# Deviations from rational beliefs:
# An investigation combining psychological and experimental economics approaches

Leslie van der Leer

Thesis submitted in fulfilment of the degree of
Doctor of Philosophy

Department of Psychology
Royal Holloway, University of London

March, 2015

# Declaration of Authorship

I, Leslie van der Leer, hereby declare that this work was carried out in accordance with the Regulations of the University of London. I declare that this submission is my own work, and does not represent the work of others, published or unpublished, except where duly acknowledged in the text. No part of this thesis has been submitted for a higher degree at another university or institution.

Signed Leslie

Date 8 March, 2015

# Abstract

This thesis investigates various deviations from rational beliefs by combining methods from psychology and experimental economics.

The first two studies focused on the jumping-to-conclusions bias, where delusional and delusion-prone individuals tend to make decisions based on less data than controls. In an incentivised and adapted "beads task" probability-reasoning paradigm, the effects of delusion-proneness on decisions and on probability ratings were investigated. All participants, but especially more delusion-prone participants, made their decisions too early. Moreover, high delusion-prone participants' probability ratings were less affected by incentives than low delusion-prone participants'.

The same paradigm was used to explore an inaccurate, but potentially evolutionarily advantageous, belief: the sexual over-perception bias, where men perceive more sexual interest in women's behaviour than women report or perceive. No evidence was found for men's over-perception of a male character's appeal to women in a belief-updating paradigm, which may reflect conceptual and methodological limitations of previous work on this topic.

Perhaps, people deviate from rationality for certain purposes (e.g., evolutionary goals), while also holding an accurate, rational belief. The fourth study examined whether people are, at some level, aware that their optimistic beliefs are inaccurate, by combining two distinct belief-updating paradigms. Participants provided repeated answers to neutral questions and questions about undesirable future outcomes. Participants were equally accurate for neutral items, but were even more optimistic on the second guess for undesirable items, suggesting that optimism involves "real" self-deception.

The last study investigated another phenomenon where people may want to avoid undesirable information. Investors are less willing to invest when playing the trust game with another player than when playing a computerised lottery with the same odds of the outcomes, which suggests that observing potential betrayal carries an additional, emotional cost. It was found that beliefs about others' trustworthiness could predict the level of such betrayal aversion.

# Dissemination of findings

## *Publications*

Van der Leer, L., & McKay, R. (under review). The optimist within. *Journal of Experimental Psychology: General.*

Van der Leer, L., Hartig, B., Goldmanis, M., & McKay, R. (in press). Delusion-proneness and jumping to conclusions: Relative and absolute effects. *Psychological Medicine.*

Van der Leer, L., & McKay, R. (2014). "Jumping-to-conclusions" in delusion-prone participants: An experimental economics approach. *Cognitive Neuropsychiatry, 19*(3), p. 257-267.

## *Conference presentations*

Van der Leer, L., Hartig, B., Goldmanis, M., & McKay, R. (2013) Do delusion-prone participants jump to conclusions? *Poster presentations at NeuroPsychoEconomics conference (Bonn, Germany) & Society for NeuroEconomics Annual Meeting conference (Lausanne, Switzerland)*

Van der Leer, L. (2012) Priors and prizes: A further investigation of the jumping-to-conclusions bias in delusion-prone participants. *Oral presentation at the Postgraduate Convention (RHUL, Egham, United Kingdom)*

# Acknowledgements

I would first and foremost like to thank my supervisor, Ryan McKay. He shared his wealth of knowledge, provided invaluable feedback, created many opportunities, and gave tremendous amounts of support both with regards to the PhD and for life in general. Besides Ryan's help, I am also grateful to Bjoern Hartig, for his constant advice regarding difficulties posed by z-Tree and by experimental economics requirements. I also thank Maris Goldmanis for calculating the optimal decision points for the study in Chapter 2. The studies in this thesis would not have been possible without their help. Furthermore, financial support was provided by a College Research Scholarship from Royal Holloway and from the Postgraduate Research Student Study Costs Scheme Award from the University of London.

In addition, I would like to thank my friends and fellow PhD students, both in the Psychology Department at Royal Holloway and beyond, for their support and friendship. A special thank you goes out to Caroline Durlik, whose friendship has enriched my life in so many ways and whose endless support throughout the last three years has kept me going through the tougher moments.

I am grateful to my parents for the comfort, help, and support they have offered throughout the years. Lastly, I want to thank David Davies for his love, patience, and understanding, and his endless ways of making me laugh.

# Table of contents

# List of tables

# List of figures

# 1 General Introduction

Across a wide range of domains, people tend to hold beliefs that are not supported by evidence. These beliefs can have disastrous consequences, and yet they persist, despite accumulation of evidence against them, making the beliefs more and more irrational. One such belief is the denial of climate change, where people are not concerned about global warming, either because they reject scientific evidence it is occurring or because they are overly optimistic that the consequences will be minor. This belief is found despite the occurrence of increasingly extreme weather conditions, such as hurricanes, droughts, and floods (Varki & Brower, 2013). Another example is the belief that the measles, mumps, and rubella (MMR) vaccine is dangerous (e.g., it is believed to cause autism). This leads to reduced immunisation rates, posing a threat to public health. Despite scientific discrediting of these beliefs and attempts to campaign for vaccinations, the beliefs about the dangers of the MMR vaccine remain (Nyhan, Reifler, Richey, & Freed, 2014). A further common example of beliefs that are not fully supported by evidence is the phenomenon of unrealistic optimism, where people underestimate their chances of experiencing unfortunate events. This could lead to increases in dangerous behaviours, such as smoking and unsafe sex (Sharot, 2011a). A related phenomenon is overconfidence, which involves overestimating one's personal qualities, which can lead to financial recessions brought on by too many risky investments, or even wars (D. D. P. Johnson & Fowler, 2011). Furthermore, men tend to believe that women are more sexually interested in them than women really are, which could lead to unwanted sexual advances, if not sexual assaults (Farris, Treat, Viken, & McFall, 2008b). Some irrational beliefs, such as delusions, are less common in the general population, but represent first-rank symptoms of disorders such as schizophrenia, which affects approximately 1% of the population  and has major consequences for society in terms of loss of functioning and increased costs of mental health care (American Psychiatric Association, 2013; Coltheart, Langdon, & McKay, 2011).

This thesis examines a selection of such deviations from rational beliefs. Some of the deviations from rational beliefs have been studied extensively, while others are more recently discovered, or hypothesised, phenomena. Across all studies, the main aim is to study the deviation from rationality with increased methodological rigour, by combining psychological and experimental economics approaches.

In this introductory chapter, the concepts of rationality and deviations from it are first defined. Next, a continuum of deviations from rationality is specified, ranging from deviations that are observed in select (e.g., clinical) groups to deviations that are common within the general population. Then, considering the focus on combining psychological and experimental economics practices, some differences between these practices and the implications for this thesis are discussed. The introductions of the empirical chapters that follow will focus on the most relevant theories and rationales of the studies presented therein, but the general overview of topics is provided in this chapter.

## 1.1 Beliefs, Rationality and Irrationality

There is no consensus about what constitutes a belief, but McKay and Dennett (2009) offer the following working definition: “a functional state of an organism that implements or embodies that organism’s endorsement of a particular state of affairs as actual” (p. 493).

A belief can be conceptualised in a binary or in a probabilistic, continuous form. Under a binary conception, one believes something ($p$) or does not believe something ($\neg p$). On the continuous conception of beliefs, a belief expresses the probability of a proposition being correct, with 1 representing absolute conviction that it is true and 0 representing absolute conviction that it is false (Caplin, Dean, Glimcher, & Rutledge, 2010; Schwitzgebel, 2014). This thesis adopts the probabilistic notion of beliefs.

Whereas binary beliefs can be correct or false, it is unclear what the notion of a false belief is on the probabilistic conception (McKay, 2012). A more useful notion, encompassing both binary and continuous conceptions, is that of "rationality", and by adopting the probabilistic conception of belief, I thus focus on whether beliefs are rational or irrational, rather than correct or false.[1] The distinction between rationality and irrationality is often disputed (Gigerenzer & Sturm, 2012). Rational beliefs must be consistent with other beliefs and intentions of the same person and they must be sensitive to available evidence and as such "conform to the best available standards of correct reasoning" (Bortolotti, 2009, p. 16). These best available standards are often considered to be provided by inductive or deductive logic rules (Gerrans, 2001). One often used logic rule is Bayes' theorem:

$$P(H|D) = \frac{P(D|H) \times P(H)}{P(D)} = \frac{P(D|H) \times P(H)}{P(D|H) \times P(H) + P(D|\neg H) \times P(\neg H)}$$

[1] On the binary conception, it is possible for beliefs to be rational, yet false. Consider, for example, the false-belief task (Wimmer & Perner, 1983), which later became known as the Sally-Anne task (Baron-Cohen, Leslie, & Frith, 1985), where a participant is told that Sally and Anne are in a room together as Sally places a marble into her basket. Sally then leaves and Anne moves the marble from the basket to a box. Participants are asked where Sally will look for the marble when she returns. Sally should look for her marble in the basket, where she would believe it to be: a rational, yet false, belief.

Where *P*(H|D) is the posterior probability of the hypothesis given the data, which integrates the probability of the data given the hypothesis (i.e., the likelihood of the data; *P*(D|H)), the prior probability of the hypothesis being true (*P*(H)) and the probability of the data (*P*(D)). The probability of the data is a combination of how likely the data is to be found if the hypothesis would be true and if it would not be true (Dienes, 2008). As an example, consider the beads task, which forms the basic paradigm for the studies in Chapters 2 and 3, and is used to investigate probabilistic reasoning (Phillips & Edwards, 1966). Participants are presented with two jars: one has more green than red beads, the other more red than green beads. Beads are drawn at random from one of the (now-hidden) jars and participants have to decide from which jar they are drawn (described in more detail later, at 1.3.1 on page 28). The prior probability pertains to the probability that either jar is selected, before seeing any evidence. The likelihood of the data refers to the probability of drawing a bead of a certain colour from either jar. This depends on the ratios of red and green beads in the jars. For example, in a jar with a ratio of 85 red: 15 green, the likelihood of drawing a red bead would be .85. The posterior probability is the belief one holds after seeing evidence. Bayes' theorem specifies the optimal procedure for arriving at this posterior probability, which is the probability used to decide which jar the beads are from.

Besides arriving at the normative probability (i.e., updating beliefs in accordance with Bayes' theorem), rationality requires an optimal consideration of the consequences of decisions made on the basis of beliefs. Often the probabilities of proposition *A* being more likely or less likely (more towards the opposite proposition $\neg A$) have different consequences, in terms of how rewarding they are expected to be if the proposition turns out to be true. For example, if climate change is man-made, this would represent a state of the world where compromises would have to be made. If climate change is not man-made, the state of the world would not require compromise and people might experience less guilt. This means that the state of the world where the proposition "climate change is man-made" is false is expected to be more rewarding than the state of the world where it is true. Here, the possible states of the world carry different expected rewards. Using this notion of expected rewards, Caplin et al. (2010) define beliefs as "the probabilities attached to the states of the world that would generate such [expected rewards]" (p. 953). This operationalisation of beliefs is used in this thesis and expected rewards are manipulated through incentives (see later, at 1.7.2 on page 73). Therefore, rationality in this thesis combines arriving at the probabilities suggested by logical reasoning (i.e., applying Bayes' theorem) and using these probabilistic beliefs to make optimal decisions (i.e., decisions that maximise expected rewards).[2]

As described above, under the probabilistic conception of beliefs, rationality requires that beliefs conform to standards of correct reasoning (Bortolotti, 2009). Deviations from rationality (i.e., irrational beliefs), then, are defined as a deviation from the probability that one *should* assign to the proposition in question according to such standards (e.g., Bayes' theorem, described earlier on page 18). Deviations from the normative probabilities, could arise due to departures from rational rules of inference (i.e., not applying Bayes' theorem correctly) or due to holding different prior probabilities or likelihoods

[2] Strictly speaking these decisions are only "optimal" from a risk-neutral perspective. Risk-seeking or risk-averse individuals may make decisions that fail to maximise their expected outcome, but that nevertheless maximise their expected utility (and thus are rational) given their risk preferences.

(Matthews, 2005). For example, people with irrational beliefs could hold a different prior belief than people with rational beliefs (e.g., McKay, 2012), so that the same amount of evidence (in terms of the likelihood ratio: $P(\mathrm{D}|\mathrm{H})/P(\mathrm{D}|\neg\mathrm{H})$) will lead to more updating of the posterior belief for one person than for another (Matthews, 2005).

It is often found that humans' decision-making deviates from that dictated by formal logic rules (Tversky & Kahneman, 1983, 1986). The irrational decisions are presumably based on irrational beliefs as Bortolotti (2009) argues that beliefs and behaviours should match.[3] These deviations are systematic rather than unsystematic; unsystematic errors may be due to performance errors, such as distraction during the task (Stanovich & West, 1998, 2000). As an example of a systematic error, consider the Linda problem: based on a description of a woman, Linda, who is "single, outspoken, and very bright" and "was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations [as a student]" (p. 297), 85% of participants thought it was more probable that she was a bank teller *and* an active feminist than that she was a bank teller (Tversky & Kahneman, 1983). This "conjunction fallacy" violates formal rules of logic, where the probability that she is a bank teller must be bigger than the probability that she is a bank teller and a feminist, as the former would include the latter. Another example is the framing effect, where people are risk-averse when the choice between a safe option and a risky option is framed in the gain domain (e.g., "200 people will be saved" versus "there is 1/3 probability that 600 people will be saved, and 2/3 probability that

[3] This match between behaviour and beliefs has been debated (e.g., McKay & Dennett, 2009), as discussed later in this thesis (e.g., at 1.2 on page 24). Especially in the case of delusions, the mismatch between beliefs and behaviour is considered to undermine the notion of delusions as beliefs.

no people will be saved"), but risk-loving in a loss domain (e.g., "400 people will die" versus "there is 1/3 probability that nobody will die, and 2/3 probability that 600 people will die"), although the safe and risky choices have equal expected values in both scenarios (Tversky & Kahneman, 1986). This switch from risk-averse to risk-loving behaviour, especially by the same participants, depending on the framing of the question appears irrational, and formal rules of logic argue that the language used should not affect the decisions made. Such findings have led some to argue that humans are not rational (Kahneman, 2003; Tversky & Kahneman, 1983, 1986).

A counter-argument to this stance is provided by defenders of bounded rationality, who emphasise that decision-making and judgment under uncertainty are influenced by the environment (e.g. limited time). Furthermore, some participants do show rational decisions and judgments, suggesting that individual differences (e.g., intelligence, knowledge, or computational power) might contribute to deviations from rationality (Stanovich & West, 1998, 2000). In the bounded-rationality view, rational behaviour is said to be shaped "by a [pair of] scissors whose two blades are the structure of task environments and the computational capabilities of the actor" (Simon, 1990, p. 7) and just as one cannot understand how scissors work by only looking at one blade, one cannot understand rationality by only considering the mind (Gigerenzer & Sturm, 2012). The ecological influence can lead to decisions deemed irrational when compared to normative rules, which assume unlimited time and cognitive resources and the application of the correct logic rules (Cosmides & Tooby, 1996; Gigerenzer & Goldstein, 1996; Gigerenzer & Sturm, 2012; Stanovich & West, 2000). Defenders of bounded rationality argue that irrational choice behaviour can often be ascribed to methodological artefacts and suggest that reasoning should be assessed with rules appropriate for both the content and context of the problem (Gigerenzer, 1996; Stanovich & West, 2000; Sturm, 2012). For example, due to ambiguous language in the Linda problem, participants might misrepresent the choice between her being “a bank teller” or “a bank teller and a feminist” as a choice between her being “a bank teller and not a feminist” or “a bank teller and a feminist”. The description of Linda would suggest the latter is more probable, and, thus, if this alternative interpretation is applied, the choice made is actually rational (Gigerenzer, 1996). This could be especially influential as the meaning of probability might be unclear to participants, as not all participants naturally take it to mean frequencies (Cosmides & Tooby, 1996; Wang, 2000). Indeed, when the Linda problem is presented as a question of frequencies (i.e., asking how many out of 200 women like Linda would be bank tellers and how many would be bank tellers and feminists), rather than of probabilities (i.e., asking the probability that Linda is a bank teller and the probability that she is a bank teller and a feminist), far fewer participants show the conjunction fallacy, arguably because they rely more on mathematical, rather than semantic, rules of inference (Hertwig & Gigerenzer, 1999). The framing problem could also have been misconstrued: participants might not take the words “will be saved” as an absolute number, but rather as a minimum number, but this is not how “will die” is interpreted; if this is the case, participants’ risk-preference reversal based on the framing would be rational (Stanovich & West, 2000).

Furthermore, as noted above, some researchers (e.g., Bortolotti, 2009) claim that rational beliefs should have matching behaviour. Others (e.g., Haselton & Buss, 2000) reason back from observed behaviour to beliefs, and infer biased beliefs from biased behaviour. Yet, biased behaviour and biased beliefs can occur jointly as well as separately from one another (Marshall, Trimmer, Houston, & McNamara, 2013). McKay and Dennett (2009) argue that biased behaviours can occur without biased beliefs. For example, although one may not have a strong belief that there is oncoming traffic, one may still check for it when crossing the street, in case there might be. Under uncertainty, nature may have maintained rational beliefs, but included a policy to behave in a way that is not in line with the rational beliefs, which may actually minimise costs (McKay & Dennett, 2009). Therefore, the inference from observed biased behaviour to biased beliefs might not be valid.

This thesis aims to focus on biased beliefs and most studies within this thesis measure whether probability estimates are systematically below or above a correct value, which Gigerenzer (2004) considers a common paradigm within social psychology. Any systematic deviation from the correct value is considered a deviation from rationality. The correct value can either be a value arrived at through Bayes theorem (Chapters 2, 3, and 4), a value taken from the literature (Chapter 5), or even a value determined by theoretical deduction (Chapter 6). Note that logic rules, such as Bayes' theorem or maximisation of expected value, are used to determine most of these correct values, or, in other words, the normative standard (Sturm, 2012).

