity and REI Experiential score also interacted, $\chi^2(3) = 10.220, p = 0.017$. For all problems, a lower REI Experiential score predicted response time ($psall < 0.05$). Therefore, a median split was performed by REI Experiential score. For participants who had a high REI Experiential score, there was no relationship between validity and response time, $\chi^2(3) = 3.288, p = 0.349$. However, for participants who had a lower score in this scale, validity was significant, $\chi^2(3) = 38.844, p < 0.001$, with the pattern of response times reflecting the main effect of validity, with necessary = possible strong > possible weak = impossible.

Training and REI Rational score interacted, $\chi^2(2) = 7.560, p = 0.023$. For untrained participants, REI rational score did not predict response time, $\chi^2(1) = 1.124, p = 0.289$. However, for trained participants, a higher REI Rational score was related to a longer response time, $\chi^2(1) = 7.747, p = 0.005$.

Confidence

Once again confidence ratings were examined as response variables, with validity, believability, feedback, training, and the individual difference measures as explanatory variables.

There was a main effect of validity, $\chi^2(3) = 69.438, p < 0.001$; participants' confidence in their responses was necessary = possible strong < possible weak < impossible. REI Rationality subscale was also significant, $\chi^2(1) = 26.451, p < 0.001$, with higher Rationality scores linked to higher confidence.

Table 6.14: Parameter Values for Confidence Model

Variable	β	SE	t	LRT-p
Validity (necessary)	-0.621	0.084	-7.334	< 0.001
Valiidty (possible strong)	-0.591	0.085	-6.943	< 0.001
Validity (possible weak)	-0.336	0.085	-3.947	< 0.001
REI Rationality	0.535	0.101	5.281	< 0.001

6.3 Discussion

Experiment 3 aimed to investigate whether the combination of feedback and training would lead to an improvement in reasoning accuracy, and whether any change would be contingent on participants' cognitive style and/or ability. Neither training, feedback, or a combination of both improved reasoning accuracy. Although training appeared at first to lead to a decrease in accuracy for distinguishing all invalid problem subtypes (possible strong, possible weak, and impossible) from valid (necessary) problems, once misclassfication was corrected for, this effect disappeared. In the adjusted comparisons of necessary and possible strong problems, - those most similar to ones used in other studies as valid and invalid problems, training affected response bias but not reasoning bias. This was also the case for the comparison of necessary and impossible problems.

These results differ to those of Prowse-Turner and Thompson (2009) as training did not improve participants' ability to discriminate between valid and invalid

syllogisms. One explanation for the lack of effect of training is that participants in the training condition did not engage in any extra effort compared to untrained participants. However, the response time data does not support this view; trained participants took significantly longer than untrained participants to respond. As well as this, training did however cause participants with more rational tendencies to spend longer on each problem. However, this result did not lead to an increase in overall accuracy or a reduction in bias. This may be consistent with a selective processing theory account of belief bias; although trained participants were more likely to be aware of alternative representations after being shown the diagrams in the training phase of the experiment, whether they constructed verifying or falsifying representations may have been affected by conclusion believability. This is consistent with Thompson et al. (2003) who found that even when participants were instructed to construct multiple representations of syllogisms (using Euler circles), they still tended to constructed a single representation. A further explanation of the lack of effect of feedback and training is the online format of the experiment. In Prowse-Turner and Thompson (2009), participants were given feedback and training face-to-face, and with extra explanations if they were unsure of an answer. Kluger and Denisi (1996) highlight how verbal feedback can have a greater effect in improving task performance than written feedback.

As training was insufficient to improve performance, it may have simply served to confuse participants - trained participants were much poorer at distinguishing between valid conclusions and any of the other conclusion type, whether possible or impossible. However, when misclassifications were corrected for, this effect of training was no longer significant.

These results also provide further support for the use of SDT models in distinguishing between reasoning and response bias. Although the endorsement rate analysis showed that believability affected endorsement for all problems, the SDT analyses showed that although believability affected response bias for all problems, belief bias only affected reasoning when looking at participants' ability to

distinguish between necessary and possible strong problems.

Interesting insights can be gained from the misclassification analysis, which showed that almost a third of responses did not conform to the rules of logical necessity. One potential explanation for this is that many participants may have been simply 'clicking through' the task without engaging with it and thus choosing answers at random. However, this seems unlikely given the extensive data screening procedure that had been used to identify such participants on the basis of both response time and answers given. An alternative explanation for this is that many participants had a poor understanding of the concept of logical necessity, and therefore responded in this manner. This finding has important implications; typically it is assumed that instructions which emphasise that a conclusion can only be valid if it is necessitated by the premises are sufficient to account for any lack of complete knowledge of logical necessity. However, the evidence presented here is not consistent with that view. When binary choice models are used, participants who realise a syllogism is 'ossible' rather than 'necessary' but don't realise that this means the conclusion is not valid, are thought to be less accurate in discriminating between valid and invalid problems. In addition, it was found that participants with a more rational thinking style, and those with higher cognitive ability were less likely to misclassify responses. These results also call in to question Stanovich (2009) classification of belief bias as being mainly the result a failure of Type 2 processing to override Type 1 processing; here there is evidence that it is the result of multiple sources, including mindware gaps, as shown in the misclassification analysis.

The results are consistent with previous individual differences approaches which argue that competent syllogistic reasoning relies on both cognitive ability and cognitive style (e.g. Trippas, Verde, & Handley, 2015). Once misclassification had been accounted for, both AH4 score and REI Rationality subscale were significant in all models for the SDT accuracy measures, and also in the response time model.

These results differ from those found in Experiment 2; in the previous experiment,

the effect of cognitive style on accuracy was contingent on problem believability. Here, there was no interaction between style and believability, but these effects separately predicted accuracy. This, however, may have been due to differences between the two samples; the previous experiment used a mix of online and lab participants, and the believability by AOT interaction was only found in the lab participants. This difference may have been due to the online participants being more engaged with the problems. The experiment presented in this chapter used only online participants, which may explain the effect of believability across all participants regardless of AOT or REI scores.

A problematic finding is that once misclassifications were accounted for, the advantage for participants with a higher AH4 score on unbelievable problems disappeared. One possible explanation for this is that participants of a higher ability simply are more likely to have a better grasp of the concept of logical necessity.

A possible criticism of this study is regarding the quality of the data due to it being collected online. Participants may have not been fully paying attention during the study. However, this seems unlikely; data were removed from participants who showed consistently unrealistically short or long response times, which constituted only 5% of participants, and individual long or short responses were removed from the response time analyses, leading to just 1% of responses being removed. The mean response time was 25.86 seconds, comparable to the mean response time of 25.01 seconds in the previous experiment.

Unexpectedly, a much smaller proportion of responses were removed after the data cleaning procedure than in Experiment 2. Another finding was that 15% of participants who began the experiment did not complete it. This may have been because many of the participants who would have simply "clicked through" in the previous experiment gave up due to the longer length of the present study. However, this may have also led to a less representative sample, as it could have been participants who found the task difficult gave up at the start. Thus, the

sample may have been distorted by containing a higher number of participants who had higher syllogistic reasoning ability than might have been expected.

There was no effect of belief bias on participants' reasoning for possible weak or impossible problems. This is reflected in the lack of effect of believability in the accuracy model for these conclusion type, and also the fact that although believability predicted response times for necessary and possible strong problems, it did not predict response times for possible weak or impossible problems. This seems to be analogous to Evans et al's (2001) finding of positive belief bias (an increased tendency to endorse believable conclusions) for possible weak problems, and negative belief bias (a decreased tendency to endorse unbelievable problems) for possible strong problems. What Evans and colleagues termed positive and negative belief bias may be roughly analogous to the response and reasoning components of belief bias.

A further explanation for the lack of training effect is differences in reasoning style; Monaghan and Stenning (2004) report that being able to decontextualise graphically and linguistically are dissociable skills. Thus, participants who employed a primarily non-graphical way of representing syllogisms would have been unlikely to benefit from training.

In summary, training and feedback did not lead to an improvement in reasoning. However, the misclassification analysis affords interesting insights into some aspects of reasoning. In the final chapter, I will discuss the empirical and theoretical implications of the research presented in this thesis.

Chapter 7

General Discussion

The main aim of this thesis was to examine methods of debiasing reasoning with a view to reducing the impact of belief bias on syllogistic inference. Debiasing reasoning is important for multiple reasons; firstly, because unbiased logical thinking is important to the improvement of critical thinking skills (Macpherson & Stanovich, 2007), and secondly because looking at how individuals improve performance may allow us deeper insight into how the underlying cognitive processes behind biased and unbiased thinking operate. This thesis aimed to untangle reasoning and response bias by applying an SDT approach to analyses in order to avoid the problems that earlier analytic methods had in conflating the two types of bias. These earlier methods, such as calculating linear indices of logic, belief and an interaction between the two led to problematic results as such analytic methods cannot distinguish between reasoning bias and response bias. Thus, the use of SDT is necessary in getting a fine-grained view of reasoning. This is especially true in the present research that aims to examine methods of debiasing reasoning; it is crucial that we are able to distinguish between methods which lead a genuine change in reasoning and those which simply alter participants' patterns of responses. Another aim was to examine the contribution of individual differences in cognitive style and ability to strategy use and susceptibility to debiasing interventions.

Experiment 1 aimed to examine the effects of feedback and practice upon

belief bias. Participants completed 16 syllogisms per session across the course of 5 sessions. It was predicted that feedback would lead to a decrease in both reasoning and response bias, and that repeated practice may also affect responding. It was found that although feedback led to changes in response bias, it had no effect upon reasoning biases. In addition, this change was immediate and remained fairly stable across the five sessions. These results were supported by an examination of response times, which showed that there was no accompanying change in response time, which one would expect if a genuine change in strategy had occurred. In addition, it was shown that changes in response bias are task-specific, with there being no differences between the feedback and no-feedback groups on a secondary conditional reasoning task.

Experiment 2 aimed to examine the effects of instructions emphasising the importance of logical necessity on reasoning and whether individual differences played a part in mediating the strength of any such effects. It was predicted that there would be an effect of instructional condition and this would be contingent on participants' cognitive ability or cognitive style. The differing instructions did not lead to a change in performance and seemed to serve only to confuse participants. Although not the original focus of the analyses, differences between online and lab participants were also examined when it became clear that they had a crucial effect on the results. Although previous research has shown similarities between online and lab participants in similar research concerning topics in heuristics and biases, the analyses presented in Experiment 2 show that differences between these groups may be masked by non-SDT approaches which cannot separate response bias and reasoning bias. Whilst the online group showed belief bias as both a reasoning and a response bias, it was only the lab participants with a more open-minded cognitive style who showed signs of reasoning bias. It was concluded that this was due to the former group engaging with more effort on the task, a view supported by evidence surrounding the motivational characteristics of online participants.

In Experiment 3, the effects of training and feedback were examined, combined

with measures of individual differences to again examine the different propensity of individuals to be affected by training and feedback. It was predicted that training and feedback would lead to a decrease in reasoning bias. However, there were no main effects of either feedback or training. Although limited in terms of the minimal effects arising from such interventions, it was nevertheless shown that giving participants training led to an increase in response time for participants with greater pre-existing tendencies towards relying upon rational thought, suggesting that such participants attempted to engage in more effort, but were still unable to increase their performance levels. Furthermore, the analyses presented in this chapter also showed how responses may be affected by some degree of lack of understanding of the concept of logical necessity, and that when results are corrected for this, it allows for a clearer picture of significant effects. The examination of multiple problem types provided further evidence that the reasoning bias and response bias components of belief bias may be analogous to what previous approaches have termed "positive belief bias" and "negative belief bias".

In all three experiments, evidence was found to support theories of belief bias which place the locus of its effects in both reasoning and response bias. Furthermore, the importance of the SDT approach was emphasised; the use of traditional analyses in all of the three experiment would have led to misleading or inaccurate results which either falsely showed differences between conditions, or hid differences between different groups of participants or different materials. These results undermine earlier theories of belief bias such as selective scrutiny, misinterpreted necessity, metacognitive uncertainty, and mental-models theory, as none of these theories predict both a reasoning and response bias effect. However, they support the SDT model suggested by Trippas et al, who argue that belief bias primarily manifests as a response bias but in the case of participants who engage in complex reasoning, belief bias is also a reasoning bias.

One limitation of the present study is that no belief-neutral premises were used. Thompson (1996) found that the believability of the premises could affect response

choice. However, later research by Thompson et al (2013) found that premise believability was irrelevant to how many models were generated, and given that syllogistic reasoning in the format presented in this thesis is thought to consist of conclusion-to-premise reasoning, it is unlikely to have had much, if any, effect upon the results.

7.1 Theoretical Implications

The research presented in this thesis provides further evidence regarding the importance of considering individual differences. In both Experiments 2 and 3, cognitive ability accurately predicted accuracy. In Experiment 2, accuracy was predicted by open-minded thinking for the lab participants, although not for the online participants. In Experiment 3, although training had no effect on improving task performance, it did lead to an increase in response times for participants with a predisposition towards rational thinking, indicating the potential importance of individual differences in developing debiasing methods.

This research also contributes to the debate over whether it is analytic cognitive style or cognitive ability which predicts reasoning performance. Although Trippas et al (2010) and Trippas et al (2014, Experiment 4) supported the theory that higher cognitive ability is linked to reasoning bias, Trippas et al (2014, Experiment 5) suggest that it is in fact cognitive style which is important, with the apparent effects of cognitive ability being the result of the high degree of correlation between analytic cognitive style and higher cognitive ability. This argument is developed further by Trippas et al (2015) who found that although cognitive ability is linked to general accuracy, it is cognitive style which predicts reasoning bias. In this thesis, Experiment 3 provided further evidence that both cognitive style and ability had an effect on overall accuracy when distinguishing between necessary and possible strong problems those commonly used in belief bias studies. However, the ability of AH4 score to predict the amount of belief bias shown disappeared once misclassifications were accounted for, implying that it is this misclassification

which drives the advantage. The effect of cognitive style remained, despite the correction, suggesting that it is really style but not ability which is predictive of belief bias as a reasoning bias.

The REI, AOT, and CRT have been used as measures of cognitive style previously, and all three were included as measures of cognitive style in Experiments 2 and 3 to examine which is the better self-reported measure of cognitive style. Given the inconsistency in their predictive effects, it seems prudent to explore what the difference is between these measures. Newstead et al (2004) caution that the REI should not be treated as a general measure of System 1 type thinking, and argue that nor should any self-reported measures. Trippas et al (2015) also emphasise the use of performance measure over self-report scales. Nevertheless, the use of all three scales produced some interesting findings. In Experiment 2, the CRT was not a better predictor than REI or AOT of any response variables (except response time) and so was not included in Experiment 3. In Experiment 2, lab participants only showed reasoning bias if they were prone to more open-minded thinking. Reasoning bias was not contingent on thinking style for online participants. However, in both Experiments 2 and 3, the REI predicted general accuracy; in Experiment 2, participants who relied more on experiential thinking were less accurate, and for all problem types in Experiment 3, those who relied more on rational thinking showed higher accuracy. Given the differences in the predictive nature of the two tests, there is clearly some distinction between what each of them measure. It seems that whilst the AOT examines the tendency to consider different alternatives, a view also supported by Stanovich & West (1998), the REI measures a general tendency to analyse at things in more depth but not necessarily to consider opposites. Thus, participants with higher REI rational or lower REI experiential scores may spend more time considering their answers (as shown by the response time analyses in Experiments 2 and 3), but may simply be searching for further justifications for their chosen response. This would be consistent with Evans et al (1983) who found that even when participants had to provide justification for their choice of a

particular response, when giving a biased response, simply referred to irrelevant information in their answer.