## 1.2 Costs and Benefits of Different Types of Irrationality

As mentioned above, different states of the world tend to carry different rewards (Caplin et al., 2010). These different rewards of different states of the world could result in biased beliefs, biased behaviour, or both, in order to maximise reward and minimise costs (Marshall et al., 2013; McKay & Dennett, 2009). For example, someone with grandiose delusions might believe that they are able to fly, a belief that could be exciting and rewarding, but yet not translate this belief into behaviour, such as jumping off a building, in order to minimise costs to survival in case the belief is wrong. This might be an example of a biased belief without a biased behaviour. In the street-crossing example, the behaviour of checking for oncoming traffic, regardless of the strength of the belief regarding whether there is such oncoming traffic, minimises the potential costs to survival if one were to be hit by oncoming traffic. This might be an example of biased behaviour without biased beliefs. Finally, believing that climate-control is not man-made reduces psychological costs of guilt and minimises efficiency costs, or even financial costs, if one behaves in line with the belief by driving rather than taking public transport or by opting out of paying the carbon offsetting fee. This might be an example of a biased belief combined with a biased behaviour.

This thesis discusses several deviations from rationality, which carry different types and levels of costs. Overall, the balance of costs and benefits of irrationality in each case might explain why some irrational beliefs and behaviours are more common than others (e.g., uncommon deviations from rationality, such as delusions, are dysfunctional). I use such differences in the prevalence of different types of irrationality to loosely place the investigated deviations from rationality along a continuum of abnormal to normal irrationality. This continuum assumes the statistical meanings of the words “normal” and “abnormal”: “ordinary or usual” versus “different from what is usual or average” (Cambridge Dictionary, 2014).

## 1.3 Delusions and the Jumping-to-Conclusions Bias

Starting with an abnormal deviation from rationality, two studies within this thesis focus on the association between the so-called jumping-to-conclusions (JTC) bias (i.e., basing decisions on minimal evidence; further described under

1.3.1 on page 28) and the formation of delusions, which “are fixed beliefs that are not amenable to change in light of conflicting evidence” according to the fifth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-V; American Psychiatric Association, 2013, p. 87). Delusions display a wide variety of contents; for example, persecutory delusions are beliefs that one is going to be harmed by others (American Psychiatric Association, 2013). Delusions form a key feature of schizophrenia spectrum and other psychotic disorders (American Psychiatric Association, 2013). They also appear in bipolar disorder, depression, dementias, or after brain damage (Coltheart et al., 2011; Garety & Freeman, 2013).

Delusions carry costs to patients and their families in terms of suffering, employment difficulties, and social isolation, as well as costs to society in terms of healthcare, with specific concerns such as high relapse rates, potentially exacerbated by low adherence to medication, which, in turn, tends to be of only medium effectiveness (Knapp, 2005). Improved understanding of psychopathological conditions could potentially be achieved by focusing on individual symptoms rather than on diagnostic syndromes (Garety & Freeman, 2013). For example, a specific focus on delusions, one positive symptom within a diagnostic syndrome such as schizophrenia, has led to various theories regarding their formation and maintenance (Garety & Freeman, 1999), such as an impaired theory of mind (for a review see Brüne, 2005) or reasoning biases, such as a bias against disconfirmatory evidence (e.g., Speechley, Ngan, Moritz, & Woodward, 2012) or the JTC bias. In this thesis, I focus on the theory that basing a decision on minimal evidence (i.e., showing the JTC bias; Huq, Garety, & Hemsley, 1988; P. Taylor, Hutton, & Dudley, 2014) contributes to the formation and maintenance of delusions.

Psychotic symptoms, such as delusions, are also found in non-clinical populations with prevalence rates of up to 20% (Peters, 2010; van Os, Linscott, Myin-Germeys, Delespaul, & Krabbendam, 2009), and their presence forms a risk factor for the development of diagnosable disorders (e.g., Heriot-Maitland, Knight, & Peters, 2012; Kelleher et al., 2012). The presence of psychotic symptoms and delusional ideation in the general population has led to the idea that psychosis might exist on a continuum, with clinical psychosis and delusions at an extreme end of the spectrum (e.g., Freeman, Pugh, & Garety, 2008; Peters, Joseph, & Garety, 1999).

Considering the presence of delusional ideation in the general population and its potential precursory role in the formation and persistence of clinical delusions (Garety & Freeman, 2013), the investigation of reasoning processes in sub-clinical samples is highly relevant (L. O. White & Mansell, 2009). The knowledge gained from sub-clinical studies could potentially be used to inform intervention treatment programs, which in turn could be cost-effective compared to existing methods (e.g., McCrone, Craig, Power, & Garety, 2010). Although Andreou, Moritz, Veith, Veckenstedt, and Naber (2013) did not find an effect of dopamine agonists or dopamine antagonists on the JTC bias (i.e., basing decisions on minimal evidence), results from studies using patient samples could be more varied due to differences between participants in medication use, duration of illness, and other cognitive impairments (Colbert & Peters, 2002; Garety & Freeman, 1999). While the literature reviewed below encompasses findings from both clinical and non-clinical samples, the studies on the JTC bias within this thesis focus on delusion-proneness in non-clinical populations. This was done in order to investigate the association between delusional ideation and the JTC bias under rigorous experimental control and at an early time in the potential developmental trajectory of a disorder.

### 1.3.1 Jumping-to-Conclusions Bias and the Beads Task

As mentioned above, a consistently found reasoning bias is known as the jumping to conclusions (JTC) bias (Garety & Freeman, 1999, 2013). A JTC bias is reflected by decision making based on minimal evidence (P. Taylor et al., 2014). The JTC bias is thought to contribute to delusion formation and maintenance, perhaps because individuals with this bias adopt bizarre and unjustified beliefs on the basis of minimal evidence and with minimal consideration of alternative conclusions (Garety & Freeman, 2013). The JTC bias has been found for clinical patients with delusions (e.g., Huq et al., 1988) and healthy controls who are delusion-prone (e.g., Colbert & Peters, 2002; McKay, Langdon, & Coltheart, 2006), although some studies fail to find the JTC bias, even for delusional patients (e.g., McKay, Langdon, & Coltheart, 2007). It must also be noted that the results for comparisons between delusion-prone and non-delusion-prone participants are generally not as robust or as extreme as for comparisons between delusional and non-delusional participants (Freeman, 2007; Zawadzki et al., 2012). Yet, in a recent literature review investigating studies since 1999, Garety and Freeman (2013) found that 74% of clinical studies found an association between delusions or psychosis and the JTC bias, while 85% of non-clinical studies found an association between delusion-proneness and the JTC bias.

As mentioned earlier (under 1.1 on page 19), the JTC bias is often investigated using the beads task, which was initially used to investigate probabilistic reasoning in the general population (Phillips & Edwards, 1966). The beads task generally consists of the presentation of two jars, each containing beads of two colours in a specific ratio. The ratios in the two jars are usually complementary (cf. Speechley, Whitman, & Woodward, 2010). For example: jar A contains 85 red beads and 15 green beads, jar B contains 15 red beads and 85 green beads. The jars are hidden from view and the experimenter draws a series of beads, ostensibly at random, from one of the jars. In reality, this pseudo-random order is predetermined. The entire series of drawn beads is from the same jar and, although often not explicitly stated, the jar from which the beads are drawn is ostensibly chosen at random (i.e., both jars are equally likely to be chosen). The participant has to decide from which jar the beads are drawn. Once a decision has been made, the drawing is terminated and the participant gives a confidence rating for the decision made. The JTC bias is operationalised as participants high in delusional thinking using significantly fewer data compared to controls (Huq et al., 1988). A different operationalisation is based on a large proportion of delusional patients deciding after one or two draws (Garety & Freeman, 1999, 2013; Moritz & Woodward, 2005). Considering a decision as JTC at one or two draws could be considered relatively arbitrary (e.g., why is deciding after three jars not JTC), and, as such, I did not use this operationalisation within this thesis.

Based on different studies with different measures, the JTC bias has been suggested to involve several components: premature decisions after minimal data gathering (e.g., Colbert & Peters, 2002; Garety, Hemsley, & Wessely, 1991), over-confidence concerning the decision made (e.g., Huq et al., 1988; McKay et al., 2006), and over-adjustment after contradictory evidence (e.g., Garety et al., 1991).

These different components are commonly assessed through different versions of the beads task. In a draws-to-decision version, participants have to decide from which jar the series of beads is drawn. After each bead, the participants are asked if they are certain which jar the beads are drawn from or if they need another draw. The trial is ended once a decision has been made (Moritz & Woodward, 2005). This version is more sensitive to the premature-decision component of the JTC bias than other versions. Another option is to request probability estimates (or graded estimates on a Likert scale) that the drawn sequence is from either of the jars after each draw for a fixed number of draws (Moritz & Woodward, 2005). This version is more sensitive to over-adjustment of probabilities after conflicting evidence than other versions. This thesis adopts both a draws-to-decision version (Chapter 2) and a probability-estimates version (Chapter 3), to investigate both decisions and beliefs that potentially underpin decisions.

#### *1.3.1.1 Variations of the Beads Task*

The JTC bias, or its components, are consistently found despite variations in the beads task. Common variations include different ratios of beads in the jars and variations in task content.

The relatively easy 85:15 ratio has been used to prevent floor effects (i.e., immediate decisions) in patients (Garety & Freeman, 1999), but with more difficult ratios, such as 60:40, the JTC bias has also been found (e.g., Warman, Lysaker, Martin, Davis, & Haudenschield, 2007). Although most studies have used one ratio, some have included different conditions with different ratios (e.g., Dudley, John, Young, & Over, 1997b; Lincoln, Ziegler, Mehl, & Rief, 2010). In these different conditions, a main effect of condition is found, so that all participants draw more beads before deciding in a 60:40 ratio condition than in an 85:15 ratio condition. This effect merely reflects task difficulty, as it does not interact with delusional status and does not abolish the JTC bias, as delusional patients consider less evidence than healthy controls in conditions with both easy and difficult ratios. Recently, using large, student samples, Cafferkey, Murphy, and Shevlin (2014) found no association between delusion-proneness and the JTC bias on a task using the 85:15 ratio (based on n=140), but there was an association between delusion-proneness and the JTC bias using the 60:40 ratio (based on n=144). This suggests that in sub-clinical samples, as used in this thesis, the ratios might have to be more difficult to elicit a JTC bias.

In terms of content, the beads task provides a relatively pure measure of reasoning in delusional participants because of its neutral nature with respect to delusional topics (Warman et al., 2007). However, Woodward, Munz, LeClerc, and Lecomte (2009) have argued that the beads task might be too abstract, which could interfere with comprehension. More concrete and realistic, yet emotionally neutral, contents have been used, without affecting the JTC bias. For example, Dudley, John, Young, and Over (1997a) asked participants to judge whether names of students came from a school mainly for boys or from a school mainly for girls, with male and female names in 60:40 ratios. Others (Speechley et al., 2010; Whitman & Woodward, 2011; Woodward et al., 2009) have adapted the content to a fisherman presenting fish (cf. beads) from one of two lakes (cf. jars) with different ratios of differently coloured fish. These contents did not affect the JTC bias.

A different story might arise for information which is emotional or self-referent, which might increase salience and lead to a more extreme JTC bias. For example, in such studies, each participant was told that two groups of people had been asked to describe them. One survey group had supplied mostly positive descriptions (e.g., describing the participant as friendly) whereas the other had supplied mostly negative descriptions (e.g., describing the participant as impatient). A series of these trait descriptions were then drawn and participants had to decide from which of the two survey groups the descriptions were being drawn. Dudley et al. (1997a) asked participants to imagine the survey groups described someone ostensibly similar to the participant. The JTC bias found for delusional patients was similar in this emotionally salient condition and in the neutral condition, which involved the standard beads-in-jars content. Using the same content, however, Warman and Martin (2006) found an association between JTC and delusion-proneness only for the emotionally salient version, but not for the neutral beads task. Warman et al. (2007) asked participants to generate several comments about themselves or to select comments from a list of suggestions which they considered to be highly reflective of themselves. These comments were then used to construct the two surveys. On the draws-to-decision measure, Warman et al. (2007) found no difference in the JTC bias between their emotionally salient condition and the standard beads condition. However, an interaction between delusional status and task content was found for confidence levels: in the standard condition, delusional, delusion-prone, and non-delusion-prone participants did not differ, but in the emotionally salient condition, delusional patients were more confident about their decisions than the delusion-prone and non-delusion-prone individuals, who did not differ from each other. Fraser, Morrison, and Wells (2006) found that emotional content does lead to faster decisions, but does so equally for clinical groups and for healthy controls.

Overall, the above studies suggest that changing the ratios or the content of the beads task does not affect the JTC bias, especially if the alternative content is also emotionally neutral. Accordingly, the draws-to-decision study in this thesis used a single ratio. However, in the second JTC study, a probability-estimates version, we manipulated ratios to investigate probability reasoning. Both JTC studies adopted the fisherman scenario as it might be more relatable and engaging than beads in jars, while at the same time remaining neutral with respect to items on the delusional ideation measure we employed.

#### *1.3.1.2 Limitations of the Beads-Task Methodology*

The methodology of the beads task carries a few potential confounds of the JTC bias. The influence of miscomprehension, working memory deficits, probability reasoning deficits, and a lack of motivation are discussed below. The focus on these potential limitations does not imply that other factors, such as mood (e.g., Lee, Barrowclough, & Lobban, 2011) or anxiety (e.g., Lincoln, Lange, Buau, Exner, & Moritz, 2010), could not influence the JTC bias, but the selected factors are most relevant to the two JTC studies in this thesis.

##### *1.3.1.2.1 Miscomprehension*

It has been suggested that the irrational responses made on the beads task (i.e., deciding after one or two beads) could be due to poor task comprehension (Moritz & Woodward, 2005). In particular, people might think that jars are being swapped throughout the sequence or that they should base their judgment on each single bead rather than on the entire sequence. Misunderstanding could explain the over-adjustment and premature decisions components of the JTC bias.

Some studies have specifically investigated the effect of miscomprehension on the JTC bias. Balzan, Delfabbro, and Galletly (2012) found that more than half of their participants misunderstood the beads task and jumped from thinking it was the suggested jar to the other after a single piece of conflicting evidence. Of the participants who did not comprehend the task (i.e., those who gave higher ratings for the jar not suggested by the sequence seen so far), 37.8% made premature decisions (i.e., a "definite" rating after one bead) and 2.19% over-adjusted, whereas of participants who did comprehend the task only 11.4% decided prematurely and 0.41% over-adjusted. Qualitative analysis indicated that participants thought the jars were being swapped (Balzan, Delfabbro, & Galletly, 2012). When Balzan, Delfabbro, Galletly, and Woodward (2012) included a clinical schizophrenic group, they again found high levels of miscomprehension. They also found high levels of premature decisions and high levels of over-adjustment for miscomprehending compared to comprehending participants.

Including explicit instructions stating that all beads come from one and the same jar does not abolish the JTC bias, as delusional patients continue to make more premature decisions and over-adjust more than controls in studies with such explicit (and sometimes directive[4]) instructions (Balzan, Delfabbro, Galletly, et al., 2012; Garety et al., 1991; Woodward et al., 2009). Whitman, Menon, Kuo, and Woodward (2012) found that when participants were to select one of three lakes as the source of a collection of fish in a downstream lake, a relatively large number of participants did not select the most likely lake at an above-chance level. This also suggests participants do not comprehend the instructions as intended.

As the JTC bias is still found despite explicit instructions, it seems likely that this bias is related to delusional reasoning. However, to ensure task comprehension

[4] Directive, explicit instructions inform participants how to behave in response to disconfirming evidence: "participants were reminded that, in addition to changing containers completely upon presentation of a contrasting bead colour, they also had the option of changing their confidence within the same container (e.g., from 'very likely' to 'probably')" (Balzan, Delfabbro, Galletly, et al., 2012, p. 536).

in this thesis, stringent methods were adopted. First, explicit instructions and several comprehension questions, which had to be answered correctly in order to continue, were included to minimise the risk of miscomprehension. Second, miscomprehension regarding the swapping of lakes within a sequence was eliminated by presenting all fish at once (second task in Chapter 2) or by only presenting one fish (Chapter 3).

#### *1.3.1.2.2 Working Memory Deficits*

Related to miscomprehension are potential working memory deficits, which would impair the ability to maintain and manipulate information, such as the rules of the task and the sequence of beads presented (Dudley et al., 1997b). Associations between the JTC bias and poor working memory have been found for participants at risk of developing a psychotic disorder (Broome et al., 2007) and for delusional patients (Freeman et al., 2014; Garety et al., 2013). Ochoa et al. (in press) also found associations between working memory and the JTC bias for both schizophrenic patients and healthy controls.

Several studies have included a memory aid in the beads task, explicitly reminding participants of the ratios of the beads in the jars and showing the colours of the previously drawn sequence of beads. Most studies have found that schizophrenic or delusional patients still requested fewer beads than psychiatric and healthy controls, despite the presence of a memory aid (e.g., Dudley et al., 1997b; Lincoln, Ziegler, Mehl, et al., 2010; Moritz & Woodward, 2005; Moritz, Woodward, & Lambert, 2007), but Menon, Pomarol-Clotet, McKenna, and McCarthy (2006) found that group differences were no longer significant in the presence of a memory aid.

It could be argued that the memory aids only help in reducing the load of maintaining information, but not for the manipulation of presented information. Impaired working memory could lead to additional noise in decision-making processes, which, in turn, could potentially account for the JTC bias (Moutoussis, Bentall, El-Deredy, & Dayan, 2011). To minimise noise as much as possible, the studies on the JTC bias in the present thesis included memory aids in the form of visually and numerically presenting the ratios of black to white fish in either lake and visually presenting the sequence of previously-seen fish.

#### *1.3.1.2.3 Probability Reasoning Deficits*

Delusional or delusion-prone participants might have general probability reasoning deficits which lead them to decide early, rather than that they jump to conclusions because of a specific cognitive bias.

In various non-beads-task paradigms, delusional or delusion-prone participants provided similar probability ratings for stimuli as controls, although different decisions were made based on these similar probability ratings (see e.g., LaRocco & Warman, 2009; McGuire, Junginger, Adams Jr., Burright, & Donovick, 2001; Moritz, Woodward, & Hausmann, 2006). However, as this thesis uses the beads task, potential differences in probability reasoning and in decision-making within this paradigm are considered.

In their reviews of the JTC bias, measured through different dependent variables (e.g., draws-to-decision, probability estimates), Garety and Freeman (1999) and Fine, Gardner, Craigie, and Gold (2007) conclude that the general probability reasoning of delusional or delusion-prone participants is not impaired, especially not when considering neutral events. It must be noted that the general population provides conservative probability estimates on the beads task, so that after seeing a bead, the likelihood is not incorporated into the posterior probability sufficiently, which, as a result, is closer to the prior probability than Bayes' theorem would suggest (Phillips & Edwards, 1966). Investigations of whether such conservative probability estimates in the general population also lead to more "conservative" decision-making in the beads task (i.e., deciding on the basis of a lot of evidence) have, to the best of my knowledge, not been conducted. The study reported in Chapter 2 addresses whether healthy participants are conservative in a draws-to-decision version of the beads task.

Although probability reasoning appears unimpaired, the aforementioned differences in decision-making point to a possible explanation for the JTC bias: delusional or delusion-prone individuals might make decisions between competing hypotheses more easily compared with controls. For example, they might accept and act upon a hypothesis at a probability level of 78%, whereas controls might not consider this probability high enough to accept the hypothesis and might require a probability of 90%. This idea, known as the liberal acceptance account (e.g., Moritz et al., 2007; L. O. White & Mansell, 2009), states that probability reasoning in delusional or delusion-prone participants is not affected, but that they have a lower decision threshold for decisions, which can be reached after two beads, for example, and this would lead to the JTC bias (see Figure 1.1).

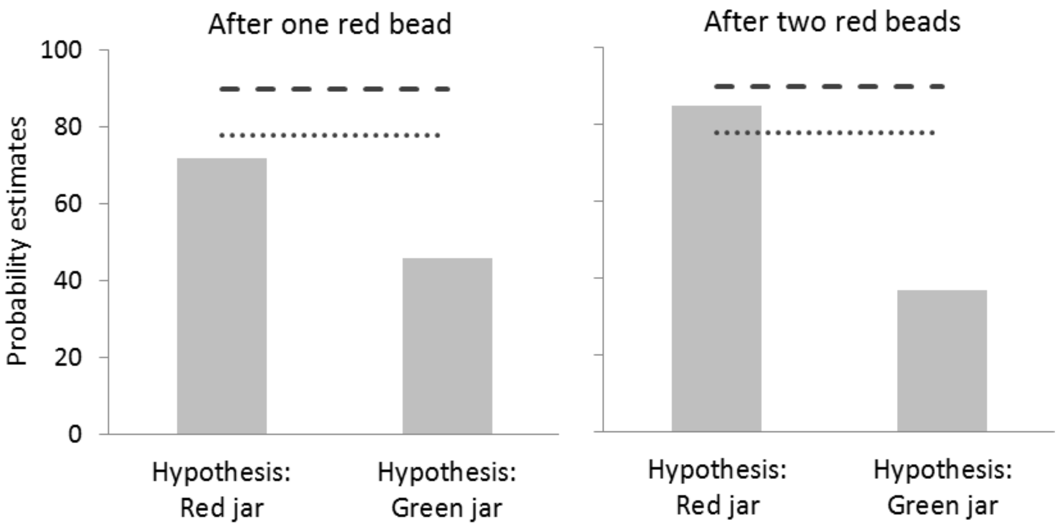


**Figure 1.1** The liberal acceptance of hypotheses account for the JTC bias (adapted from Fig. 2 in Moritz et al., 2007). In a standard beads task, one jar contains 85% red beads and 15% green beads, and vice versa in the other jar. After seeing one or two red beads, controls and delusional or delusion-prone participants provide the same (conservative) probability estimates (grey bars) for either jar. However, controls have a higher decision threshold (dashed line) than delusional or delusion-prone participants (dotted line). After one red bead, the probability estimates for either jar do not reach either group's decision threshold. After two red beads, however, the probability that they are from the red jar is high enough for delusional or delusion-prone participants to decide, but not yet for controls. This leads to the JTC bias, while probability reasoning is intact.

Following this account, Moritz et al. (2007) suggested that in a more ambiguous situation, the JTC bias should be abolished as no hypothesis will reach the (lower) decision threshold. In their study, participants completed three tasks. The first two tasks were standard graded-estimates and draws-to-decision versions, respectively. In the first task, participants were also asked to indicate whether their provided probability judgment would be sufficient evidence for a decision. The third task involved probability judgments after each bead, but this time for each of four jars, each with different ratios of beads. As there were multiple alternatives, with similar ratios, no hypothesis should stand out, and therefore no JTC should occur. In the first two tasks, the decision threshold was lower for schizophrenic patients than for controls; in the third task, however, no group differences occurred. This supports the liberal acceptance account (Moritz et al., 2007). However, L. O. White and Mansell (2009) did not find support for this account. In their study, delusion-prone participants and control participants performed several draws-to-decision versions of the beads task, including one with multiple jars. Delusion-prone participants decided on the basis of fewer beads compared with the controls in all conditions. L. O. White and Mansell (2009) argue that the discrepancy between theirs and Moritz et al.'s (2007) results might be due to Moritz et al. using a graded-estimates version, where a decision would not terminate trials. This might be less sensitive to a JTC reasoning style (Fine et al., 2007; Garety & Freeman, 1999), perhaps because participants decide faster when their decisions truly affect the trials.

The JTC studies in this thesis included draws-to-decision (in Chapter 2) and probability-estimates (in Chapter 3) versions of the task. These different measures were used to investigate whether potential differences in probability estimates were associated with potential differences in decision-making between low-delusion-prone and high-delusion-prone participants.

#### *1.3.1.2.4 A Lack of Motivation*

Finally, the JTC bias could be affected by low levels of motivation. The standard beads task presents no incentive to arrive at the correct decision regarding which jar is the source of the sequence of beads, other than to impress the experimenter. Considering the reduced working-memory capacity of delusional or delusion-prone participants, they might be relatively unmotivated to engage with a pointless task and might be in a "rush" to finish the study (L. O. White & Mansell, 2009). This lack of motivation, rather than an interesting cognitive bias, could then lead to the same JTC behaviour on the beads task, especially considering the notion that participants might only exert the required effort to reach rational decisions on reasoning tasks for which they are highly motivated (Kühberger, 2000). Perhaps, this lower motivation to arrive at a normatively correct conclusion about the jars could lead to the lowered decision-threshold suggested under the liberal acceptance account described above.

This confound could be addressed by introducing incentives in the beads task, which has been done to a very limited extent. Woodward et al. (2009) incorporated two conditions in which correct responses were rewarded with $0.25 or with $5. The analyses showed no significant difference between conditions, so that rewarding correct responses did not seem to affect patients' JTC bias. Woodward et al. (2009) suggested that this lack of an effect could be due to patients' ceiling performance, where patients already performed the best they could and monetary incentives could not improve performance above this threshold.

Lincoln, Ziegler, Mehl, et al. (2010) also used monetary incentives and did not find a difference in performance in this incentivised condition versus a non-incentivised condition. Lincoln, Ziegler, Mehl, et al. (2010) gave participants an initial ten tokens, each worth €0.25 at the end of the experiment. For a correct decision, they could earn one token; an incorrect decision would cost five tokens. Furthermore, participants were instructed that at some point no more beads would be drawn and the absence of a decision would leave the number of tokens unchanged. In order to equal losses and gains, five out of six decisions had to be correct and, therefore, a decision should only be made after 83.3% (5/6=.833) certainty was reached. Participants in all three groups decided faster than the Bayesian conditional probabilities would dictate. Lincoln, Ziegler, Mehl, et al. (2010) suggest that, despite not yet reaching a certainty level of

83.3%, participants may have been afraid the trial would be terminated, leading them to decide early so as not to miss an opportunity to win a token.

These studies do not seem to suggest that monetary incentives abolish the JTC bias, although neither tested specifically for the effect of such incentives. The two JTC studies in this thesis investigate the role of incentives further, as described in more detail in the introductions of Chapters 2 and 3. Furthermore, in order to investigate whether decision-making is really “premature” versus “conservative” (a question mentioned earlier, see 1.3.1.2.3 on page 37), the decision should be compared against an optimal decision point, much like conservative probability reasoning has been compared against the Bayesian posterior probability (Phillips & Edwards, 1966). In order for such a point to exist, one must consider both costs and benefits of deciding early, as elaborated upon in Chapter 2. Without such an objective point, it is not valid to label decisions as “premature” or “conservative”. Compare, for example, a situation where one compares the incomes of bankers and of movie stars. One could observe that bankers earn less than movie stars, but it would not be warranted to conclude from this that bankers are poor. Likewise, by just comparing people who use fewer data in their decision-making to people who use more data, one cannot label decisions made by the former as “premature” or those made by the latter as “conservative”. The study in Chapter 2 addressed this absolute, rather than relative, JTC bias. Differences between relative and absolute operationalisations have also been found for the optimism bias (discussed later at 1.5.1 on page 51).

The JTC bias could be a precursor to delusions. Delusions are a psychologically and, arguably, *biologically* maladaptive phenomenon (McKay & Dennett, 2009), with the potential of harm to the self or others (e.g., in the Cotard delusion, where one believes that one is dead, self-harm might be evoked to "prove" that one would not bleed) or of social isolation (e.g., when one believes others are conspiring against them, one may choose not to leave the house). Therefore, the JTC bias, which is associated with delusion formation, might represent an evolutionarily maladaptive deviation from rationality present in a minority of the population.[5] From such maladaptive deviations from rationality, the thesis continues to investigate other types of biases, some of which might actually be biologically adaptive, as discussed below.

## 1.4 Evolutionary Biases and Error-Management Theory

Perfectly rational decisions require maintenance of accurate representations of all probabilities involved, as well as representations of the costs of erroneous decisions. The biological limitations of human brain capacity complicate such optimal decisions and, instead, lead to deviations from rational belief (D. D. P. Johnson, Blumstein, Fowler, & Haselton, 2013). Not all of these deviations need to be evolutionarily maladaptive, as is presumably the case with the JTC bias (McKay & Dennett, 2009). Sometimes, deviations from rational beliefs might be evolutionarily adaptive. In such cases, normative rationality is determined by utility maximisation given the goals and beliefs of the individual, while "evolutionary rationality", with "rationality" at the locus of the genes themselves, is determined by maximisation of reproductive fitness (Stanovich & West, 2000). The idea of evolutionary influence on reasoning behaviour and belief revision is supported by the notion that there is variation in participants' responses on the reasoning problems used to study rationality, such as the

[5] Zolotova and Brüne (2006) found support for the notion that persecutory delusions are pathological exaggerations of threat recognition systems, which were adaptive in the ancestral environment. However, being pathological exaggerations, delusions should still be considered evolutionarily maladaptive.

Linda problem described earlier (under 1.1 on page 21). Some people do provide the normatively rational response. Variation in responses would suit the purpose of natural selection, which can select for the behaviours most adaptive in ever-changing environments (Greene & Levy, 2000). DeKay, Haselton, and Kirkpatrick (2000) argue that it is no surprise that human reasoning, formed through millennia of natural selection favouring processes which increased fitness, can fall short of the normative reasoning required in the problems used in rationality research, which are a few decades old, at best. Hence, sometimes, what is thought to be irrational compared to the normative response could be adaptive in an evolutionary sense.

To determine whether responses are biologically adaptive, one should consider the costs and benefits of various decisions. For example, foraging for food would carry the costs of potential time and energy wasted looking for food in new places where there might be none, but the potential benefits include finding more food than available in the current location, which could then sustain a larger family and thus increase evolutionary fitness. Deciding not to forage could lead to starvation once the food source in the current location is depleted. However, given the uncertainty regarding whether there are other food sources, errors could be made. The two errors that could be made in this particular case are not to forage when better food sources are available, and to forage when no better food sources are available (Greene & Levy, 2000).

As mentioned above, uncertainty can lead to errors in decision-making. The costs of different types of errors are often recurrently asymmetrical (D. D. P. Johnson et al., 2013). Although it might increase overall error rates, a bias against committing the more costly error(s) would be evolutionarily adaptive

(Haselton & Nettle, 2006; McKay & Efferson, 2010). This theory is known as error management theory (EMT; Haselton & Buss, 2000).

EMT has been applied to many domains to explain various biases (Haselton & Nettle, 2006; D. D. P. Johnson et al., 2013). For example, in the perception domain, it would be evolutionarily less costly to anticipate the arrival of the source of a tone (e.g., a predator) too early and have ample time to prepare than to prepare too late (Neuhoff, 2001). Hence, the mind might be biased towards interpreting approaching sounds to be nearer than they are. Indeed, Neuhoff (2001) found that participants judge an approaching tone to be closer than it really is. Another example is the illusion of control, where people have a superstitious belief that their behaviour influences outcomes which are not truly contingent on their behaviour (Alloy & Abramson, 1982; A. J. L. Harris & Osman, 2012; Rudski, 2001). Believing that one's behaviour can control outcomes can be beneficial, as an absence of this illusion of control, or even learned helplessness, is found in depressed participants (Alloy & Abramson, 1982). Generally, believing that one's behaviour controls outcomes when it does not, especially when the behaviour is not effortful, may be less costly than believing one's behaviour cannot control outcomes while it can (A. J. L. Harris & Osman, 2012; Rudski, 2001). The costs of expending some energy to pointlessly press a lever are arguably less than the psychological costs of assuming chaos, which may lead to learned helplessness, which, in turn, is linked to depression. Therefore, the illusion of control might be an adaptive bias, and possibly explained by EMT.

### 1.4.1 Sexual Over-Perception

The most prominent example EMT has been applied to is the sexual over-perception bias, where men perceive more sexual interest from a woman than the woman herself reports or other women perceive (Haselton & Buss, 2000). In the mating domain, sexual interest is interpreted from behavioural signals, which are contaminated by noise (e.g., if playing "hard-to-get", sexually-interested women might not display sexually-interested behaviours; Jonason & Li, 2013). This leads to the four alternatives in signal detection theory (see Figure 1.2): correct rejections, where no sexual interest is perceived from the behaviour of non-sexually-interested women; hits, where sexual interest is perceived from the behaviour of sexually-interested women; false alarms, where sexual interest is perceived from the behaviour of non-sexually-interested women; and misses, where no sexual interest is perceived from the behaviour of sexually-interested women.

Given that the perception of a woman's sexual interest influences men's courtship-initiating behaviour (Choi & Hur, 2013), the two errors a man can make in the perception of a potential mate's sexual interest carry different costs. A false alarm would result in a waste of time and effort when making futile advances, while a miss would result in a missed opportunity to reproduce. According to EMT theorists, the costs of these errors are highly asymmetrical, and EMT accordingly predicts that errors should be biased towards false alarms rather than misses, to minimise overall fitness costs. Hence, men are predicted to over-perceive sexual interest from women, perhaps by adopting a more liberal threshold for assuming sexual interest (Farris, Treat, Viken, & McFall, 2008a; Shotland & Craig, 1988).

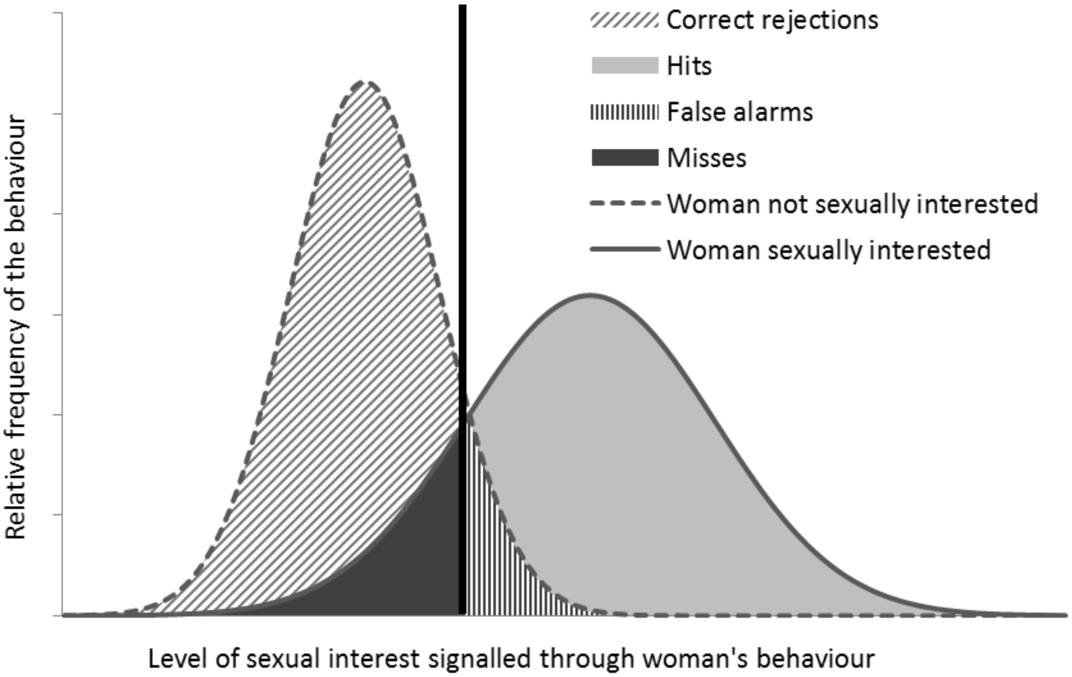


**Figure 1.2** A signal detection model of the four alternatives in interpreting a woman's sexual interest (inspired by Fig. 2 in Farris et al., 2008b). A woman's different behaviours towards a man can signal different levels of sexual interest (x-axis; ranging from, for example, making eye contact to touching a man's genitals). Some behaviour is shown both by women who are not sexually interested (dashed line) and who are sexually interested (solid line), and the signalled level of sexual interest is ambiguous. Men have to interpret this ambiguous information and decide when they assume sexual interest (indicated by the black vertical line: assume a woman is not sexually interested with behaviours left to the line, assume a woman is sexually interested with behaviours right to the line). Correct interpretations can result in correct rejections (diagonally striped area) or in hits (light grey area). Incorrect interpretations lead to false alarms (vertically striped area) or to misses (dark grey area).