More recent research by Pennycook et al (2015) suggests similar roles for cognitive style and ability, albeit under a different framework. Their theory posits three stages in reasoning. Firstly, Type 1 processing can cue multiple responses. In the second stage, a conflict between multiple responses may be detected. In the third stage, if conflict is detected, two sorts of Type 2 outputs are possible; firstly, the Type 1 response might be supported by (inaccurate) Type 2 responses, or secondly, a normatively correct Type 2 response may be given as the result of further simulation. Stages 2 and 3 may be akin to cognitive style and ability; participants with a more open-minded cognitive style may be more likely to consider a wider range of alternatives, and so more likely to detect conflict. During stage 3, cognitive ability comes into play; those able to simulate multiple possibilities succeed and those unable to do so proceed to rationalise their chosen salient response. This model is distinct from some earlier dual-process models in that it highlights the fact that Type 2 processing may still have occurred in non-normative responding, and thus contributes to the 'quality' argument (Evans 2007); that differences in levels of bias shown are contingent on the *kind* of reasoning rather than the *amount*.

In other words, if one is more open-minded, one can detect conflict better - as a wider range of possible models are considered. But any effect on accuracy is contingent on having higher cognitive ability and be able to separate belief from logic, otherwise the original heuristic decision is simply further justified. The findings of Experiment 2 do not support this theory; one would expect longer response times for those of a more open-minded or reflective cognitive style, but given that both cognitive style and ability were included in the analyses, it should be cognitive ability which predicts the effects of reasoning bias. However, it was found that for the lab participants, it was in fact cognitive style which predicted the degree of reasoning bias shown. Experiment 3, however, does lend some support to this theory; for participants with more rational thinking dispositions, training

led to longer response times, indicating that it encouraged these participants to consider more options. However, this was not accompanied by an increase in accuracy; perhaps these participants simply expended more effort rationalising these decisions. This would also be consistent with the view discussed above that whilst the AOT measures the tendency to consider alternatives, the REI simply measures the tendency to rationalise one's choices.

7.2 Methodology

7.2.1 Online Testing

The experiments presented in this thesis highlighted important methodological considerations for belief bias research. Firstly, differences between the data collected in the lab compared to online highlight differences in motivations for these participants. In Experiment 2, online participants all showed reasoning bias, whereas the lab participants' results were more in line with those seen in comparable studies (e.g. Trippas et al., 2013); only participants with a more open-minded thinking style were more accurate on unbelievable problems. A similar result was also found in Experiment 3, where there was no interaction between believability and thinking style or cognitive ability in any of the analyses focussing on accuracy. Previous research has shown that participants recruited via Amazon Turk often show high levels of attention and task motivation. The online participants in Experiment 2 showed a greater degree of reasoning bias, and given the similarities in terms of individual difference measures, one may assume that this is the consequence of these participants being more engaged with the task. Given the speed in which large volumes of data can be collected in this manner, online testing may be of great use in belief bias research. Although concerns have been raised previously about the quality of such data, the use of participant screening and data cleaning alleviates such concerns as long as these procedures involve clearly defined and theoretically based criteria for inclusion and exclusion. One limitation of online

testing, however, is task familiarity. In Experiment 2, the online participants score significantly higher in the CRT than the lab participants. Given that the CRT is widely used in psychology experiments and attention has already been drawn to the fact that many participants are already familiar with its contents (Toplak, West & Stanovich, 2013), this highlights the importance of developing materials for online tests that will not be familiar to the participant pool, some of whom will have participated in a large number of psychology experiments (Buhrmester et al, 2011) and thus been exposed on numerous occasions to many widely used tests.

Despite online testing being advantageous due to being an inexpensive way of rapdily collecting large amounts of data, an obvious disadvantage is the sacrifice in level of experimenter control, compared to that which would be found in a laboratory experiment. Despite instructions indicating otherwise, participants may have been completing the experiment whilst otherwise distracted. Linked to this is the lack of opportunity for further elaboration and explanation by the experimenter. Given the positive results of training found by Prowse-Turner & Thompson (2009) which was very similar to the one presented in the present study, one might have expected similar results in Experiment 3. However, it may have been the lack of in-depth explanation which prevented participants from benefiting from the training in the present study. Therefore, the decision of whether to test participants in a lab setting or online should take into account the nature of the experimental factors being manipulated in deciding if online testing is appropriate or not.

7.2.2 Analyses

A number of the key contribution of this thesis centre around combining established methods with an SDT approach in order to distinguish between reasoning bias and response bias. In Experiment 1, feedback affected response but not reasoning bias, and this distinction was only apparent due to the use of SDT analyses. Using traditional measures of accuracy would have led to the conclusion that feedback had

improved reasoning performance. Similarly, the SDT approach was also important in Experiment 2 in showing the differences between the two groups. Whilst the SDT approach showed that only the online participants showed a general effect of belief bias as a reasoning bias but lab participants only showed signs of response bias unless they had higher cognitive ability, the traditional analyses simply implied a general effect of belief bias on all participants.

Another key contributions of this thesis was to demonstrate the importance of triangulation of measures when examining belief bias, the importance of which has been emphasised by Stupple & Ball (2014). The use of response times and confidence ratings allowed an examination of changes in performance, even when accuracy and response bias remained constant. For example, in Experiment 1, the case for a lack of an accuracy effect of feedback was strengthened by the lack of a response time effect of feedback. In Experiment 3, it was found that those with more rational tendencies were more likely to spend longer on each problem, but did not show an increase in accuracy.

One potential limitation of the research presented in this thesis is the method of analysis used to calculated the SDT accuracy value. Due to issues with getting models to fit data which had empty cells (i.e. if participants didn't use the full range of the confidence rating scale), there were problems with using existing software packages, and a method described by Trippas et al (2014) as "more straightforward if less reliable", the regression of z scores for hits against the z scores for false alarms, was used. However, I tested this methodology on example data sets from textbooks (e.g. Wickens, 2002), and no significant differences were found between this analysis, and ones reported , and so this limitation should not have had any real effects on the results.

7.3 Debiasing

In Experiment 1, participants were given feedback and opportunities to practice solving syllogisms. These interventions may have caused participants to become

more aware of their own biased responding; however, an attempt to change strategy led to simply a change in response bias, as there was not a complete understanding of the reason for error.

In Experiment 2, extra emphasis on the importance of logical necessity was still insufficient to debias reasoning, despite the fact that in Experiment 3 it was demonstrated that some of the variation in performance was attributable to an incorrect understanding of logical necessity. The final experiment found that even with training on constructing alternative representations of syllogisms, there was no increase in accuracy for these participants generally or specifically on believable problems. However, training did lead to an increase in effort with this effect amplified in participants who had greater rational tendencies. This suggests that the consideration of individual differences between participants may be a crucial consideration in developing an effective method of debiasing reasoning.

Although not shown in the results presented in this thesis, it still remains feasible that training could improve reasoning and reduce reliance on belief. Prowse-Turner & Thompson (2009) included a training manipulation which involved one-to-one instruction and careful elaboration by an experimenter, which was found to improve performance on a syllogistic reasoning task containing neutral content. Future research should examine whether a similarly elaborate training procedure to that used by Prowse-Turner & Thompson (2009) can lead to the reduction of belief bias. Such research could also involve a detailed qualitative analysis of any diagrams which participants create in order to examine whether, as predicted by selective processing theory, participants are likely to construct confirming or disconfirming models on the basis of belief, and whether it is more effective for a successful training method to encourage them to rely on constructing a disconfirming model, or exhaustively represent all possible models. Such analysis may provide clear arbitration between different theories of belief bias.

A further possibility could be tailored training or different forms of training. Although it is tempting to suggest that training could be tailored to the individual

on the basis of individual differences, it may be simpler to allow individuals to have more control over their own learning. Johnson-Laird (2015) claims that the key way to improve reasoning is to teach individuals to consider all possibilities. He argues that this can be done through intensive critical thinking training, or through the use of diagrams. Johnson-Laird recommends what is termed the 'model method' over general diagrams, claiming that the model method has the advantages of not needing to be represented graphically, is simple to learn, and applicable to multiple contexts. The model method was devised by Bell (1999) and consists of a single instruction: "Try to construct all the possibilities consistent with the given information". Participants are then given a conditional reasoning problem, and are shown a worked out example to guide them in constructing representations, but are allowed to construct their own representations using any form of diagram that they choose. Although Johnson-Laird (2015) concedes that it remains to be seen whether the model method would be effective for syllogisms, this method sounds feasible. It is expected that given that reasoners are sufficiently motivated, the greatest benefit would be seen in those who have a lower score on the AOT as those with a higher score are already likely to think more flexibly.