Evidence in support of this prediction has accumulated, especially for behaviours that could be expressed in platonic as well as sexually-intended interactions (Lindgren, Parkhill, George, & Hendershot, 2008). Abbey (1982) first found this bias in an experiment where a male and a female actor had a conversation, while unbeknown to them, a male and a female observer observed the conversation. All four participants then judged the male and female actors on several traits, including promiscuity and flirtatiousness. Actors indicated whether they were sexually attracted to and if they would want to date the other actor (i.e., their interaction partner). Observers judged whether they thought the two actors were sexually attracted to and would want to date each other. Overall, it was found that male actors and observers rated the female actor as more seductive and promiscuous than female actors rated themselves or than female observers rated the female actors. Male observers also considered the female actor to be more sexually attracted to and willing to date the male actor than did female observers (Abbey, 1982). Males also rated other men more highly on the sexual-intention-related items, suggesting that, perhaps, men generally perceive more sexual intentions in people's behaviour compared to women (Abbey, 1982).

Further research has also shown this male over-perception of sexual interest when males read a vignette describing interactions between male and female targets (Abbey & Harnish, 1995), instead of observing a dyad interacting (Abbey, 1982). Furthermore, the sexual over-perception bias can be exacerbated by situational cues, such as alcohol use and provocative clothing (Farris et al., 2008b). Perilloux, Easton, and Buss (2012) found support for this sexual over-perception bias and also showed that the level of sexual over-perception is moderated by attractiveness of both the perceiver and the actor. Men who considered themselves as more attractive also over-perceived more sexual interest from their female interaction partners. Moreover, the more attractive a female was rated by men, the more her sexual interest was over-perceived (Perilloux et al., 2012). Recently, Fletcher, Kerr, Li, and Valentine (2014) found that men, besides perceiving more sexual interest, also perceive more romantic interest, assuming females had a stronger desire to get to know them better or to go on a further date after a speed-date, than was actually the case.

After a review of the literature, Lindgren et al. (2008) argue that there is reliable support for the sexual over-perception bias shown by men. They note some limitations, however, such as that perceptions are measured only once and no information is available about possible changes in perception. Furthermore, although EMT offers a potential explanation of the sexual over-perception bias, different socialisation and cultural expectations can also lead men to have more sexual expectancies than women, leading them to perceive more sexual interest (Lindgren et al., 2008). The sexual over-perception study in this thesis addresses some limitations of both EMT's theoretical underpinnings and of the evidence for the sexual over-perception bias, described in further detail in Chapter 4.

To summarise, the sexual over-perception bias involves men having biased estimates of their prospects in the mating domain. Error management theory suggests that this bias has evolved because it leads to potential benefits in terms of biological fitness, at the expense of small costs, such as the psychological cost of rejection when pursuing females who are not interested. The focus of this thesis now shifts to a bias in estimates of future prospects present in both genders.

## 1.5 Self-Deception

Much like sexual over-perception has been theorised to be a result of evolutionary influences, von Hippel and Trivers (2011) argue that evolution may have promoted a capacity for self-deception. Self-deception is the motivated acquisition and retention of a belief in the face of countervailing evidence (Deweese-Boyd, 2012). There are various instances of self-deception (von Hippel & Trivers, 2011), including defensive strategies where hypotheses are maintained despite disconfirming evidence (Gur & Sackeim, 1979), and forms of self-enhancement. Self-deceptive self-enhancement describes the phenomenon where the self is considered better, in some way, than appears justified (Alicke & Sedikides, 2009). If this better version of oneself is then signalled to others, it can increase one's appeal as a social or reproductive partner, in turn enhancing reproductive fitness. This would mean self-enhancement is evolutionarily adaptive (von Hippel & Trivers, 2011). Recently, Lamba and Nityananda (2014) have provided empirical support for the theory that self-deception is associated with others' deception, as others are equally overconfident about someone's abilities as the person is themselves.

One example of self-enhancement is the "better-than-average" effect, where most people consider themselves above average on various desirable traits (Alicke, 1985). This is not just reported to impress others, without truly believing oneself to be more creative, intelligent, or mature. Williams and Gilovich (2008) found that participants were just as willing to play a bet where winning required them to score higher than a random other participant on a positive trait as they were to play a bet of a random draw using their percentile ratings. For example, if they indicated their score would be in the 60th percentile, their chances of winning in the random draw would be 60 out of 100. If they had not truly believed their reported percentiles, they should have favoured the percentile bet over the bet where they had to score higher than a random other person.

Another example of self-enhancement is that people appear to think they are more attractive than they really are. Epley and Whitchurch (2008) presented participants with a set of pictures: one was an undistorted picture of the participant; others had been morphed into more attractive and less attractive representations of the participant. Participants had to indicate which picture was the true representation of their own face, and tended to select slightly more attractive faces than their own.

Unrealistic optimism, the phenomenon where good future outcomes are expected to be more likely, and bad future outcomes less likely, than indicated by an objective standard (Segerstrom, 2007; Shepperd, Klein, Waters, & Weinstein, 2013), could be considered another form of self-deception. People might self-deceive in thinking that they are more able to avoid misfortune and attract fortune than others. This bias is elaborated on below.

Despite abundant examples of self-deception, the process underlying self-deception is a long-debated issue (Mijovic-Prelec & Prelec, 2010). Several competing interpretations have been put forward.

One interpretation draws an analogy between self-deception and interpersonal deception. Here, "person A deceives person B (where B may or may not be the same person as A) into believing that *p* only if A knows, or at least believes truly, that ¬*p* and causes B to believe that *p*" (Mele, 1997, p. 92). Hence, one belief (¬*p*) must be held, but holding the other belief (*p*) must be desired. If self-deception succeeds, one arrives at the belief *p*; if it fails, one maintains belief ¬*p* (D. L. Smith, 2011). A major paradox with this interpersonal analogy conception of self-deception is that one would have to hold beliefs *p* and ¬*p* at the same time (the "static" paradox). Another paradox (the "dynamic" paradox) is that one would have to deceive oneself into believing something that one knows to be false, and the knowledge of the deception process should undermine it and prevent adoption of the desired belief (Mele, 1997; Mijovic-Prelec & Prelec, 2010). One response to these paradoxes is that one can hold two opposite beliefs (i.e., *p* and ¬*p*) at the same time, as long as one is held consciously and the other subconsciously (Bandura, 2011; Gur & Sackeim, 1979), perhaps in different, autonomous modules of the mind (Kurzban, 2011).

An alternative, “deflationary” account of self-deception avoids the paradoxes within the interpersonal-analogy account, without invoking several sub-selves at different levels of consciousness. Here, it is argued that there is no self-deception when people are not aware of the falsity of what they are saying (*p*), because the truth ($\neg p$) is not known (Fridland, 2011). The truth can remain unknown through biased information processing, and even through unintentional, unmotivated, biasing processes, such as salience of certain types of information (Mele, 1997).

Consider these two different conceptions within the optimism bias. On the classic conception of self-deception, a heavy smoker who believes her future health prospects are good may also represent a more accurate, and less rosy, state of affairs. In contrast, proponents of the deflationary view might argue that there is no need to suppose that she carries two conflicting representations. She may be processing evidence about the health implications of smoking in a biased fashion (Sharot, 2011) to arrive at one false representation.

The study presented in Chapter 5 aimed to tease apart these two accounts of the processes underlying self-deception, focusing on the optimism bias.

### 1.5.1 Unrealistic Optimism

Unrealistic optimism is the phenomenon where good future outcomes are expected to be more likely, and bad future outcomes less likely, than indicated by an objective standard (Segerstrom, 2007; Shepperd et al., 2013). In moderate amounts, optimism can be beneficial as it has been associated with various positive outcomes, such as showing helpful financial behaviours for the future (e.g., saving; Puri & Robinson, 2007), improving adjustment during life transitions due to better perceived social support (Brissette, Scheier, & Carver, 2002), reducing stress (Solberg Nes & Segerstrom, 2006), and physical health (S. E. Taylor, Kemeny, Reed, Bower, & Gruenewald, 2000). Furthermore, people with major depressive disorder do not show these optimistic tendencies (Korn, Sharot, Walter, Heekeren, & Dolan, 2014), again suggesting that the optimism bias (in moderate amounts) is beneficial. However, this bias may also have significant drawbacks. For example, some authors have argued that unrealistic optimism has contributed to global economic crises (e.g., Sevincer, Wagner, Kalvelage, & Oettingen, 2014; Ubel, 2009).

Unrealistic optimism is shown in many domains (Sharot, 2011a). Unrealistic comparative optimism is found when people report their chances of good outcomes compared to others' chances (Shepperd et al., 2013). For example, people think their own marriages are less likely to result in divorce than are others' marriages (Baker & Emery, 1993). Unrealistic absolute optimism is found when people report their chances compared to reality (Shepperd et al., 2013). For example, people expect to finish tasks faster than they do (Buehler, Griffin, & MacDonald, 1997; Buehler, Griffin, & Ross, 1995), and people underestimate their own chances of having alcohol-related problems later in life (Dillard, Midboe, & Klein, 2009). Optimism tends to be greatest for events which have not (yet) been personally experienced, which are rare, which are controllable, and which show symptoms early on (Weinstein, 1989).

One potential argument is that unrealistic optimism is a reporting bias, where people do not truly believe what they report. However, much like participants still showed the better-than-average effect in Williams and Gilovich's (2008) incentivised paradigm (discussed on page 49), Simmons and Massey (2012) showed that, even when promised a large reward for accuracy, people reported optimistic beliefs that their favourite football team would win. This suggests that people truly believe their stated expectancies. These results do not clarify whether the false belief is held alongside another, sub-consciously held, veridical belief, or whether it is the only belief, arrived at through biased information processing. The former case would support the interpersonal-analogy account of self-deception. The latter would support the deflationary account of self-deception. The study in Chapter 5 aimed to tease these two options apart, while incentivising the reporting of truly-held beliefs.

The optimism studied in this thesis centres on the optimism bias for future (mis)fortune. Weinstein (1980) showed that people expected their own chances of experiencing positive events to be above average, while they estimated their chances of experiencing negative events as below average. Although some studies focus on overestimating the chances of experiencing positive events (e.g., Hoorens, Smits, & Shepperd, 2008), most focus on underestimating the chances of experiencing negative events. Optimism for such events tends to be stronger (Shepperd et al., 2013) and tends to carry more relevant consequences as people might not take the necessary precautions to protect against harmful behaviours (Weinstein & Klein, 1995). For these reasons, this thesis also focused on negative events.

A. J. L. Harris and Hahn (2011) noted that statistical artefacts may have confounded unrealistic-optimism studies, especially comparative-optimism studies using rare events (Shepperd et al., 2013). These statistical artefacts include scale attenuation, minority undersampling, and base-rate regression. Scale attenuation could explain unrealistic optimism if a restricted response set, such as a 5-point Likert scale running from “much below average” to “much above average", is used. For a rare event, most people are at a below-average risk. Similarly, for rare events, the participant sample might genuinely not contain anyone from the minority sample who would experience the negative event. The participants in the study's sample may thus be correct in their low expectations. Finally, when estimating an unknown, average person's chances of experiencing a rare negative event, people's estimates are often not extreme enough (i.e., they are not low enough/regressive). Yet, when estimating for oneself, enough information is available about behaviours that would decrease risks, so that low, non-regressive estimates are provided. This would then lead to unrealistic comparative optimism, where one's own risk is considered to be lower than the average person's risk (A. J. L. Harris & Hahn, 2011; Shepperd et al., 2013). Given these potential confounds, the unrealistic optimism studied in this thesis is limited to a paradigm least affected by such artefacts (Shah, 2012): optimistic belief updating (Sharot, Korn, & Dolan, 2011).

In the optimistic belief-updating paradigm (Sharot, 2011a; Sharot, Guitart-Masip, Korn, Chowdhury, & Dolan, 2012; Sharot et al., 2011) participants first indicate their chances of experiencing various adverse events. Participants are then presented with the base rate of the event happening to their demographic group; this constitutes the average person's chance of experiencing the negative event. The base rate can provide desirable information (i.e., the event is less likely than initially thought) or undesirable information (i.e., the event is more likely than initially thought). Then, after having been presented with the base rates, participants are asked to provide their own chances again. Optimistic updating is investigated by comparing the amount of belief updating based on desirable information versus the amount of belief updating based on undesirable information. This paradigm includes adverse events with base rates ranging from 10% to 70% and participants indicate their chances on a 0% to 100% scale, avoiding rare events and minimising influences of scale attenuation. Furthermore, although base rates are presented, no comparison is made between estimates for the average person and for oneself, but rather for how desirable versus undesirable information is attended to, minimising the artefacts of minority undersampling and base-rate regression.

Findings from this paradigm show that participants update their initial beliefs more when the base rate is lower than their initial guess (i.e., the base rate provides desirable information) than when the base rate is higher than their initial guess (i.e., it provides undesirable information), demonstrating an optimistic, selective updating of beliefs (Sharot et al., 2011). As such, the optimism bias is not just a bias of considering desirable events as more likely, and undesirable events as less likely, than they are (Shepperd et al., 2013). It also involves biased updating of beliefs about the probability of negative events, with an optimistic bias towards desirable information (i.e., information that indicates that undesirable events are less likely than initially thought).

Such selective updating strongly suggests the bias is due to biased information processing, which could be carried out intentionally or unintentionally (Mele, 1997). Intentionally biasing information processing would suggest the more veridical belief is actively avoided and thus support the interpersonal-analogy account of self-deception. Unintentional biased information processing would support the deflationary account of self-deception.

To investigate whether biased information processing in the optimism bias is intentional or not, one could compare belief updating for undesirable items to belief updating for neutral, impersonal items. Biased processing, either unintentionally or intentionally, is more likely for the former than for the latter; if the latter would be biased, it is unlikely that this would be due to intentional processes. For undesirable items, the base rates form an objective standard, where deviations from this standard on an individual level are still possible, given individuating information (e.g., a family history of a certain disease). Yet, for neutral, impersonal items, the objective standard would be the correct answer and deviations should not be found after having been presented with the answer, especially when participants are incentivised to provide accurate answers. Hence, based on wanting to maximise expected value by giving the correct answer for their own chances, participants would be justified in deviating from the base rate in the second estimate for undesirable items, due to individuating information they might have. They would not be justified in deviating from the correct answer for neutral items. As such, a difference in updating for the two types of items could simply be due to a maximisation of expected value, rather than due to self-deceptive optimism.

In devising a possible solution to this problem of justifiable deviations in the second estimate for personal, negative items, but not for impersonal, neutral items within the belief-updating paradigm, we take inspiration from the crowd-within literature (Herzog & Hertwig, 2009; Vul & Pashler, 2008). In this paradigm participants provide an estimated answer to a neutral question (e.g., "what percentage of the world's airports are in the United States?"). They are then told to assume their initial answer was wrong, and asked to provide a second, alternative estimate (Vul & Pashler, 2008). Without directional feedback (cf. base rates in the optimistic belief-updating paradigm), the second estimates can be higher or lower than the first estimate. One's intentions and desires might influence whether one gives a higher or lower second estimate compared to the first estimate in the crowd-within paradigm. If there is no desire about the answer being either high or low, the second estimates are equally likely to be lower as they are to be higher than the first estimate. If there is a desire to hold a belief in a certain direction, second estimates might be systematically different from first estimates. We hypothesised there would be no desired directions for neutral questions, but there would be desired directions for undesirable questions. This was investigated in the study described in Chapter 5.

Just as people appear to have irrationally optimistic beliefs about the future, some people have irrationally pessimistic, if not paranoid, beliefs about the intentions of others. Haselton and Nettle (2006) argued that both such biases can co-exist and be explained by error-management theory (EMT). This theory, as described earlier, holds that humans may have evolved cognitive systems biased towards making less costly errors when the costs of different types of errors are recurrently asymmetric. The optimism bias discussed above could be due to people overestimating the effectiveness of their own efforts to avoid misfortune. Thinking one's efforts are effective, when they are not, may be overall less costly than assuming one's efforts are not effective, when they are, which may lead one to not undertake action to prevent misfortune. Paranoid beliefs and distrust might arise because, over time, it would be beneficial to minimise the chance of failing to detect others' negative intentions, at the cost of increasing the chance of inferring negative intentions when these are not really there (Haselton & Nettle, 2006). This thesis moves from optimistic beliefs about the future to more pessimistic beliefs about others' intentions. In particular, the last part of the thesis will focus on apparent irrationality in trust behaviour.

## 1.6 Trusting in Trust Games

Trust is generally defined as "a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behaviour of another" (Rousseau, Sitkin, Burt, & Camerer, 1998, p. 395). Trust could be measured through survey questions, such as "Generally speaking, would you say that most people can be trusted or that you can't be too careful in dealing with people?" (Glaeser, Laibson, Scheinkman, & Soutter, 2000, p. 812). However, there is evidence that attitudinal self-report measures of trust, like answers to the question above, do not correlate with actual trust behaviours and correlate only weakly with one's own trustworthy behaviour (Glaeser et al., 2000).

Given the relative lack of validity and reliability of survey measures of trust, another measure of trust focuses on behaviour in an economic game known as the trust game. The standard trust game involves two anonymous players interacting in two stages (Berg, Dickhaut, & McCabe, 1995). In the first stage, one participant, the sender, is endowed with a certain amount of money (e.g., £10) and can decide to send any portion ($x$), including nothing at all, of the endowment to the other player, the trustee. Whatever is not sent (£10-$x$) is kept by the sender. The portion sent ($x$) is then tripled ($3x$) and passed on to the trustee. In the second stage, the trustee can decide to return any portion ($y$), including nothing at all, of the received money ($3x$) to the sender. Panel A of Figure 1.3 shows the structure of this standard trust game. In a binary-choice version of the trust game (Camerer & Weigelt, 1988), the sender can decide to send a set amount of money (e.g., £10) or not in the first stage. If it is not sent, the game ends and both players receive a small reward (or in some cases the trustee does not receive anything). If it is sent, in the second stage the trustee receives a multiple (multiplied by a factor >1) amount of the money sent (e.g., £30). The trustee can then decide to reward trust and send a set, fair amount of money back (e.g., £15) or to betray trust and keep (a large sum of) the received money (e.g., £22; Hong & Bohnet, 2007). Panel B of Figure 1.3 shows the structure of this binary trust game. The proportion of the endowment sent, or the choice to send a set amount, is considered a measure of trust. The returned proportion of the money received, or the choice to reward trust by returning a set, fair amount, is a measure of trustworthiness (N. D. Johnson & Mislin, 2011).

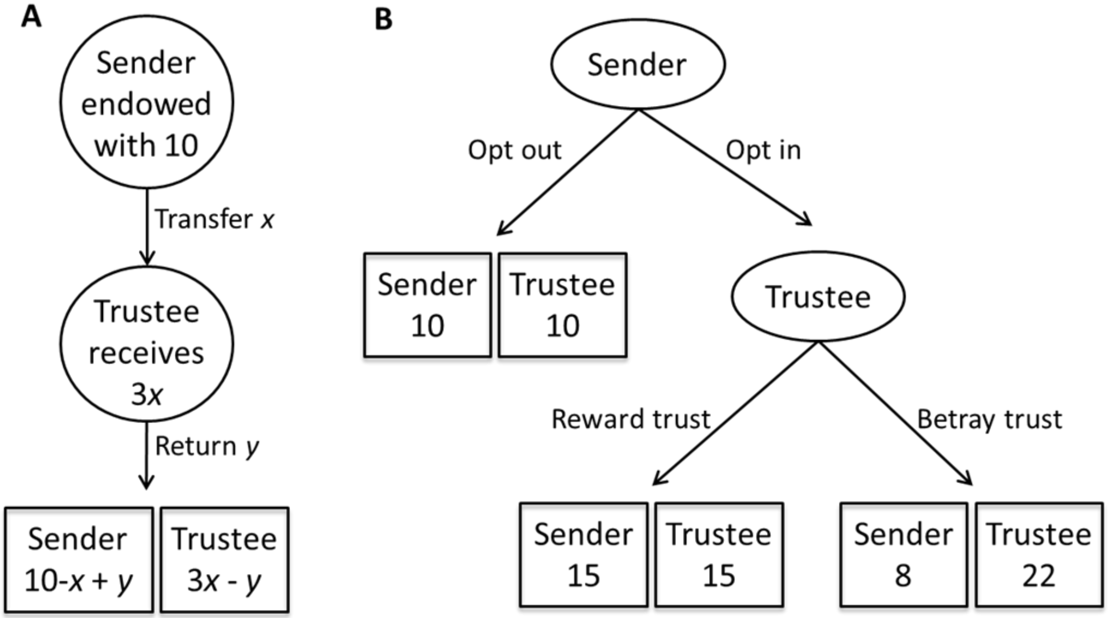


**Figure 1.3** The structure of the standard (panel A) and of the binary (panel B) trust game (panel B based on Figure 1 in Bohnet, Greig, Herrmann, & Zeckhauser, 2008). Stages at which a player makes a decision are shown in ovals; actions are represented by arrows; outcomes are depicted in rectangles. Note that in the standard trust game, the values *x* and *y* could be zero. Note also that the numbers provided are examples and studies have varied these amounts, including versions with no payment for the trustee if the sender opts out of the binary trust game (see the meta-analysis in N. D. Johnson & Mislin, 2011).

From a neo-classical economics point of view, with assumptions of rationality based on maximisation of expected value, participants in the sender role in an anonymous, one-shot trust game, should not trust, as participants in the trustee role have no incentive to reciprocate and thus should not return anything (Manapat, Nowak, & Rand, 2013; Weber, Malhotra, & Murnighan, 2004). Yet, Berg et al. (1995) found that, on average, senders sent 51.6% of their endowment, which was reciprocated by more than half of the trustees (sending back what had been invested or more). This is a common finding, as a meta-analysis of 162 trust game scenarios showed that, on average, senders sent half of their endowment to the trustee, and the trustees return, on average, a bit more than the sum invested (N. D. Johnson & Mislin, 2011).

As noted before, a normative standard like maximisation of expected value might be ignoring the influence of the environment (e.g., Sturm, 2012). In this case, the environment might have led to the evolution of trustworthy behaviour. Despite potential financial costs of being exploited, trust and trustworthiness tend to be beneficial for the economy and society at large (Hong & Bohnet, 2007; N. D. Johnson & Mislin, 2011). For example, correlational evidence shows that countries with higher levels of trust also have higher and more equal incomes (Knack & Keefer, 1997). Indeed, theories that trust is adaptive, leading to it being naturally selected for throughout evolution, have been put forth (Manapat et al., 2013; McNamara, Stephens, Dall, & Houston, 2009).

Using the binary trust game as a model of naturally occurring cooperative interactions, McNamara et al. (2009) found that social awareness of variation in trusting and trustworthy behaviours can maintain such variation. At the extreme ends of trusting behaviour's variation, senders would never or always trust, without gathering more information about the trustee's trustworthiness. However, due to random mutations, some senders might be willing to gather and use more information in their decision whether to trust or not. Some trustees might be aware of the possibility that a future sender might obtain information about the trustee's past behaviours. This should then lead trustees to show trustworthy behaviour, at least on some occasions, so as to build a relatively trustworthy reputation. Senders, in turn, being aware of such variation in trustees' trustworthiness, should be more willing to investigate the trustee's past behaviours. With individual differences in the awareness of such variation in trusting and trustworthy behaviours, both are then maintained (McNamara et al., 2009). For such behaviour to robustly evolve, senders also need to use the information obtained about trustees' trustworthiness in their decisions about which trustees to trust (Manapat et al., 2013). This then creates a free market where trustees need to meet the minimal demanded level of trustworthiness in order to be selected over competitor trustees (whose levels of trustworthiness may or may not be known), while making profit by occasionally betraying trust. Naturally occurring interactions in real life, throughout evolution, have often involved access to information regarding previous trustworthiness and involved choices of interaction partners. This might explain why trusting and trustworthy behaviours are still found in the (generally) anonymous, one-shot trust game (Manapat et al., 2013).

Other factors than reputation management, such as geographical location, student samples, and whether people play against a real, other person or a computer, have been shown to influence trust and trustworthiness (N. D. Johnson & Mislin, 2011). This last influence, whether people play against another person or a computer, is highly relevant to the last study reported in this thesis. Different behaviours for playing against human players or a computer have been highlighted in the ultimatum game (Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003). Here, two players have to divide an amount of money between them. One player is the proposer and can propose an offer of how to divide the total sum. The other player, the responder, can either accept or reject the offer. If accepted, both players are paid the amount stated in the offer; if rejected, neither player gets paid anything. Participants in the role of the responder have been found to reject unfair offers (i.e., unequal splits with more money for the proposer) made by another human player at a significantly higher rate than that at which they reject computerised unfair offers (Sanfey et al., 2003). Rejecting any non-zero offer is irrational from the view of a standard economic model, as any non-zero offer would increase monetary payoff. However, unfair offers in the ultimatum game are often rejected, arguably because people want to punish the proposer for their unfairness and maintain a social reputation so as not to be exploited in the future (Nowak, Page, & Sigmund, 2000). Besides reputational concerns, Fehr and Gächter (2002) found that costly punishment can be an altruistic act, as it increases the probability that others will cooperate in future encounters, even when these encounters are with different people. These reasons may explain why Sanfey et al. (2003) found more rejections of unfair offers made by humans, who could be punished and might change their behaviour in the future, than of unfair offers made by a computer, which would not change its (random) choice based on having been punished.

Overall, the theories explaining the differences in response to human and computerised unfair offers in the ultimatum game tend to extend to all cooperative interactions. Hence, differences between interacting with a human or with a computer should occur in the trust game as well. Indeed, a phenomenon known as betrayal aversion, described in more detail below, has been found when playing against humans, but not computers, in the trust game.

The fifth study of this thesis, reported in Chapter 6, focused on this phenomenon.

### 1.6.1 Betrayal Aversion

Betrayal aversion is the phenomenon where people avoid risk more when a person determines an uncertain outcome compared to when a random mechanism, such as a (computerised) lottery, determines the outcome (Aimone & Houser, 2012).

Bohnet and Zeckhauser (2004) first reported this phenomenon. They used three variants of a one-shot game, in a between-subjects design: the decision problem, the risky dictator game, and the trust game. The decision problem (panel A in Figure 1.4) is a measure of risk-taking behaviour. Participants, as senders, have two options: opt out and receive a certain 10 points or opt into a lottery with potential outcomes of 8 points or 15 points with unknown probabilities 1-*p* and *p*, respectively. The risky dictator game (panel B in Figure 1.4) is similar to the decision problem, with the addition of another player who is a passive recipient. The passive recipient would also get 10 points if the sender opts out, or 22 or 15 points if the sender opts in and receives 8 or 15 points, respectively. Finally, Bohnet and Zeckhauser (2004) used a binary trust game (panel B in Figure 1.3). In comparison to the risky dictator game, in the binary trust game the passive recipient becomes the trustee, an active player, and the computerised lottery determining the outcomes of the risky choice is replaced by the trustee's decision between the two outcomes (Bohnet & Zeckhauser, 2004; Hong & Bohnet, 2007).

In all three games, participants provided the minimal value they needed the probability of the good outcome (*p*) to be in order to opt in. Note that the outcomes for the sender are the same across these three variants. Yet, the minimal acceptable probability to opt in was higher in the trust game than for the decision problem or the risky dictator game, while the latter two did not differ. Bohnet and Zeckhauser (2004) concluded that the bad outcome was more costly when chosen by the trustee than when selected in a lottery. Betrayal costs, which are "costs that make it less attractive to rely on a [t]rustee than a random device offering the same probabilities" (Bohnet & Zeckhauser, 2004, p. 478), are additional to the cost of choosing a risky option with the chance of receiving a worse outcome than would have been received if one had opted out. The reduced willingness to opt into a trust game, compared to the lotteries in the other games, was considered evidence of betrayal aversion.

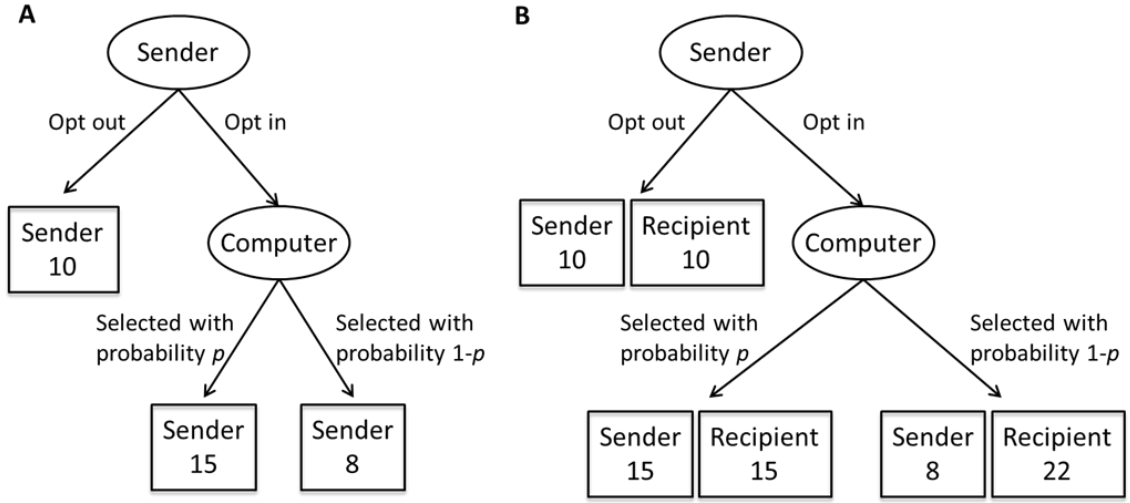


**Figure 1.4** The decision problem (panel A) and the risky dictator game (panel B), which are contrasted with the binary trust game (panel B of Figure 1.3) to investigate betrayal aversion (based on Figure 1 in Bohnet et al., 2008). Stages at which the player or a computer makes a decision are shown in ovals; actions are represented by arrows; outcomes are depicted in rectangles.

Using the same three games, Bohnet et al. (2008) replicated their result across a range of countries, consisting of the United States, Brazil, China, Oman, Switzerland, and Turkey. Hong and Bohnet (2007) and Aimone and Houser

(2011, 2013) also found betrayal aversion in their studies. However, Fetchenhauer and Dunning (2012) did not find evidence for the betrayal aversion phenomenon. There are a few methodological differences between these studies, discussed further in Chapter 6, which might explain why evidence of betrayal aversion has been found in some studies but not in others.

Besides seeking support for betrayal aversion in a paradigm free from potential methodological confounds, the study in Chapter 6 tested a theory to explain the phenomenon, as few previous studied have attempted to do this. Instead, previous studies on betrayal aversion have generally focused on its presence (Bohnet & Zeckhauser, 2004; Fetchenhauer & Dunning, 2012), on the extent to which it can be generalised to other cultures (Bohnet et al., 2008), on whether it is influenced by status (Hong & Bohnet, 2007), or on how its presence can increase levels of trustworthiness (Aimone & Houser, 2011, 2013).

Aimone and Houser (2012) form an exception to this and offer an explanation of betrayal aversion: in an attempt to regulate their emotions, participants might opt out of the trust game to avoid taking an action that could have unpleasant emotional consequences. Participants trusted less when they could be exposed to “personal betrayal” compared to conditions with potential exposure to “general betrayal” (Aimone & Houser, 2012). Personal betrayal was present when the focal participant’s assigned trustee selected the bad outcome after learning their assigned sender had opted in. In other conditions, participants might have been exposed to general betrayal, because they received the bad outcome from the computer, which, in turn, selected good and bad outcomes with probabilities equal to the proportion of trustees who selected the good and bad outcomes. Aimone and Houser (2012) claim to have accounted for inequality aversion confounding trust (i.e., opting out to avoid potential inequality of outcomes) by including a condition where the computer selected an outcome, but still showing the assigned trustee's payoff, and thus potential inequality of payoffs. However, a sender who receives the bad outcome from the computer still knows that some trustee chose the bad outcome. This trustee, whether it is the specific assigned trustee or not, would be paid more than the sender, and so inequality aversion might still affect the results in this study.

The study in Chapter 6 avoids confounds of signalling distrust and of inequality aversion. Furthermore, it further develops the notion of emotional costs of betrayal, along with a prediction that these costs depend on prior beliefs regarding others' trustworthiness. This theoretical prediction is elaborated on in Chapter 6.

## 1.7 Experimental Practices in Economics versus Psychology

All studies reported in this thesis involve some methods from experimental economics and, as such, straddle the theoretical and methodological divide between experimental psychology and behavioural economics (as in, e.g., Wischniewski & Brüne, 2011). In this section, some of the key methodological differences between these disciplines are reviewed. Furthermore, it is outlined how the different approaches were consolidated into the studies reported in the thesis.

Whereas economists have tended to focus on rationality, psychologists have emphasised cognitive limitations and have highlighted how choices are sensitive to context, leading to irrational behaviour (Camerer, Loewenstein, & Prelec, 2004). This thesis focuses on deviations from rationality from a psychological perspective, using certain methods from behavioural economics, in an attempt to achieve the best of both disciplines. All reported studies were conducted in the Economics Department's EconLab at Royal Holloway, University of London (RHUL). Here, participants are seated at separate cubicles, with interconnected computers using Zurich Toolbox for Readymade Economic Experiments (z-Tree) software (Fischbacher, 2007). Participants do not speak to one another, they cannot see the monitors of other participants, and all identities remain anonymous, with everyone known only by their computer's number. All experiments had detailed written instructions outlining participants' tasks and the number of trials, and where necessary their assigned roles (i.e., a script; Hertwig & Ortmann, 2001), as well as comprehension questions. Participants' earnings were kept confidential by using individual receipts.

The two differences between economics and psychology methodologies most relevant to this thesis are the use of deception and the use of incentives.

### 1.7.1 Deception

Both fields seem to agree on the distinction between "real deception" and "deception by omission". The former involves actively and intentionally misrepresenting aspects of the study to participants. The latter involves withholding information about certain aspects of the study, such as not informing participants of the exact hypotheses under investigation (Hertwig & Ortmann, 2008b). Whereas deception by omission is allowed in both fields (Ortmann & Hertwig, 2002), and might minimise the risk of demand effects (Bonetti, 1998), the two fields diverge with respect to their stances on the use of real deception (Hertwig & Ortmann, 2001).