7.4 Conclusion

The results presented in this thesis show that debiasing reasoning is not a trivial task and any successful interventions must do more than simply encourage participants to engage in deeper reflection on their answers. The reasoning component of belief bias is an extremely persistent cognitive bias and a successful debiasing intervention must necessarily be in-depth and comprehensive. It is also important to carefully consider the effects of the particularly methodology used; the sample of participants used may be as important as the debiasing method attempted. In addition, SDT analyses are crucial for distinguishing between reasoning and response bias, and for getting a clear and accurate picture of results. Finally, it is clear that previous approaches which equated higher accuracy or longer response times alone as a

simple metric of Type 2 processing are insufficient, and that both must be viewed alongside one another if we are to make useful inferences about these processes. In addition, exploring individual differences is also an important consideration if we wish to get an accurate picture of what is going on. Debiasing reasoning is a noble aim for science, but also a deeply complicated one, and the nuances of the particular approach, sample, or methodology are key to the effectiveness of any debiasing intervention.

Appendices

Appendix A

One Complete Set of Syllogisms

Number	Content			Structure					
	A	B	C	Figure	Mood	VB	Figure	Mood	VU
1	metals	soft things	steel	ABBC	IE	Some metals are soft things No soft things are steel	BACB	EI	No soft things are metals Some steel are soft things
				AC		Therefore, some metals are not steel	CA		Therefore, some steel are not metals
2	animals	reptiles	cats	BACB	IE	Some reptiles are animals No cats are reptiles	ABBC	EI	No animals are reptiles Some reptiles are cats
				AC		Therefore, some animals are not cats	CA		Therefore, some cats are not animals
3	parrots	tame creatures	birds	BACB	EI	No tame creatures are parrots Some birds are tame creatures	ABBC	IE	Some parrots are tame creatures No tame creatures are birds
				CA		Therefore, some birds are not parrots	AC		Therefore, some parrots are not birds
4	crocuses	grey things	flowers	ABBC	EI	No crocuses are grey things Some grey things are flowers	BACB	IE	Some grey things are crocuses No flowers are grey things
				CA		Therefore, some flowers are not crocuses	AC		Therefore, some crocuses are not flowers
5	daffodils	blue things	flowers	BACB	EI	No blue things are daffodils Some flowers are blue things	ABBC	IE	Some daffodils are blue things No blue things are flowers
				CA		Therefore, some flowers are not daffodils	AC		Therefore, some daffodils are not flowers
6	pedigree dogs	small things	mammals	ABBC	EI	No pedigree dogs are small things Some small things are mammals	BACB	IE	Some small things are pedigree dogs No mammals are small things
				CA		Therefore, some mammals are not pedigree dogs	AC		Therefore, some pedigree dogs are not mammals
7	vehicles	machines	trams	ABBC	IE	Some vehicles are machines No machines are trams	BACB	EI	No machines are vehicles Some trams are machines
				AC		Therefore, some vehicles are not trams	CA		Therefore, some trams are not vehicles
8	birds	cartoon characters	sparrows	BACB	IE	Some cartoon characters are birds No sparrows are cartoon characters	ABBC	EI	No birds are cartoon characters Some cartoon characters are sparrows
				AC		Therefore, some birds are not sparrows			Therefore, some sparrows are not birds
9	men	boys	reigning kings	ABBC	IE	Some men are boys No boys are reigning kings	BACB	EI	No boys are men Some reigning kings are boys
				AC		Therefore, some men are not reigning kings			Therefore, some reigning kings are not men
10	snakes	venomous things	pythons	BACB	IE	Some venomous things are snakes No pythons are venomous things	ABBC	EI	No snakes are venomous things Some venomous things are pythons
				AC		Therefore, some snakes are not pythons			Therefore, some pythons are not snakes
11	taxi cabs	inexpensive things	vehicles	BACB	EI	No inexpensive things are taxi cabs Some vehicles are inexpensive things	ABBC	IE	Some taxi cabs are inexpensive things No inexpensive things are vehicles
				CA		Therefore, some vehicles are not taxi cabs			Therefore, some taxi cabs are not vehicles
12	deep sea divers	smokers	good swimmers	ABBC	EI	No deep sea divers are smokers Some smokers are good swimmers	BACB	IE	Some smokers are deep sea divers No good swimmers are smokers
				CA		Therefore, some good swimmers are not deep sea divers			Therefore, some deep sea divers are not good swimmers
13	lobsters	ocean dwellers	crustaceans	BACB	EI	No ocean dwellers are lobsters Some crustaceans are ocean dwellers	ABBC	IE	Some lobsters are ocean dwellers No ocean dwellers are crustaceans
				CA		Therefore, some crustaceans are not lobsters			Therefore, some lobsters are not crustaceans
14	strawberries	sour food	fruit	ABBC	EI	No strawberries are sour food Some sour food are fruit	BACB	IE	Some sour food are strawberries No fruit are sour food
				CA		Therefore, some fruit are not strawberries			Therefore, some strawberries are not fruit
15	fish	oily food	tuna	ABBC	IE	Some fish are oily food No oily food are tuna	BACB	EI	No oily food are fish Some tuna are oily food
				AC		Therefore, some fish are not tuna			Therefore, some tuna are not fish
16	reptiles	green animals	alligators	BACB	IE	Some green animals are reptiles No alligators are green animals	ABBC	EI	No reptiles are green animals Some green animals are alligators
				AC		Therefore, some reptiles are not alligators			Therefore, some alligators are not reptiles

One complete set of syllogisms. The coloured backgrounds indicate subsets (i.e. so all syllogisms with a yellow background belong to subset 1, all syllogisms with a green background belong to subset 2 etc). Continued overleaf.

Figure	Mood	IB	Figure	Mood	IU
ABBC	EI	No metals are soft things Some soft things are steel AC	BACB	IE	Some soft things are metals No steel are soft things CA
BACB	EI	No reptiles are animals Some cats are reptiles AC	ABBC	IE	Some animals are reptiles No reptiles are cats CA
BACB	IE	Some tame creatures are parrots No birds are tame creatures CA	ABBC	EI	No parrots are tame creatures Some tame creatures are birds AC
ABBC	IE	Some crocuses are grey things No grey things are flowers CA	BACB	EI	No grey things are crocuses Some flowers are grey things AC
BACB	IE	Some blue things are daffodils No flowers are blue things CA	ABBC	EI	No daffodils are blue things Some blue things are flowers AC
ABBC	IE	Some pedigree dogs are small things No small things are mammals CA	BACB	EI	No small things are pedigree dogs Some mammals are small things AC
ABBC	EI	No vehicles are machines Some machines are trams AC	BACB	IE	Some machines are vehicles No trams are machines CA
BACB	EI	No cartoon characters are birds Some sparrows are cartoon characters	ABBC	IE	Some birds are cartoon characters No cartoon characters are sparrows
ABBC	EI	No men are boys Some boys are reigning kings	BACB	IE	Some boys are men No reigning kings are boys
BACB	EI	No venomous things are snakes Some pythons are venomous things	ABBC	IE	Some snakes are venomous things No venomous things are pythons
BACB	IE	Some inexpensive things are taxi cabs No vehicles are inexpensive things	ABBC	EI	No taxi cabs are inexpensive things Some inexpensive things are vehicles
ABBC	IE	Some deep sea divers are smokers No smokers are good swimmers	BACB	EI	No smokers are deep sea divers Some good swimmers are smokers
BACB	IE	Some ocean dwellers are lobsters No crustaceans are ocean dwellers	ABBC	EI	No lobsters are ocean dwellers Some ocean dwellers are crustaceans
ABBC	IE	Some strawberries are sour food No sour food are fruit	BACB	EI	No sour food are strawberries Some fruit are sour food
ABBC	EI	No fish are oily food Some oily food are tuna	BACB	IE	Some oily food are fish No tuna are oily food
BACB	EI	No green animals are reptiles Some alligators are green animals	ABBC	IE	Some reptiles are green animals No green animals are alligators

Continued from previous page.