Psychologists argue that deception is often needed to create the context for their topic of investigation (Ariely & Norton, 2007). For example, helping behaviour in emergencies can only be studied with experimental control if these emergencies are created, rather than naturally occurring (Hertwig & Ortmann, 2008a). Deception is often used to distract from the true purpose of the study, making the behaviour of interest more natural (Bonetti, 1998). Bröder (1998) provides an example of a topic of investigation that requires deception: incidental learning. First, participants are asked to perform an irrelevant task, such as rating the emotionality of a list of words. Then, they are unexpectedly asked to recall the words that were on the list. If they had been told there would be a memory test, the learning during the first stage would not have been incidental. Of course, this deception might be considered simply deception by omission, rather than active real deception (Ortmann & Hertwig, 1998). Nevertheless, other factors could influence findings if no active deception about the true purpose is used. For example, in studies without a clear, objectively correct answer (e.g., questionnaires), social desirability might influence participants' self-presentation, such as avoiding racial stereotyping behaviour, and here active deception could be beneficial in masking the true purpose of the study and counter this effect (Weiss, 2001).

Economists, however, consider the use of (active) deception in experiments a risk to the subject pool. Economists worry that if participants find out they have been deceived in one experiment, either through first-hand experience or by being (inadvertently) informed by peers, they will be suspicious in future experiments (Hertwig & Ortmann, 2001). Contamination of the subject pool can also occur through second-hand experience with deception, which could arise from coverage of psychological studies in classes or even in the media (Hertwig & Ortmann, 2008a). Moreover, the "spill-over hypothesis" (Barrera & Simpson, 2012) argues that the participant pool may contain participants who take part in a psychology study that uses deception, and subsequently become suspicious of all experiments, whether conducted in an economics or a psychology lab. Another potential problem with the use of deception is that experimenters may face a more and more difficult challenge in making their participants believe the cover story for future studies as participants become more and more suspicious (Baron, 2001; Weiss, 2001). Although not negating the notion that participants might become suspicious after finding out they have been deceived, participants have been reported to generally understand and accept the need for deception, still cooperate in future investigations, and even enjoy studies despite the use of deception (Kimmel, 1998). But, as Ortmann and Hertwig (1998) point out, this only provides reassurance regarding participants' attitudes towards (future) psychological experiments, and not regarding their actual behaviour in such experiments.

The evidence for the behavioural effects of deception is mixed. After having reviewed relevant literature, Bonetti (1998) concluded that the use of deception does not alter participants' behaviour and can enable the study of more natural behaviour. However, after a more extensive systematic literature review, Hertwig and Ortmann (2008a) concluded that first-hand experience with deception can alter behaviour, or at least increase the suspicion of being deceived again in future experiments.

Hertwig and Ortmann (2008a) argue that different dependent variables can be differentially influenced by previous experience with deception. On the basis of a literature review, they argue that suspicious participants show less social conformity than non-suspicious participants (Hertwig & Ortmann, 2008b). Here, the conformity consisted of contributing a certain amount to a common good, for example, and suspicion that one is not truly playing against other players may lead to more selfish behaviour (less conformity), leading to a systematic effect of suspicion. Yet, Jamison, Karlan, and Schechter (2008) found that, of the participants who returned after an initial experiment in which some were deceived and some were not, previously-deceived and previously-non-deceived participants did not differ in their generous behaviour in the prisoner's dilemma or in the dictator game in the second experiment. Barrera and Simpson (2012) also did not find differences in dictator-game or trust-game behaviour between previously-deceived and previously-non-deceived participants.

When it is clear what action to take if being deceived (e.g., keeping the entire amount in a dictator game without experiencing guilt, because one does not believe there is another player), the effect of suspected deception can be systematic. When it is unclear what action to take if being deceived, the suspicion of deception might increase non-systematic (i.e., random) variability (Hertwig & Ortmann, 2008b), which makes statistical tests conservative (Barrera & Simpson, 2012). Indeed, Jamison et al. (2008) found that previously-deceived participants showed more inconsistent behaviour on a risk-preferences measure in a subsequent study than non-previously-deceived participants. The variance in the number of safe gambles chosen in the former group was also higher than in the latter. Jamison et al. (2008) suggest that participants were not taking the experiments seriously as a result of having been deceived previously. However, Barrera and Simpson (2012) did not find a difference in variance or in the proportion of inconsistent responders between previously-deceived and previously-non-deceived participants on this same risk-preferences task. Possibly, the results presented by Jamison et al. (2008) were confounded by selection effects, as they found that previously-deceived females were less likely to participate in the second study than previously-non-deceived females. Barrera and Simpson (2012) avoided this potential influence of selection effects by making participants' participation credit contingent on participating in two studies and noted drop-out rates of 8% compared to 40% in the Jamison et al. (2008) study.

Although the evidence regarding the effects of deception is far from conclusive, the studies reported in this thesis did not use deception. Overall, avoiding the potential risk of polluting the subject pool with suspicion about the use of deception in future studies was considered more important than the convenience that deception would offer. An additional reason to avoid deception is that the use of deception would jeopardise the reputation of the EconLab at RHUL and would prohibit publication in economics journals (Cook & Yamagishi, 2008). Avoiding deception required creative and convoluted solutions to problems that would have easily been solved through deception. For example, most experiments necessitated an additional round to be used for payment. In this round, the presented information was drawn truly at random, at the time the experiment was conducted. The other rounds had predetermined sequences of information, which were equal across all sessions and across participants, to facilitate analyses (Bardsley, 2000). The studies in this thesis are unique in the length to which they go to maintain methodological purity, which, to the best of my knowledge, is not done to such an extent in other studies on the topics of this thesis (e.g., in the extensive literature on the beads task). Economics only proscribes active deception (Hertwig & Ortmann, 2008b), while withholding information is permitted (Barrera & Simpson, 2012). So, for several studies, it was possible to omit information: for example, participants were told they were paid for one of the rounds or for several questions, but not told which ones these were. Then, in the debrief it was explained that the round they were paid for was based on a random selection of information, while the other rounds were predetermined, yet still based on an initial random draw, to facilitate analyses.

The possibility of avoiding deception suggests that psychologists' frequent use of deception is unwarranted and violates the American Psychological Association's rule of only using deception in those cases where it is necessary and where non-deceptive methodologies will not suffice (Hertwig & Ortmann, 2008a). Indeed, deception is often used for convenience (Hertwig & Ortmann, 2008b). For example, in interactive games, when the behaviour of interest is that of only one of the two players, psychologists often deceive their participants into thinking that they are playing against another person, where in reality the other person's choice is simulated by a computer or is omitted altogether. This is done to minimise researchers' expenses for participants whose behaviour they are not interested in. For example, Shariff and Norenzayan (2007) investigated whether religious primes would affect monetary donations in the dictator game. Instead of truly randomly allocating participants to the roles of dictator or receiver, a confederate acted as the receiver, so that all participants were dictators. In this case, research expenses and logistics were minimised through deception.

This same expenses problem was encountered for the last study of this thesis, where the behaviour of interest was that of the sender in the trust game, but trustees had to be real people, rather than a computer. One possibility would have been to have people play both the role of sender and of trustees, and then randomly pair each participant with another player and randomly assign roles and pay the outcome made in the assigned roles. However, this has been found to negatively affect trustworthiness compared to trustee's decisions in trust games where each participant plays one role only (N. D. Johnson & Mislin, 2011). Therefore, our preference was to have participants play one role. This could have meant that a lot of data would have been collected and money would have been spent on participants whose data was not analysed (Cook & Yamagishi, 2008). Instead, after the trustees had made their relevant decisions for the trust game, they continued to provide data for an unrelated pilot study. This illustrates one way to avoid wasting researchers' money for data that is not directly related to the behaviour of interest, while maintaining methodological diligence.

### 1.7.2 Monetary Incentives

Economics and psychology also have different views on the use of incentives in experiments, so that generally economics experiments involve real monetary incentives, while psychology experiments involve no or merely hypothetical monetary incentives (Hertwig & Ortmann, 2001). Economists argue that monetary incentives motivate good performance (without leading to satiety, which might be the case for other incentives). They also provide a direct translation of economic normative theory, which assumes that maximisation of expected value drives behaviour (Ariely & Norton, 2007), to laboratory-based experiments (Hertwig & Ortmann, 2001). Incentives are also used to create an ecologically valid environment in the lab, as different decisions in the real world carry different costs and benefits (Rosenboim & Shavit, 2012). Furthermore, investigations of the effects of incentives have shown that incentives improve performance, when they have an effect, and in any case they reduce variability in performance, especially in studies on judgments and decision-making (Hertwig & Ortmann, 2001; V. L. Smith, 1991). On the basis of these arguments, economists argue that many observed deviations from rational behaviour can be explained by a lack of monetary incentives in the relevant studies (V. L. Smith, 1991).

Psychologists, on the other hand, tend to argue that the participants in most experiments' samples are cooperative, intrinsically motivated, and achievement-oriented, which will lead to maximal performance without the need for monetary incentives (Hertwig & Ortmann, 2001). External monetary incentives could reduce this intrinsic motivation. Furthermore, psychologists seem to argue that explicitly defining costs and rewards renders decisions made in the lab less ecologically valid (Ariely & Norton, 2007).

Investigations into the effects of extrinsic, monetary rewards have not been conclusive, but tend to suggest that their inclusion aids research on judgment and decision-making, if only by reducing variability (Hertwig & Ortmann, 2001). Differences between real or hypothetical rewards have not been found in delay discounting tasks (Madden et al., 2004), but have been found for risk preference tasks, albeit only for high rewards (Holt & Laury, 2002). Yet, in decisions involving social interaction, real rewards form a stronger incentive than hypothetical rewards and participants might overestimate their own and their partner's cooperation in hypothetical situations compared to real situations (Vlaev, 2012). For example, Parco, Rapoport, and Stein (2002) have shown that the size of incentives can affect behaviour in trust scenarios, which implies that the presence versus absence of incentives must also influence behaviour in such scenarios. One way in which incentives could influence behaviour is by increasing concentration on task stimuli and details, which may impair performance on attentional blink paradigms where increased attention is given to distracters (Bijleveld, Custers, & Aarts, 2011), but may potentially aid performance in judgment and decision-making paradigms. Another reason to include explicit extrinsic incentives is that performance of participants participating for course credit might vary across the term time, due to variation in intrinsic motivation across the term time, as various external pressures such as exams tap limited cognitive resources. Nicholls, Loveless, Thomas, Loetscher, and Churches (2014) used a sustained-attention paradigm, where participants had to respond to the presentation of all digits except "3". They found that course-credit participants' performance dropped from the start compared to the end of term, compared to paid participants. The contradictory effects of incentives might be explained by the type of monetary incentive. In a meta-analysis, Eisenberger and Cameron (1996) found that only rewards which were independent of performance (e.g., show-up fees) were detrimental to intrinsic motivation, while performance-based rewards, which focus on the quality of the performance, improved attitudes towards the task. Rewards for simply completing the task did not seem to affect intrinsic motivation.

In this thesis, Hertwig and Ortmann's (2001) recommendations to include incentives and avoid deception were followed, considering the potential impact they might have on the types of studies reported and the fact that studies were performed at different times during the year. Furthermore, in the study in Chapter 3, the inclusion of incentives was investigated as one of the independent variables, to address the question of the effect of incentives on the JTC bias empirically. Performance-based or decision-based incentives were included for all studies (except for one condition in the study in Chapter 3, where a participation incentive was used, to compare its effects to performance-based incentives). The average payment in each study was planned to be £8 to £10 per hour, in line with EconLab regulations. These real incentives could improve performance (Bardsley et al., 2010), and should reduce variability

(Hertwig & Ortmann, 2001). These incentives necessarily posed certain financial constraints on the amount of data that could be collected (Baron, 2001), in terms of sample size, study duration, and number of studies conducted within the thesis. Sample size was prioritised, as several studies required enough variation on one of the measures to assess how it related to other variables (e.g., delusion-proneness, beliefs regarding others' trustworthiness). Several variables were manipulated within subjects, as the inclusion of extra trials was deemed financially more efficient than adding another between-subject condition where additional time would be needed for participants to read instructions and receive payment.

## 1.8 Thesis Overview

In the next five empirical chapters, the specific theories and rationales relevant to the studies in question are elaborated upon. Here a brief overview is provided.

The first two studies (Chapters 2 and 3) focus on the widely cited claim that delusion-prone individuals "jump to conclusions" (JTC). Specifically, the studies involved *incentivised* draws-to-decision (Chapter 2) and probability-estimates (Chapter 3) versions of the beads task. The use of incentives is crucial given the potential confounding effect the lack of incentives may have had on motivation in previous studies. Moreover, in Chapter 2 incentives are used to generate optimal decision points, enabling the first evidence of absolute "jumping to conclusions" (as opposed to the standard, relative finding that delusion-prone individuals reach conclusions earlier than controls). These studies also minimised the influence of potentially impaired working memory on the JTC bias.

Chapter 2 contains two variants of the draws-to-decision version of the beads task. The first task involved a dynamic updating component to decision-making as information was presented sequentially and participants had to indicate at each step if they wanted to gather more information or decide. The second task required participants to indicate how much information they wanted to see, all at once, before seeing any information, in order to make their decision.

Chapter 3 investigates the effects of incentives on beliefs in a probability-estimates version of the beads task, while also investigating potential differences in reasoning with regards to prior probabilities and likelihoods. Again, the JTC bias is considered in a relative and absolute sense.

The study in Chapter 4 addresses potential limitations to predictions made by error-management theory (EMT) with regards to the sexual over-perception bias and to evidence adduced in support of it. First, the sexual over-perception bias might not be a biased belief, but rather could be merely behavioural in nature (McKay & Dennett, 2009). Second, results in support of this bias are found in paradigms that only present one piece of evidence, so that different prior beliefs might explain the different posterior beliefs, rather than a theoretically interesting cognitive bias. The study in this thesis investigates the sexual over-perception bias further in an adaptation of the updating paradigm used in Chapters 2 and 3.

In Chapter 5, it is explored whether the optimism bias is found in a crowd-within paradigm. If, in this updating paradigm without directional feedback, second estimates are systematically different from first estimates, this suggests estimates are updated in a desired direction. The study in Chapter 5 tests if such second estimates are systematically rosier than first estimates for undesirable questions, but not for neutral questions. In this way we shed light on the processes underpinning self-deception.

The last study, reported in Chapter 6, investigates whether betrayal aversion in the trust game is found in a design free from the methodological confounds of inequality aversion and the signalling of distrust, which may have led to mixed evidence for betrayal aversion. Furthermore, it tests whether the level of betrayal aversion could be related to prior beliefs about people's trustworthiness.

Finally, the various studies are discussed, implications and future suggestions are considered, and conclusions are reached in Chapter 7.

On a general note, an attempt has been made to move toward the "new statistics" (Cumming, 2014), by reporting effect sizes for all effects, including non-significant effects, and by reporting 95% confidence intervals (95%-CIs) of differences (and, where specified, of means) throughout the thesis. However, traditional null-hypothesis significance testing (NHST) is still employed as the main statistical analysis, because although Bayesian statistics are becoming more accessible (e.g., Masson, 2011), these accessible Bayesian methods still need to be fully validated (personal communication with Masson, April 29, 2014).

For the experiments reported in this thesis, established effect sizes were typically unavailable, limiting the utility of power calculations to determine the required sample sizes. Instead, we based our sample sizes on those used in the most relevant literature. For example, the sample sizes in the studies investigating the association between delusion-proneness and the JTC bias (Chapters 2 and 3) were equal to or larger than those reported in other JTC studies using sub-clinical populations (e.g., Balzan, Delfabbro, & Galletly, 2012; Colbert & Peters, 2002; LaRocco & Warman, 2009; McKay et al., 2006; Warman et al., 2007). As another example of determining the most relevant literature, consider the optimism study (Chapter 5), where the literature on the optimistic belief-updating bias consists largely of imaging studies, which tend to have small sample sizes (e.g., n=20 in Sharot et al., 2011). We therefore determined our sample size based on those used in the crowd-within literature, specifically on equivalent laboratory-based, rather than online, studies (n=101 in Herzog & Hertwig, 2009).

# 2 Delusion-Proneness and Data Gathering Biases[6]

## 2.1 Background

That delusional and delusion-prone individuals “jump to conclusions” (JTC) is one of the most important and influential claims in the literature on cognitive theories of delusions. As described in Chapter 1, the beads task is generally used to investigate the JTC bias. In this task, participants are shown two jars filled with beads of different colours, in opposite ratios across the jars. The jars are hidden and a sequence of beads is drawn from one of the two jars. In the standard draws-to-decision version of the beads task, participants have to decide which of the two jars the beads come from by requesting as many beads as they want from the sequence. Compared to control participants, delusional and delusion-prone participants make this decision on the basis of less evidence. As a result, a tendency to gather insufficient evidence when forming beliefs and making decisions is thought to be a core cognitive component of delusion formation (Fine et al., 2007; Garety & Freeman, 2013).

However, although the JTC effect is well replicated and robust to many modifications of the basic beads-task paradigm, there are some fundamental limitations with the way this task is typically administered that call the above interpretation into question. The key problem is that the terminology of “jumping to conclusions” implies that people gather *insufficient* evidence and reach decisions *prematurely*, yet the standard JTC bias, reported in over 50 studies (see reviews by Garety & Freeman, 1999, 2013), is of a *relative* nature:

[6] Part of this chapter has been published in Van der Leer, Hartig, Goldmanis, and McKay (in press).

delusional or delusion-prone participants request fewer pieces of evidence before deciding than healthy controls. However, just as one cannot conclude that bankers are poor because they earn less than movie stars, one cannot conclude delusional or delusion-prone participants decide *too early* because they decide earlier than healthy controls. The notions of premature or late decisions are only meaningful if there is an optimal point at which a rational person should decide. For an optimal decision point to exist, there needs to be both an (opportunity) cost of incorrect decisions and a cost associated with gathering more information. Investigations using the beads task typically do not incentivise participants (cf. Lincoln, Ziegler, Mehl, et al., 2010; Woodward et al., 2009), and no previous study has incorporated both of these elements.[7] Therefore, the most that can be said is that deluded and delusion-prone participants reach conclusions on this task more quickly than control participants, but the standard, non-incentivised, paradigm cannot justify the suggestion that they "jump to conclusions".