Appendix B

SDT measures code in R

```
# Gets Az given an intercept and coefficient
get_az=function(intercept, coef){
  topline = sqrt(2)*intercept
  bottomline = sqrt(1+coef^2)
  da = topline/bottomline
  az = pnorm(da/sqrt(2))
  return(az)
}

# Gets Az given the proportions of hits and false alarms
# Used when SDT unequal variance model is a poor fit
get_az_dprime=function(prop_h, prop_f){
  d_prime = qnorm(prop_h)-qnorm(prop_f)
  az = pnorm(d_prime/sqrt(2))
  return(az)
}

# Gets ca  - the response criterion -
```

```
# for given slope coefficient and proportion of hits and false alarms
get_ca=function(coef, prop_h, prop_f){
  topline = - (sqrt(2)*coef)
  bottomline = sqrt(1+coef^2)*(1+coef)
  multiplier = qnorm(prop_h)+qnorm(prop_f)
  ca = (topline/bottomline)*multiplier
  return(ca)
}

# The ca - response criterion for equal variance models
get_ca_eq=function(prop_h,prop_f){
  return(-0.5*(qnorm(prop_h)+qnorm(prop_f)))
}
```

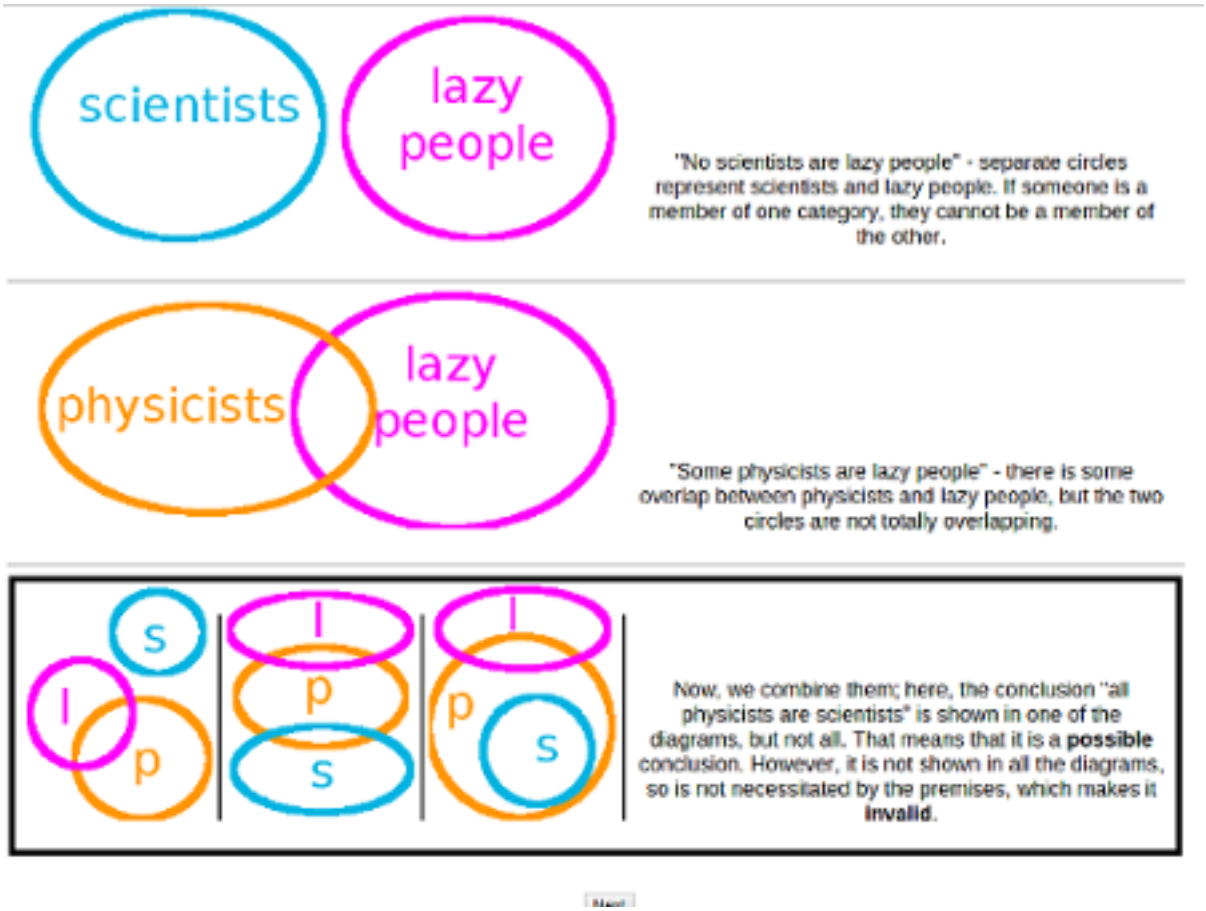
Appendix C

Slow Bootstrap

```
# Adapted from Long (2012)
slow.b=function(x,y,z){
  chisq.star=numeric(x)
  for(i in 1:x){
    simDV=simulate(y)
    full.s=refit(z,simDV[,1])
    reduced.s=refit(y,simDV[,1])
    chisq.star[i]=-2*(logLik(reduced.s)-logLik(full.s))
  }
  mean(anova(y,z)[2,6]<chisq.star)
}
```

Appendix D

Training



References

Alter, A. L., Oppenheimer, D. M., & Zemla, J. C. (2010). Missing the trees for the forest: A construal level account of the illusion of explanatory depth. *Journal of Personality and Social Psychology*, *99*(3), 436.

Arkes, H. R. (1991). Costs and benefits of judgment errors: Implications for debiasing. *Psychological Bulletin*(3), 486–498.

Arkes, H. R., Dawes, R. M., & Christensen, C. (1986). Factors influencing the use of a decision rule in a probabilistic task. *Organizational Behavior and Human Decision Processes*, *37*(1), 93–110.

Arkes, H. R., Faust, D., Guihnette, T. J., & Hart, K. (1988). Eliminating the hindsight bias. *Journal of Applied Psychology*, *73*, 305-307.

Ball, L. J. (2013). Microgenetic evidence for the beneficial effects of feedback and practice on belief bias. *Journal of Cognitive Psychology*, *25*(2), 183–191.

Ball, L. J., Hoyle, A., & Towse, A. (2010). The facilitatory effect of negative feedback on the emergence of analogical reasoning abilities. *British Journal of Developmental Psychology*, *2*, 583–603.

Ball, L. J., Phillips, P., Wade, C. N., & Quayle, J. D. (2006). Effects of Belief and Logic on Syllogistic Reasoning: Eye-Movement Evidence for Selective Processing Models. *Experimental Psychology*, *53*(1), 77–86.

Baron, J. (1995). Myside bias in thinking about abortion. *Thinking and Reasoning*, *1*, 221-235.

Baron, J. (2008). *Thinking and deciding*. Cambridge University Press.

Bazerman, M. H., Loewenstein, G. F., & White, S. B. (1992). Reversals of preference in allocation decisions: Judging an alternative versus choosing among alternatives. *Administrative Science Quarterly*, 220–240.

Braine, M. D. S., & OBrien, D. P. (1991). A theory of if: A lexical entry, reasoning program, and pragmatic principles. *Psychology Review*, *98*(2), 182-203.

Brainerd, C. J., & Reyna, V. F. (2001). Fuzzy-trace theory: Dual processes in memory, reasoning, and cognitive neuroscience. In H. W. Reese & R. Kail (Eds.), *Advances in child development and behavior* (pp. 41–100). San Diego, CA: Academic Press.

Bransford, J., Sherwood, R., Vye, N., & Rieser, J. (1986). Teaching thinking and problem solving: Research foundations. *American Psychologist*, *41*(10), 1078.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on*

Psychological Science, *6*(1), 3–5.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*(1), 116.