### 2.1.1 The Present Study

The present study consisted of two separate, but related tasks.

The first task ("dynamic task") incorporated a dynamic decision-making process where, after seeing each fish, participants had to choose whether to make their

[7] Lincoln, Ziegler, Mehl, et al. (2010) did create a decision threshold. Participants were given an initial ten tokens, each worth €0.25, and they could earn one more token by deciding on the correct jar or lose five tokens by deciding on the incorrect jar. To maintain or increase the level of rewards, five out of six decisions had to be correct, which meant that a decision should be made only *after* a certainty level of 83.3%. However, this incentive scheme did not generate a stopping rule. Without costs for gathering information, one should draw infinitely to reach the maximal level of certainty to base any decision on, which is the level participants are generally instructed to reach: "as certain as they could be as to which of the jars the beads were being drawn from" (Broome et al., 2007, p. s39).

decision about which lake the fish were coming from, or to see another fish. In this sense, it was similar to the standard draws-to-decision version of the beads task. New elements to the task were the rewards for a correct decision and small costs to see more data. Together, these two elements produced an optimal decision point within each sequence of fish, i.e., a point at which expected payoff would be maximal (see Appendix A).[8] Previous fish were shown as a memory aid. After each choice between more data or deciding on a lake, participants also rated their confidence that the fish were coming from either lake.

In contrast to the dynamic decision-making process in the first task, the second task (“static task”) had a one-shot decision-making process. Participants had to indicate how many fish they wanted to see before they saw any fish. Again, a small price was requested for each fish and a reward was provided for correct decisions, leading to optimal numbers of fish to request (see Appendix A). After indicating how many fish they wanted to see, participants were shown their requested number of fish all at once and were required to decide on one of the two lakes. As in the dynamic task, participants rated their confidence in the two lakes after making their decision regarding the two lakes.

In this second task, the effect of miscomprehension about swapping sources of the information was eliminated. It also eliminated the possibility that any JTC bias might be due to the additional time and effort required to see more information, as the one-shot nature of the task meant that equal amounts of time and effort were required to see a few or many fish.

[8] Strictly speaking these decision points were only “optimal” from a risk-neutral perspective. Risk-seeking or risk-averse individuals may have made decisions that failed to maximise their expected outcome, but that nevertheless maximised their expected utility (and thus were rational) given their risk preferences.

Both these tasks used incentives to create optimal decisions points to enable investigation of an absolute JTC bias, in addition to a relative JTC bias. This investigation formed the primary aim of the study presented in this chapter.

A secondary aim was to minimise several confounds, including mis-comprehension, motivation, and working memory. Both tasks address these confounds, but to different degrees. The dynamic task is very closely related to the standard paradigm, but minimises miscomprehension (see Balzan, Delfabbro, & Galletly, 2012) through detailed instructions and comprehension checks, and increases motivation by including incentives. The static task departs from the standard paradigm, but eliminates the confound of a lack of motivation in terms of effort and time. Furthermore, it minimises the working-memory load, as information does not need to be updated, but instead is presented all at once. An interesting effect of the means of presenting the data has been found for probability estimates provided by healthy controls: probability estimates were affected more strongly when information is presented gradually (i.e., stepwise selections presented sequentially) than when the entire selection is presented instantaneously (Whitman & Woodward, 2011). This was investigated only in healthy participants, so it is unknown if and how this manipulation would affect the JTC bias. However, the probability-estimates version used by Whitman and Woodward (2011) might be too weak to elicit the JTC bias, as the JTC bias is more robustly found with the draws-to-decision version (Fine et al., 2007; L. O. White & Mansell, 2009). Our second task is equivalent to a draws-to-decision version in which all the evidence is presented instantaneously and can thus shed light on the effect of presenting information instantaneously.

Finally, risk-aversion was measured and accounted for in analyses (cf. Lincoln, Ziegler, Mehl, et al., 2010). Additionally, given the unclear role of intelligence in performance on the beads task (Ziegler, Rief, Werner, Mehl, & Lincoln, 2008), intelligence was also measured and accounted for in analyses.

### 2.1.2 Hypotheses

The following findings were predicted across both tasks:

- First, a relative effect of delusion-proneness on draws-to-decision was expected such that the more delusion-prone participants were, the less evidence they would gather for their decision.
- Second, an absolute effect of delusion-proneness was also expected, where high-delusion-prone participants would jump to conclusions with respect to the optimal decision point, but low-delusion-prone participants would not.
- Third, a relative effect of delusion-proneness on confidence levels was expected, such that the more delusion-prone participants were, the more confident they would be in their decision (as in, e.g., McKay et al., 2006).

## 2.2 Methods

### 2.2.1 Participants

Participants (n=115, 60 females, 55 males; mean (SD) age = 19.92 (2.9) years) were students from RHUL, recruited using the Online Recruitment System for Economic Experiments (Greiner, 2004). Participants received £6 for participation and a performance-based bonus between £0 and £7.50 (mean (SD) bonus = £6.35 (£1.36)). The Psychology Department Ethics Committee of RHUL approved this study.

### 2.2.2 Materials

#### 2.2.2.1 *Fisherman-Adapted Beads Task*

Before the start of the experiment, participants were issued written instructions providing information about the ratios of black to white fish in the two lakes (i.e., 25%:75% or 75%:25%), the facts that the fisherman never visited both lakes on one trip and was equally likely to visit Lake A as he was to visit Lake B, the rewards and costs, and the means of responding (see Appendix B). Figure 2.1 shows an example of a trial in this task. Participants had to correctly answer eleven comprehension questions before starting the experiment (see Appendix B). Answers were checked and participants were referred back to the written instructions when an incorrect answer was given. Visual stimuli were based on those used by Speechley et al. (2010).

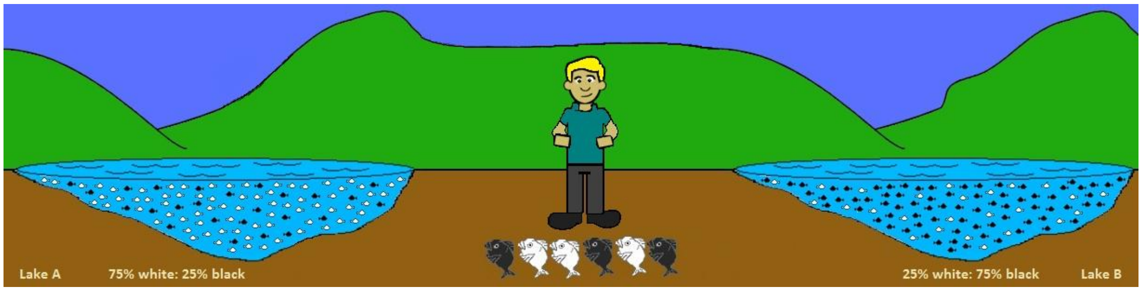


**Figure 2.1** An example of a trial in the dynamic beads task. The fisherman displays a series of six fish (B-W-W-B-W-B), which he has caught from one of the two lakes. Lake A has more white fish than black fish (represented visually and by a stated ratio), while Lake B has more black than white fish.

In the dynamic task, participants saw one fish and were then given the option to choose between deciding on a lake and winning a reward if correct, or seeing another fish for a small price. If participants chose to see another fish, this procedure was repeated. In each series, they could request to see at least eleven more fish; with some series going up to seventeen fish, although this maximum was not conveyed to participants to avoid decisions being made based on the expected ends of a trial (cf. Lincoln, Ziegler, Mehl, et al., 2010). Upon seeing the last fish of a trial, participants had to make a decision between Lake A and Lake B. After seeing each fish, participants rated their confidence in either lake. Separate confidence ratings were requested for Lake A and for Lake B (Speechley et al., 2010). This was repeated for seven series of fish; each series was presented as a new fishing trip, further indicated by the fisherman's shirt changing colour.

The first trial consisted of a randomly drawn series of fish[9], which was the same for all participants in one session, but differed across sessions. This trial was used for payment: participants were informed that the series of fish on one trial would be drawn at random and that they would be paid based on their performance on this trial; they did not, however, know which of the trials was the random trial. By including this truly random sequence we avoided deceiving our participants and thus conformed to a key methodological principle of experimental economics (Hertwig & Ortmann, 2001). This trial was not used for analyses, as it differed per session.

The next five trials (trials A-E in Table 2.1) were tailored to have varying optimal decision points, which were reached at a difference of three fish of the majority colour over the minority colour (see Appendix A). The sequences of the trials were drawn based on the same process as that used for the randomly drawn sequence used for payment, except that after the optimal decision point, all subsequent fish in the series were changed to the majority colour. This strengthened the stopping rule and we no longer introduced contradictory

[9] One lake was selected at random, each with a probability of .5. The sequence of fish was drawn at random from the selected lake, in accordance with the ratios of black and white fish in that lake (e.g., if the mostly black lake was selected, black fish were presented with a probability of .75 and white fish with a probability of .25).

evidence after the optimal decision point. The seventh trial (trial F in Table 2.1) only showed fish of one colour, in order to investigate decision making without contradictory evidence. For all trials, the potential bonus after seeing one fish was 100 points; the cost per fish was 2 points. Each experimental point was worth £0.05.

**Table 2.1** The sequences of fish in the trials of the dynamic task; optimal decision points are underlined.

| Trial | #1 | #2 | #3 | #4 | #5 | #6 | #7 | #8 | #9 | #10 | #11 | #12 | #13 | #14 | #15 | #16 | #17 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **A** | B | W | B | B | $\underline{B}$ | B | B | B | B | B | B | B | B | B | B | | |
| **B** | B | W | W | W | $\underline{W}$ | W | W | W | W | W | W | W | W | W | | | |
| **C** | B | W | B | B | W | B | $\underline{B}$ | B | B | B | B | B | | | | | |
| **D** | W | B | B | W | B | B | $\underline{B}$ | B | B | B | B | B | B | B | B | B | B |
| **E** | B | W | B | W | W | B | B | B | $\underline{B}$ | B | B | B | B | B | B | | |
| **F** | W | W | $\underline{W}$ | W | W | W | W | W | W | W | W | W | W | | | | |

In the static task, participants saw the fisherman and the ratios of black to white fish in the two lakes; these ratios were the same as in the dynamic task. Before seeing any fish, participants were then asked to indicate how many fish they would like to see before deciding between Lake A and B. They could choose between one and ten fish. Next, they were shown their requested number of fish and were required to decide between the two lakes. Participants rated their confidence in either lake after their decision between the lakes. Separate confidence ratings were requested for Lake A and for Lake B (Speechley et al., 2010). This was repeated for five series (trials G-K in Table 2.2); each series was presented as a new fishing trip on which the potential reward and the costs per fish differed. The combination of black and white fish shown for a given requested number of fish was randomly drawn based on the 25:75 ratios before the experiment began; this sequence was then used in all sessions to facilitate analyses of confidence levels. A sixth trial with a truly randomly drawn combination of fish of the requested number was included for payment, but not included in the analyses as it differed per session. Table 2.2 shows the predetermined fish, the rewards and costs, and the optimal decision points.

**Table 2.2** The rewards and costs and the colours of the fish in the trials of the static task, optimal decision points are underlined.

| Trial | Reward | Price per fish | #1 | #2 | #3 | #4 | #5 | #6 | #7 | #8 | #9 | #10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **G** | 100 | 2 | W | B | W | W | W | W | W | B | W | W |
| **H** | 50 | 3 | B | B | B | B | W | W | B | B | B | B |
| **I** | 200 | 2 | B | W | B | W | W | W | W | W | B | W |
| **J** | 50 | 1 | W | B | W | W | W | W | W | W | W | B |
| **K** | 50 | 2 | B | W | B | W | B | B | B | B | B | B |

### 2.2.2.2 *Risk Aversion*

We administered a computerised risk-aversion measure (Holt & Laury, 2002), see Table 2.3. This measure involved ten decisions between two gambles. For example, the seventh decision was between Option A "A 7/10 chance of winning £2.00, a 3/10 chance of winning £1.60" and Option B "A 7/10 chance of winning £3.85, a 3/10 chance of winning £0.10" (Holt & Laury, 2002, based on p. 1645). Starting at 10%, the probability of the high payoff outcome in each gamble increased by 10% in each successive decision, such that the expected value of the "risky" Option B increased. Each participant's risk preference score was defined by their switching point from Option A to Option B (i.e., the number of times they chose Option A). Risk-neutral responding implied choosing Option A for the first four decisions, and choosing Option B thereafter. Choosing Option A fewer than four times implied risk-seeking behaviour, while choosing it more than four times indicated risk-averse behaviour.

**Table 2.3** The risk aversion measure by Holt and Laury (2002, p. 1645). The choices indicate risk neutrality.

| Option A | | Option B |
|---|---|---|
| 1/10 chance of £2.00,<br>9/10 chance of £1.60 | ← | 1/10 chance of £3.85,<br>9/10 chance of £0.10 |
| 2/10 chance of £2.00,<br>8/10 chance of £1.60 | ← | 2/10 chance of £3.85,<br>8/10 chance of £0.10 |
| 3/10 chance of £2.00,<br>7/10 chance of £1.60 | ← | 3/10 chance of £3.85,<br>7/10 chance of £0.10 |
| 4/10 chance of £2.00,<br>6/10 chance of £1.60 | ← | 4/10 chance of £3.85,<br>6/10 chance of £0.10 |
| 5/10 chance of £2.00,<br>5/10 chance of £1.60 | → | 5/10 chance of £3.85,<br>5/10 chance of £0.10 |
| 6/10 chance of £2.00,<br>4/10 chance of £1.60 | → | 6/10 chance of £3.85,<br>4/10 chance of £0.10 |
| 7/10 chance of £2.00,<br>3/10 chance of £1.60 | → | 7/10 chance of £3.85,<br>3/10 chance of £0.10 |
| 8/10 chance of £2.00,<br>2/10 chance of £1.60 | → | 8/10 chance of £3.85,<br>2/10 chance of £0.10 |
| 9/10 chance of £2.00,<br>1/10 chance of £1.60 | → | 9/10 chance of £3.85,<br>1/10 chance of £0.10 |
| 10/10 chance of £2.00,<br>0/10 chance of £1.60 | → | 10/10 chance of £3.85,<br>0/10 chance of £0.10 |

#### 2.2.2.3 *Delusion Proneness*

We used the 21-item Peters et al. Delusions Inventory (PDI; Peters, Joseph, Day, & Garety, 2004) to measure delusional ideation, with items such as "Have your thoughts ever been so vivid that you were worried other people would hear them?" and "Do you ever feel as if there is a conspiracy against you?". If an item was endorsed, participants also had to indicate their level of distress, preoccupation and conviction for that item on 5-point scales. The yes/no endorsement summed scores could range from 0-21; the separate dimensions' summed scores from 0-105; and the total summed score from 0-336.

#### 2.2.2.4 *Intelligence*

We used the short 12-item version of Raven's advanced progressive matrices (APM; Arthur & Day, 1994) to measure intelligence. This measure has a low sensitivity to age, gender, and cultural differences (Colbert & Peters, 2002), which are factors not expected to affect the experimental tasks. The 12 items in the short form of the APM are items 1, 4, 8, 11, 15, 18, 21, 23, 25, 30, 31, and 35 from Set II of the long, 36-item APM. The psychometric properties of these items are similar to all 36 items of the full APM (Arthur & Day, 1994). A time limit of 15 minutes was set for these twelve items, based on the average duration Arthur and Day (1994) measured. The example item was item 8 from Set I of the APM, while verbal instructions based on the manual (Raven, Court, & Raven, 1992) were also displayed on the screen. Intelligence was defined as the total number of correctly answered items of the Raven's 12-APM (APM-scores).

### 2.2.3 Procedure

The experiment lasted approximately 75 minutes. Participants were tested in groups ranging from 20-26 people. All sessions were conducted on a local computer network using z-Tree software (Fischbacher, 2007) in the EconLab at RHUL. Before the experiment began, participants provided written informed consent. Next, they were issued detailed instructions and after correctly answering the comprehension questions, participants completed the dynamic task. They then read instructions regarding the static task and completed the task after correctly answering the comprehension questions. The static task was always conducted after the dynamic task, to avoid confusion about the rules of the task, in view of the contribution miscomprehension seems to make to the JTC bias (Balzan, Delfabbro, & Galletly, 2012; Balzan, Delfabbro, Galletly, et al., 2012). The tasks were analysed separately, given potential order effects.

The middle five trials of the dynamic task and the first five trials of the static task were presented in counterbalanced order across testing sessions. Incomplete counterbalancing was accomplished through Latin square counterbalancing, using maximally different orders for each session (see Table 2.4). To discourage participants from making formal calculations (e.g., using phones), each decision (i.e., seeing more fish or choosing a lake) had a time limit of 20 seconds. The first decision of each task had a time bonus of 40 seconds, to allow participants to get acquainted with the task. After the two tasks, participants completed the risk-aversion measure, the 12-APM, the PDI, and some demographic questions (gender, age, field of study). At the conclusion of the sessions, participants were paid their earnings (i.e., show-up fee and potential bonus) and dismissed.