Carroll, J. M., & Kay, D. S. (1988). Prompting, feedback and error correction in the design of a scenario machine. *International Journal of Man-Machine Studies*, *28*(1), 11–27.

Chandler, D., & Kapelner, A. (2013). Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior and Organization*, *90*, 123–133.

Chapman, L. J., & Chapman, A. P. (1959). Atmopshere effect re-examined. *Journal of Experimental Psychology*, *58*, 220-226.

Chater, N., & Oaksford, M. (1999). The Probability Heuristics Model of Syllogistic Reasoning. *Cognitive Psychology*, *258*, 191–258.

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.

Cheshire, A., Ball, L. J., & Lewis, C. (2005). Self-explanation, feedback and the development of analogical reasoning skills: Microgenetic evidence for a metacognitive processing account. In *Proceedings of the twenty-second annual conference of the cognitive science society* (pp. 435–41).

Clifford, S., & Jerit, J. (2014). Is there a cost to convenience? an experimental comparison of data quality in laboratory and online studies. *Journal of Experimental Political Science*, *1*(02), 120–131.

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

De Neys, W. (2012). Bias and Conflict: A Case for Logical Intuitions. *Perspectives on Psychological Science*, *7*(1), 28–38.

De Neys, W. (2013). Conflict detection, dual processes, and logical intuitions: Some clarifications. *Thinking and Reasoning*, *1*, 1–19.

De Neys, W., & Bonnefon, J.-F. (2013). The whys and whens of individual differences in thinking biases. *Trends in Cognitive Sciences*, 1–7.

De Neys, W., & Franssens, S. (2009). Belief inhibition during thinking: not always winning but at least taking part. *Cognition*, *113*(1), 45–61.

De Neys, W., Moyens, E., & Vansteenwegen, D. (2010). Feeling we're biased: autonomic arousal and reasoning conflict. *Cognitive, Affective and Behavioral Neuroscience*, *10*(2), 208–16.

DeStefano, F. (2007). Vaccines and autism: Evidence does not support a causal association. *Clinical Pharmacology and Therapeutics*, *82*(6), 756–759.

Dicksten, L. S. (1975). The effects of figure on syllogistic reasoning. *Journal of Experimental Psychology: Human Learning and Memory*, *104*, 376-384.

Doja, A., & Roberts, W. (2006). Immunizations and autism: A review of the literature. *The Canadian Journal of Neurological Sciences*, *33*, 341–346.

Dube, C., Rotello, C. M., & Heit, E. (2010). Assessing the belief bias effect with ROCs: it's a response bias effect. *Psychological Review*, *117*(3), 831–63.

Dube, C., Rotello, C. M., & Heit, E. (2011). The belief bias effect is aptly named: a reply to Klauer and Kellen (2011). *Psychological Review*, *118*(1), 155–63.

Elqayam, S., & Evans, J. S. B. T. (2011). Subtracting "ought" from "is": descriptivism versus normativism in the study of human thinking. *The Behavioral and Brain Sciences*, *34*(5), 233–48; discussion 249–90.

Epstein, S., Pacini, R., Denes-Raj, V., & Heier, H. (1996). Individual differences in intuitive–experiential and analytical–rational thinking styles. *Journal of Personality and Social Psychology*, *71*(2), 390.

Eriksson, K., Simpson, B., et al. (2010). Emotional reactions to losing explain gender differences in entering a risky lottery. *Judgment and Decision Making*, *5*(3), 159–163.

Evans, J. S. B. T. (1989). *Bias in human reasoning: Causes and consequences*. London: Erlbaum.

Evans, J. S. B. T. (2000). Thinking and believing. In J. Garcia-Madruga, N. Carriedo, & M. J. Gonzales-Labra (Eds.), *Mental models in reasoning* (pp. 41–55). Madrid, Spain: Universidad Nacional de Educacio n a Distancia.

Evans, J. S. B. T., Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory and Cognition*, *11*(3), 295–306.

Evans, J. S. B. T., & Curtis-Holmes, J. (2005). Rapid responding increases belief bias: Evidence for the dual-process theory of reasoning. *Thinking and Reasoning*, *11*(4), 382–389.

Evans, J. S. B. T., Handley, S., & Harper, C. (2001). Necessity, possibility and belief: A study of syllogistic reasoning. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, 37–41.

Evans, J. S. B. T., Jonathan, Handley, S. J., Harper, C. N. J., & Johnson-Laird, P. N. (1999). Reasoning about necessity and possibility: A test of the mental model theory of deduction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(6), 1495–1513.

Evans, J. S. B. T., Newstead, S. E., & Allen, J. (1994). Debiasing by instruction: The case of belief bias. *European Journal of Cognitive Psychology*, *6*, 263–285.

Evans, J. S. B. T., & Pollard, P. (1990). Belief bias and problem complexity in deductive reasoning. *Advances in Psychology*, *68*, 131–154.

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science*, *8*(3), 223–241.

Fischhoff, B. (1975). Hindsight not equal to foresight: the effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology*, *12*(4), 304–12.

Fong, G. T., Krantz, D. H., & Nisbett, R. E. (1986). The effects of statistical training on thinking about everyday problems. *Cognitive Psychology*, *18*(3), 253–292.

Ford, M. (1995, January). Two modes of mental representation and problem solution in syllogistic reasoning. *Cognition*, *54*(1), 1–71.

Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of*

Economic Perspectives, *19*(4), 25–42.

Gorman, M. E., & Gorman, M. E. (1984). A comparison of disconfirmatory, confirmatory and control strategies on wason's 2–4–6 task. *The Quarterly Journal of Experimental Psychology*, *36*(4), 629–648.

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *The American Psychologist*, *59*(2), 93–104.

Handley, S. J., Newstead, S. E., & Trippas, D. (2011). Logic, beliefs, and instruction: a test of the default interventionist account of belief bias. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *37*(1), 28–43.

Handley, S. J., & Trippas, D. (2015). Dual processes, knowledge, and structure: A critical evaluation of the Default Interventionist account of biases in reasoning and judgment. *Psychology of Learning and Motivation*, *62*, 34–75.

Hauser, D. J., & Schwarz, N. (2015). Attentive turkers: Mturk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, 1–8.

Hayes, B. K., Heit, E., & Rotello, C. M. (2014). Memory, reasoning, and categorization: parallels and common mechanisms. *Frontiers in Psychology*, 529.

Heijltjes, A., Gog, T. V. A. N., & Paas, F. (2014). Improving Students Critical Thinking : Empirical Support for Explicit Instructions Combined with Practice. , *530*, 518–530.

Heim, A. W. (1970). *Group test of general intelligence*. UK: NFER-Nelson, Windsor.

Heit, E., & Rotello, C. M. (2014). Traditional difference-score analyses of reasoning are flawed. *Cognition*, 1–59.

Hilbert, M., & López, P. (2011). The worlds technological capacity to store, communicate, and compute information. *Science*, *332*(6025), 60–65.

Hoch, S. J. (1985). Counterfactual reasoning and accuracy in predicting personal events. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*, 719-731.

Hogarth, R. M., Gibbs, B. J., McKenzie, C. R., & Marquis, M. A. (1991). Learning from feedback: exactingness and incentives. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(4), 734.

Houdé, O., Zago, L., Mellet, E., Moutier, S., Pineau, A., Mazoyer, B., & Tzourio-Mazoyer, N. (2000). Shifting from the perceptual brain to the logical brain: the neural impact of cognitive inhibition training. *Journal of cognitive neuroscience*, *12*(5), 721–8.

Jefferson, T., Price, D., Demicheli, V., & Bianco, E. (2003). Unintended events following immunization with mmr: a systematic review. *Vaccine*, *21*(25), 3954–3960.

Johnson-Laird, P. N. (1975). Models of Deduction. In *Reasoning, representation and process in children and adults* (pp. 7–54). New York: John Wiley and Sons.

Johnson-Laird, P. N., & Bara, B. G. (1984). Syllogistic inference. *Cognition*, *16*.

Johnson-Laird, P. N., & Byrne, R. (1991). *Deduction.* Hillsdale, NJ: Erlbaum.

Johnson-Laird, P. N., & Steedman, M. (1978). The psychology of syllogisms. *Cognitive Psychology*, *10*(1), 64–99.

Kahneman, D., Slovic, P., & Tverky, A. (1982). *Judgement under uncertainty: heuristics and biases.* Cambridge: Cambridge University Press.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*, 237–252.

Kelley, C. M., & McLaughlin, A. C. (2012). Individual differences in the benefits of feedback for learning. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *54*(1), 26–35.

Khemlani, S. S., & Johnson-Laird, P. N. (2012). Theories of the syllogism: A meta-analysis. *Psychological Bulletin*, 1–31.

Klauer, K. C., & Kellen, D. (2011). Assessing the belief bias effect with ROCs: reply to Dube, Rotello, and Heit (2010). *Psychological review*, *118*(1), 164–73.

Klauer, K. C., Musch, J., & Naumer, B. (2000). On belief bias in syllogistic reasoning. *Psychological Review*, *107*(4), 852–884.

Kluger, A. N., & Denisi, A. (1996). The Effects of Feedback Interventions on Performance : A Historical Review , a Meta-Analysis , and a Preliminary Feedback Intervention Theory. *Psychological Bulletin*, *2*(2), 254–284.

Komaki, J., Heinzmann, A. T., & Lawson, L. (1980). Effect of training and feedback: component analysis of a behavioral safety program. *Journal of Applied Psychology*, *65*(3), 261.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 107-118.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236.

Kuhn, D., & Udell, W. (2007). Coordinating own and other perspectives in argument. *Thinking and Reasoning*, *13*, 90–104.

Larrick, R. P. (2004). Debiasing. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgement and decision making.* UK: Blackwell.

Larrick, R. P., Morgan, J. N., & Nisbett, R. E. (1990). Teaching the use of cost-benefit reasoning in everyday life. *Psychological Science*, *1*(6), 362–370.

Leighton, J. P. (2006). Teaching and assessing deductive reasoning skills. *The Journal of Experimental Education*, *74*(2), 107–136.

Lilienfeld, S. O., Ammirati, R., & Landfield, K. (2009). Giving Debiasing Away: Can Psychological Research on Correcting Cognitive Errors Promote Human Welfare? *Perspectives on Psychological Science*, *4*(4), 390–398.

Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: a corrective strategy for social judgment. *Journal of Personality and Social Psychology*, *47*(6), 1231–43.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*(11), 2098–2109.

Maclaren, V. V., Fugelsang, J. A., Harrigan, K. A., & Dixon, M. J. (2012). Effects of impulsivity , reinforcement sensitivity , and cognitive style on Pathological Gambling symptoms among frequent slot machine players. *Personality and Individual Differences*, *52*(3), 390–394.

Macpherson, R., & Stanovich, K. E. (2007). Cognitive ability, thinking dispositions, and instructional set as predictors of critical thinking. *Learning and Individual Differences*, *17*, 115-127.

Markovits, H., & Handley, S. (2005). Is inferential reasoning just probabilistic reasoning in disguise? *Memory and Cognition*, *33*, 1315-1323.

Markovits, H., & Nantel, G. (1989). The belief-bias effect in the production and evaluation of logical conclusions. *Memory and cognition*, *17*(1), 11–7.

Mevel, K., Poirel, N., Rossi, S., Cassotti, M., Simon, G., Houdé, O., & De Neys, W. (2014). Bias detection: Response confidence evidence for conflict sensitivity in the ratio bias task. *Journal of Cognitive Psychology*, *27*(February), 227–237.

Meyer, A., Frederick, S., Burnham, T. C., Guevara Pinto, J. D., Boyer, T. W., Ball, L. J., ... Schuldt, J. P. (2015). Disfluent fonts don't help people solve math problems. *Journal of Experimental Psychology: General*, *144*(2), e16.

Mikulincer, M., Glaubman, H., Ben-Artzi, E., & Grossman, S. (1991). The cognitive specificity of learned helplessness and depression deficits: The role of self-focused cognitions. *Anxiety Research*, *3*(4), 273–290.

Milkman, K. L., Chugh, D., & Bazerman, M. H. (2009). How Can Decision Making Be Improved? *Perspectives on Psychological Science*, *4*, 379–383.

Monaghan, P., & Stenning, K. (2004). Generalising Individual Differences and Strategies Across Different Deductive Reasoning Domains. In D. Hardman & L. Macchi (Eds.), . John Wiley and Sons.

Moutier, S., & Angeard, N. (2010). Deductive reasoning and matching-bias inhibition training : Evidence from a debiasing paradigm. *Thinking and Reasoning*, 37–41.

Moutier, S., Angeard, N., & Houdé, O. (2002). Deductive reasoning and matching-bias inhibition training: Evidence from a debiasing paradigm. *Thinking and reasoning*, *8*(3), 205–224.

Moutier, S., & Houdé, O. (2003). Judgement under uncertainty and conjunction fallacy inhibition training. *Thinking and Reasoning*, *9*(3), 185–201.

Newstead, S. E., Handley, S. J., Harley, C., Wright, H., & Farrelly, D. (2004). Individual differences in deductive reasoning. *The Quarterly Journal of Experimental Psychology: A, Human Experimental Psychology*, *57*(1), 33–60.

Newstead, S. E., Pollard, P., Evans, J. S. B. T., & Allen, L. J. (1992). The Source of Belief Bias in Syllogistic Reasoning. *Cognition*, *45*, 257–284.

Newton, E. J., & Roberts, M. J. (2000). An experimental study of strategy development. *Memory and Cognition*, *28*(4), 565–573.

Oakhill, J., & Johnson-Laird, P. N. (1985). The Effects of Belief on the Production of Syllogistic Conclusions. *Quarterly Journal of Experimental Psychology*, *37A*, 553–569.

Oakhill, J., & Johnson-Laird, P. N. (1989). Believability and syllogistic reasoning.

Cognition, *31*, 117–140.

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*(6), 972.

Paolacci, G., Chandler, J., & Stern, L. N. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*(5), 411–419.

Peer, E., Vosgerau, J., & Acquisti, A. (2013). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods*.

Pennycook, G., Trippas, D., Handley, S. J., & Thompson, V. A. (2014). Base rates: Both neglected and intuitive. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *40*(2), 544–54.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. New York: Springer.

Polk, T. a., & Newell, A. (1995). Deduction as verbal reasoning. *Psychological Review*, *102*, 533–566.

Prowse-Turner, J. A., & Thompson, V. A. (2009). The role of training, alternative models, and logical necessity in determining confidence in syllogistic reasoning. *Thinking and Reasoning*, *15*(1), 69–100.

Quayle, J. D., & Ball, L. J. (2000). Working memory, metacognitive uncertainty, and belief bias in syllogistic reasoning. *The Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology*, *53*(4), 1202–23.

Renner, C. H., & Renner, M. J. (2001). But i thought i knew that: Using confidence estimation as a debiasing technique to improve classroom performance. *Applied Cognitive Psychology*, *15*(1), 23–32.

Revlin, R., Leirer, V., Yopp, H., & Yopp, R. (1980). The belief-bias effect in formal reasoning : The influence of knowledge on logic. *Memory and Cognition*, *8*(6), 584–592.

Revlis, R. (1975). Two models of syllogistic reasoning: Feature selection and conversion. *Journal of Verbal Learning and Verbal Behavior*, *14*, 180–195.

Rips, L. J. (1994). *The psychology of proof: Deductive reasoning in human thinking*. Cambridge,MA: MIT Press.

Rips, L. J. (2001). Two kinds of reasoning. *Psychological Science*, *12*, 129-134.

Rips, L. J. (2002). Circular reasoning. *Cognitive Science*, *26*, 767-795.

Roberts, M. J., & Newton, E. J. (2003). Individual Differences in the Development of Reasoning Strategies. In D. Hardman & E. J. Newton (Eds.), *Thinking: Psychological perspectives on reasoning, judgment and decision making*. Wiley.

Roberts, M. J., & Sykes, D. A., Elizabeth. (2003). Belief bias and relational reasoning. *The Quarterly Journal of Experimental Psychology. A: Human Experimental Psychology*, *56*(1), 131–53.

Rotello, C. M., Heit, E., & Way, H. (2014). The neural correlates of belief bias : activation in inferior frontal cortex reflects response rate differences. *Frontiers in Human Neuroscience*.

Santamaría, C., García-Madruga, J. A., & Johnson-Laird, P. N. (1998). Reasoning

From Double Conditionals : The Effects of Logical Structure and Believability. *Thinking and Reasoning*, *4*, 97–122.

Sargis, E. G., Skitka, L. J., & McKeever, W. (2013). The internet as psychological laboratory revisited: Best practices, challenges, and solutions. In Y. Amichai-Hamburger (Ed.), *The social net: Understanding our online behavior.* Oxford University Press.

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology*, *51*(3), 515.

Shannon, C., & Weaver, W. W. (1949). *The mathematical theory of communication.* University of Illinois Press, Urbana, IL.

Siegler, R. S. (1995). How does change occur: A microgenetic study of number conservation. *Cognitive Psychology*, *28*(3), 225–273.

Siegler, R. S., & Chen, Z. (1998). Developmental Differences in Rule Learning : A Microgenetic Analysis. *Cognitive Psychology*, *310*(36), 273–310.

Sloman, S. A. (1996). The Empirical Case for Two Systems of Reasoning. *Psychological Bulletin*, *119*(1), 3–22.

Slovic, P., & Fischhoff, B. (1977). On the psychology of experimental surprises. *Journal of Experimental Psychology: Human Perception and Performance*, *3*, 544–551.

Solcz, S. (2011). *Not All Syllogisms Are Created Equal: Varying Premise Believability Reveals Differences Between Conditional and Categorical Syllogisms* (Unpublished doctoral dissertation). University of Waterloo, Ontario, Canada.

Sprouse, J. (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods*, *43*(1), 155–67.

Stanovich, K. E. (2009). Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory. In J. S. B. Evans & K. E. Frankish (Eds.), *In two minds: Dual processes and beyond.* Oxford University Press.

Stanovich, K. E., & Stanovich, P. J. (2010). A framework for critical thinking, rational thinking, and intelligence. In D. D. Preiss (Ed.), *Innovations in educational psychology: Perspectives on learning, teaching and human development* (pp. 195–237). Springer New York, NY.

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*(2), 342–357.

Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of Experimental Psychology: General*, *127*(2), 161–188.

Stanovich, K. E., & West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology*, *94*(4), 672–95.

Stenning, K., & Oberlander, J. (1997). A cognitive theory of graphical and linguistic reasoning: logic and implementation. Cognitive Science. , 97–140.

Stenning, K., & Yule, P. (1997). Image and language in human reasoning: a

syllogistic illustration. *Cognitive Psychology*, *34*, 109159.

Stupple, E. J. N., & Ball, L. J. (2008). Belieflogic conflict resolution in syllogistic reasoning: Inspection-time evidence for a parallel-process model. *Thinking and Reasoning*, *14*(2), 168–181.

Stupple, E. J. N., & Ball, L. J. (2014). The intersection between Descriptivism and Meliorism in reasoning research: further proposals in support of 'soft normativism'. *Frontiers in Psychology*, *5*, 1269.

Stupple, E. J. N., Ball, L. J., & Ellis, D. (2012). Matching bias in syllogistic reasoning: Evidence for a dual-process account from response times and confidence ratings. *Thinking and Reasoning*, 1–24.

Stupple, E. J. N., Ball, L. J., Evans, J. S. B. T., & Kamal-Smith, E. (2011). When logic and belief collide: Individual differences in reasoning times support a selective processing model. *Journal of Cognitive Psychology*, *23*(8), 931–941.

Taylor, L. E., Swerdfeger, A. L., & Eslick, G. D. (2014). Vaccines are not associated with autism: An evidence-based meta-analysis of case-control and cohort studies. *Vaccine*, *32*(29), 3623–3629.

Thompson, V. A., & Johnson, S. C. (2014). Conflict, metacognition, and analytic thinking. *Thinking and Reasoning*, 1–30.

Thompson, V. A., Morley, N. J., & Newstead, S. E. (2011). Methodological and theoretical issues in belief-bias: Implications for dual process theories. In K. I. Manktelow, D. E. Over, & S. Elqayam (Eds.), *The science of reason: A festschrift for Jonathan St. B. T Evans* (pp. 309–338). Hove: Psychology Press.

Thompson, V. A., Prowse, J. A., & Pennycook, G. (2011). Intuition, Reason, and Metacognition. *Cognitive psychology*, *63*(3), 107–40.

Thompson, V. A., Prowse-Turner, J. A., Pennycook, G., Ball, L. J., Brack, H., Ophir, Y., & Ackerman, R. (2013). The role of answer fluency and perceptual fluency as metacognitive cues for initiating analytic thinking. *Cognition*, *128*(2), 237–51.

Thompson, V. A., Striemer, C. L., Reikoff, R., Gunter, R. W., & Campbell, J. I. D. (2003). Syllogistic reasoning time: disconfirmation disconfirmed. *Psychonomic Bulletin and Review*, *10*, 184–9.

Thompson, V. A., Turner, J. A. P., Pennycook, G., Ball, L. J., Brack, H., Ophir, Y., & Ackerman, R. (2013). The role of answer fluency and perceptual fluency as metacognitive cues for initiating analytic thinking. *Cognition*, *128*(2), 237–251.

Toplak, M. E., & Stanovich, K. E. (2003). Associations between myside bias on an informal reasoning task and amount of post-secondary education. *Applied Cognitive Psychology*, *17*(7), 851–860.

Toplak, M. E., West, R. F., & Keith, E. (2013). Assessing miserly information processing : An expansion of the Cognitive Reflection Test. *Thinking and Reasoning*, *20*(2), 1–22.

Torrens, D., Thompson, V. A., & Cramer, K. M. (1999). Individual Differences and the Belief Bias Effect : Mental Models , Logical Necessity , and Abstract

Reasoning. *Thinking and Reasoning*, *5*(1), 1–28.

Trippas, D., Handley, S. J., & Verde, M. F. (2013). The SDT model of belief bias: complexity, time, and cognitive ability mediate the effects of believability. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *39*(5), 1393–402.

Trippas, D., Handley, S. J., & Verde, M. F. (2014). Fluency and belief bias in deductive reasoning: new indices for old effects. *Frontiers in Psychology*, *5*, 631.

Trippas, D., Pennycook, G., Verde, M. F., & Handley, S. J. (2015). Better but still biased: Analytic cognitive style and belief bias. *Thinking and Reasoning*, 1–15.

Trippas, D., Verde, M. F., & Handley, S. J. (2014). Using forced choice to test belief bias in syllogistic reasoning. *Cognition*, *133*(3), 586–600.

Trippas, D., Verde, M. F., & Handley, S. J. (2015). Alleviating the concerns with the SDT approach to reasoning: reply to Singmann and Kellen (2014). *Frontiers in Psychology*(2014).

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124-1131.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293.

Vollmeyer, R., & Rheinberg, F. (2005). A surprising effect of feedback on learning. *Learning and Instruction*, *15*(6), 589–602.

Wagenaar, W., & Keren, G. B. (1986). Does the expert know? The reliability of predictions and confidence ratings of experts. In E. Hollnagel, G. Mancini, & D. Woods (Eds.), *Intelligent decision support in process environments* (pp. 87–103). Berlin: Springer-Verlag.

Wakefield, A. J., Murch, S. H., Anthony, A., Linnell, J., Casson, D., Malik, M., ... Harvey, P. (1998). Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. *The Lancet*, *351*(9103), 637–641.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, *12*, 129-140.

Wetherick, N. E., & Gilhooly, K. J. (1995). Atmosphere, matching, and logic in syllogistic reasoning. *Current Psychology*, *14*, 169–178.

Woodworth, R. S., & Sells, S. B. (1935). An atmosphere effect in syllogistic reasoning. *Journal of Experimental Psychology*(18), 451-460.