**Table 2.4** Latin-square counterbalanced orders of the analysed trials in the dynamic and static task across the testing sessions.

| **Session** | **Task** | **First trial** | **Second trial** | **Third trial** | **Fourth trial** | **Fifth trial** |
|---|---|---|---|---|---|---|
| **1** | Dynamic | A | B | C | D | E |
| | Static | G | H | I | J | K |
| **2** | Dynamic | B | D | A | E | C |
| | Static | H | I | K | G | J |
| **3** | Dynamic | C | E | D | B | A |
| | Static | I | K | J | H | G |
| **4** | Dynamic | D | A | E | C | B |
| | Static | J | G | H | K | I |
| **5** | Dynamic | E | C | B | A | D |
| | Static | K | J | G | I | H |

### 2.2.4 Statistical Analyses

The dependent variable on each trial was the number of fish seen before making a decision (i.e., draws to decision in the dynamic task and number of fish requested in the static task). The number of fish seen before making a decision was averaged across the five counterbalanced trials (separately for the dynamic and static tasks). Furthermore, the number of fish seen before deciding in a completely white sequence was a separate dependent variable. We analysed this sequence separately because it may have been prone to order effects as it was always presented as the last trial (in case the sequence raised suspicion about the randomness of the draw).

To investigate the relative JTC bias, regression analyses were conducted with delusion-proneness (i.e., continuous PDI-scores) as the predictor and draws to decision as the outcome variable (McKay et al., 2006). Hierarchical regression analyses were also conducted, where the first model accounted for intelligence (APM-scores), which may influence the ability to calculate the optimal decision point, and for risk-aversion, as decisions had financial consequences. The second model investigated if delusion-proneness could explain additional variance in the draws to decision.

Furthermore, we used one-sample *t*-tests to investigate an absolute JTC bias for the whole sample, and for low-delusion-prone and high-delusion-prone participants separately. To make this latter distinction, we used a median split to convert the continuous PDI-scores into a categorical factor, excluding participants who fell at the median (LaRocco & Warman, 2009; Warman et al., 2007). The mean draws-to-decision of each group was compared to the optimal number of draws before deciding.

Finally, we investigated whether decisions were made with higher confidence by high-delusion-prone participants than by low-delusion-prone participants. The confidence in the chosen lake was used as the dependent variable. We also investigated whether draws-to-decision and confidence levels were correlated. Due to insufficient power, we could not analyse confidence levels for each fish in a sequence, except for the first fish in each sequence. We analysed if delusion-proneness was associated with increased confidence levels for the first fish, averaged across the five sequences in the dynamic task, and for the completely white sequence separately.

## 2.3 Results

### 2.3.1 Data Screening

First, the data was inspected for outliers based on Cook's distances, Mahalanobis' distances, and standardised residuals with values outside the range from -2 to 2. Three participants were outliers across the three analyses. These were excluded, reducing the sample size to 112. One participant's PDI-score was an outlier according to the Mahalanobis distance. However, this participant's score (i.e., 253) was well within the full possible range (i.e., 1-336) of the variable of interest. Therefore, the PDI-scores were square-root transformed. After this transformation no more outliers were detected through Cook's and Mahalanobis' distances and the standardised residuals indicated that fewer than the allowed 5% of the participants were outliers (Field, 2009).

Further statistical assumptions were checked for n=112. Absence of multicollinearity was confirmed by the fact that there were no strong correlations between predictors, the tolerance values were >.966, and the VIF values were all <1.036. The predictors were linearly related with the outcome, as inclusion of the squared predictors did not lead to significantly better models. The standardised residuals were normally distributed. Homoscedasticity was confirmed as the plots of standardised residuals and predicted scores showed that the variance was equal across the range of the predicted scores. No violations of assumptions were detected and analyses were thus conducted using this sample of n=112.

### 2.3.2 Descriptive Statistics

Table 2.5 shows descriptive statistics for age, PDI-scores, APM-scores, risk-aversion, gender, and subject of study.

**Table 2.5** Descriptive statistics for age, PDI-scores, APM-scores, and risk-aversion.

| Variable | Subcategory | N | Median | Mean | Standard deviation | Range |
|---|---|---|---|---|---|---|
| **Age** | | 112 | 19 | 19.94 | 2.920 | 17-35 |
| **PDI-scores** | | 112 | 68.5 | 75.06 | 40.611 | 10-253 |
| **APM-scores** | | 112 | 7 | 6.62 | 2.945 | 1-12 |
| **Risk-aversion** | | 112 | 6 | 5.41 | 1.732 | 0-9 |
| **Gender** | Female | 59 (52.7%) | | | | |
| | Male | 53 (47.3%) | | | | |
| **Subject studied** | Economics | 19 (17.0%) | | | | |
| | Mathematics | 11 (9.8%) | | | | |
| | Psychology | 12 (10.7%) | | | | |
| | Various | 70 (62.5%) | | | | |

For the one-sample *t*-tests used to investigate absolute JTC, we employed a median split to classify participants as low-delusion-prone or high-delusion-prone. Although the median in the current sample (i.e., 68.5) was lower than that generally reported in the literature (e.g., Balzan, Delfabbro, & Galletly, 2012; LaRocco & Warman, 2009; Warman et al., 2007; 75.5, 90, and 97.5, respectively), it was higher than the median score of 49 originally found by Peters et al. (2004).

Participants from certain educational backgrounds (i.e., economics, psychology, or mathematics) might have approached the tasks in this experiment differently. However, these three subjects were equally represented ($\chi^2(2)=2.714$, $p=.257$, $\phi_C=.110$), with the majority of the sample coming from various other educational backgrounds that were not expected to influence performance in this experiment (e.g., management, history, or geology; see Table 2.5).

## 2.3.3 Dynamic Task

### *2.3.3.1 Draws to Decision*

A linear regression showed that delusion-proneness was a significant predictor of the number of draws to decision ($F(1,110)=5.520$, $p=.021$, $R^2_{\text{ADJUSTED}}=.039$; see Table 2.6, model 1). The more delusion-prone a participant was, the fewer draws they requested before deciding. A hierarchical linear regression accounted for risk-aversion and intelligence. The first step included risk-aversion and intelligence, which significantly explained 19.4% of the variance ($F(2,109)=14.373$, $p<.001$, $R^2_{\text{ADJUSTED}}=.194$). Risk-aversion was not a significant predictor, while intelligence was (see Table 2.6, model 2.1). Adding delusion-proneness as a third predictor significantly improved the model ($\Delta F(1,108)=6.128$, $\Delta p=.015$, $\Delta R^2=.042$; $F(3,108)=12.076$, $p<.001$, $R^2_{\text{ADJUSTED}}=.230$). Delusion-proneness and intelligence were significant predictors in this model, while risk-aversion was not (see Table 2.6, model 2.2). Figure 2.2 shows the relationships between the predictors and draws to decision in the dynamic task.

One-sample *t*-tests were conducted to investigate the absolute JTC bias. These tests indicated that the mean draws to decision, across trials, were significantly different from optimal (i.e., 6.6) for the whole sample (mean (SE) difference = -2.566 (.165); $t(111)=15.506$, $p<.001$, $d=2.944$, 95%-CI [-2.894, -2.238]). Moreover, fewer fish than optimal were considered by both low-delusion-prone participants (-2.286 (0.220); $t(55)=10.377$, $p<.001$, $d=2.798$, 95%-CI [-2.727, -1.844]) and high-delusion-prone participants (-2.846 (0.243); $t(55)=11.702$, $p<.001$, *d*=3.156, 95%-CI [-3.334, -2.359]). This indicates that *both* groups decided *prematurely* in comparison with an objective, rationally optimal decision point.

**Table 2.6** B-values, standard errors (SE), β-values, and *p*-values for each of the predictors in the steps of the regression models for the draws to decision in the dynamic task. Square-root transformed PDI-scores represented delusion-proneness; APM-scores represented intelligence.

| Model | Predictor | b | 95%-CI of b | SE | β | *p* |
|---|---|---|---|---|---|---|
| **1** | Delusion-proneness | -.166 | [-0.307, -0.026] | .071 | -.219 | .021 |
| **2.1** | Intelligence | .271 | [0.171, 0.372] | .051 | .456 | <.001 |
| | Risk-aversion | .009 | [-0.162, 0.180] | .086 | .009 | .913 |
| **2.1** | Intelligence | .269 | [0.170, 0.367] | .050 | .452 | <.001 |
| | Risk-aversion | -.029 | [-0.199, 0.141] | .086 | -.029 | .737 |
| | Delusion-proneness | -.160 | [-0.287, -0.032] | .064 | -.210 | .015 |

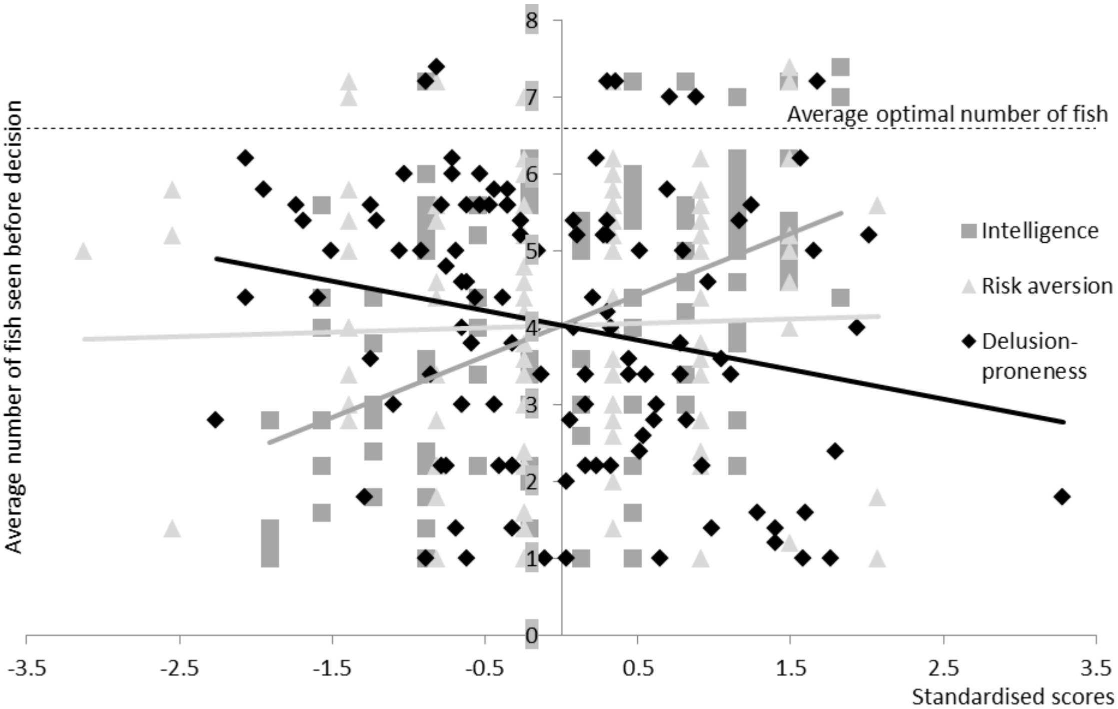


**Figure 2.2** Relationships between the predictors (squares: intelligence as represented by APM-scores; triangles: risk-aversion; diamonds: delusions-proneness as represented by square-root transformed PDI-scores) and the number of fish seen before deciding in the dynamic task. The distances from the dotted line (the optimal number of fish to have seen before deciding) indicates the deviation from the optimal decision point.

#### 2.3.3.2 *Confidence Levels*

Confidence levels at the moment of deciding were significantly correlated with draws to decision (*r*(112)=.304, *p*=.001), so that the more fish were seen before deciding, the more confident participants felt. A linear regression showed that delusion-proneness did not significantly predict confidence levels at the moment of deciding (*F*(1,110)=.182, *p*=.671, $R^2_{ADJUSTED}$=-.007; *b* (SE)=.245 (.575), 95%-CI of *b* [-0.894, 0.385], β=.041). Another linear regression showed that delusion-proneness also did not significantly predict confidence levels after seeing the first fish in the sequence (*F*(1,110)=2.063, *p*=.154, $R^2_{ADJUSTED}$=.009; *b* (SE)=1.113 (.775), 95%-CI of *b* [-0.422, 2.647], β=.154).

#### 2.3.3.3 *Completely White Sequence of the Dynamic Task*

This completely white sequence minimises the need to integrate information, which is thought to be one impairment contributing to the JTC bias (Fine et al., 2007; Young & Bentall, 1995). As this sequence (trial F) was always presented last, there may have been practice effects or fatigue on this trial compared to the counterbalanced trials A-E and it was thus analysed separately.

##### *2.3.3.3.1 Draws to Decision*

As with the above analyses for the dynamic task's sequences with contradictory information, a linear regression conducted for the sequence of only white fish indicated that delusion-proneness was a significant predictor of draws to decision (*F*(1,110)=7.429, *p*=.007, $R^2_{ADJUSTED}$=.055; see Table 2.7, model 1). The more delusion-prone a participant was, the fewer fish they saw before deciding. As above, a hierarchical linear regression accounting for risk-aversion and intelligence was run to investigate if adding delusion-proneness could predict additional variance not already accounted for. The first step of this model, including risk-aversion and intelligence, significantly explained 8.7% of the variance (*F*(2,109)=6.321, *p*=.003, $R^2_{ADJUSTED}$=.087). Risk-aversion was not a significant predictor, while intelligence was (see Table 2.7, model 2.1). Adding delusion-proneness as a third predictor significantly improved the model (Δ*F*(1,108)=6.128, Δ*p*=.008, Δ$R^2$=.057; *F*(3,108)=6.883, *p*<.001, $R^2_{ADJUSTED}$=.137). Delusion-proneness and intelligence were significant predictors in this model, while risk-aversion was not (see Table 2.7, model 2.2).

One-sample *t*-tests were conducted to investigate an absolute JTC bias on this sequence as was done for the sequences with contradictory information above. These tests indicated that the mean draws to decision were significantly different from optimal (i.e., 3) for the whole sample (mean (SE) difference = -0.357 (.165); *t*(111)=3.702, *p*<.001, *d*=.703, 95%-CI [-0.548, -0.166]). When split by delusion-proneness, only high-delusion-prone participants showed an absolute JTC bias (-.554 (0.146); *t*(55)=3.786, *p*<.001, *d*=0.719, 95%-CI [-0.847, 0.261]). Low-delusion-prone participants requested an optimal number of fish on this completely white sequence (-.161 (0.122); *t*(55)=1.322, *p*=.192, *d*=0.251, 95%-CI [-0.404, -0.083]).

**Table 2.7** B-values, standard errors (SE), β-values, and *p*-values for each of the predictors in the steps of the regression models for the draws to decision in the completely white sequence of the dynamic task. Square-root transformed PDI-scores represented delusion-proneness; APM-scores represented intelligence.

| Model | Predictor | b | 95%-CI for b | SE | β | *p* |
|---|---|---|---|---|---|---|
| **1** | Delusion-proneness | -.112 | [-0.193, -0.030] | .041 | -.252 | .007 |
| **2.1** | Intelligence | .110 | [0.048, 0.172] | .031 | .317 | .001 |
| | Risk-aversion | .025 | [-0.081, 0.131] | .054 | .043 | .639 |
| **2.2** | Intelligence | .108 | [0.048, 0.169] | .031 | .312 | .001 |
| | Risk-aversion | -.001 | [-0.105, 0.104] | .053 | -.001 | .992 |
| | Delusion-proneness | -.107 | [-0.186, -0.028] | .040 | -.242 | .008 |

#### *2.3.3.3.2 Confidence Levels*

As for the sequences with contradictory information above, confidence levels at the moment of deciding were significantly correlated with the draws to decision in the completely white sequence (*r*(112)=.400, *p*<.001), so that the more fish were seen before deciding, the more confident participants felt. A linear regression showed that delusion-proneness did not significantly predict confidence levels at the moment of deciding (*F*(1,110)=.014, *p*=.906, $R^2_{ADJUSTED}$= -.009; *b* (SE)=.092 (.776), 95%-CI of *b* [-1.447, 1.630], β=.011). Another linear regression showed that delusion-proneness significantly predicted confidence levels after seeing the first fish in the sequence (*F*(1,110)=5.803, *p*=.018, $R^2_{ADJUSTED}$=.050; *b* (SE)=1.979 (.822), 95%-CI of *b* [0.351, 3.607], β=.224).

## 2.3.4 Static Task

### *2.3.4.1 Draws to Decision*

As for the dynamic task above, a linear regression showed that delusion-proneness was a significant predictor of draws to decision (*F*(1,110)=5.054, *p*=.027, $R^2_{ADJUSTED}$=.035; see Table 2.8, model 1). The more delusion-prone participants were, the fewer fish they requested. Again, a hierarchical linear regression accounted for risk-aversion and intelligence. The first step, including risk-aversion and intelligence, significantly explained 10.8% of the variance (*F*(2,109)=14.373, *p*=.001, $R^2_{ADJUSTED}$=.108). Risk-aversion was not a significant predictor, while intelligence was (see Table 2.8, model 2.1). Adding delusion-proneness as a third predictor significantly improved the model (Δ*F*(1,108)=4.288, *p*=.041, Δ$R^2$=.033; *F*(3,108)=6.748, *p*<.001, $R^2_{ADJUSTED}$=.134). Intelligence and delusion-proneness were significant predictors in this model, while risk-aversion was not (see Table 2.8, model 2.2). Figure 2.3 shows the relationships between the predictors and draws to decisions in the static task.

**Table 2.8** B-values, standard errors (SE), β-values, and *p*-values for each of the predictors in the steps of the regression models for mean deviations from optimal decisions in the static task.

| **Model** | **Predictor** | **b** | **95%-CI for b** | **SE** | **β** | ***p*** |
|---|---|---|---|---|---|---|
| **1** | Delusion-proneness | -.167 | [-0.313, -0.020] | .074 | -.210 | .027 |
| **2.1** | APM-scores | .205 | [0.095, 0.316] | .056 | .331 | <.001 |
| | Risk-aversion | .113 | [-0.074, 0.301] | .095 | .107 | .234 |
| **2.2** | APM-scores | .203 | [0.094, 0.312] | .055 | .327 | <.001 |
| | Risk-aversion | .078 | [-0.110, 0.266] | .095 | .074 | .413 |
| | Delusion-proneness | -.148 | [-0.289, -0.006] | .071 | -.186 | .041 |

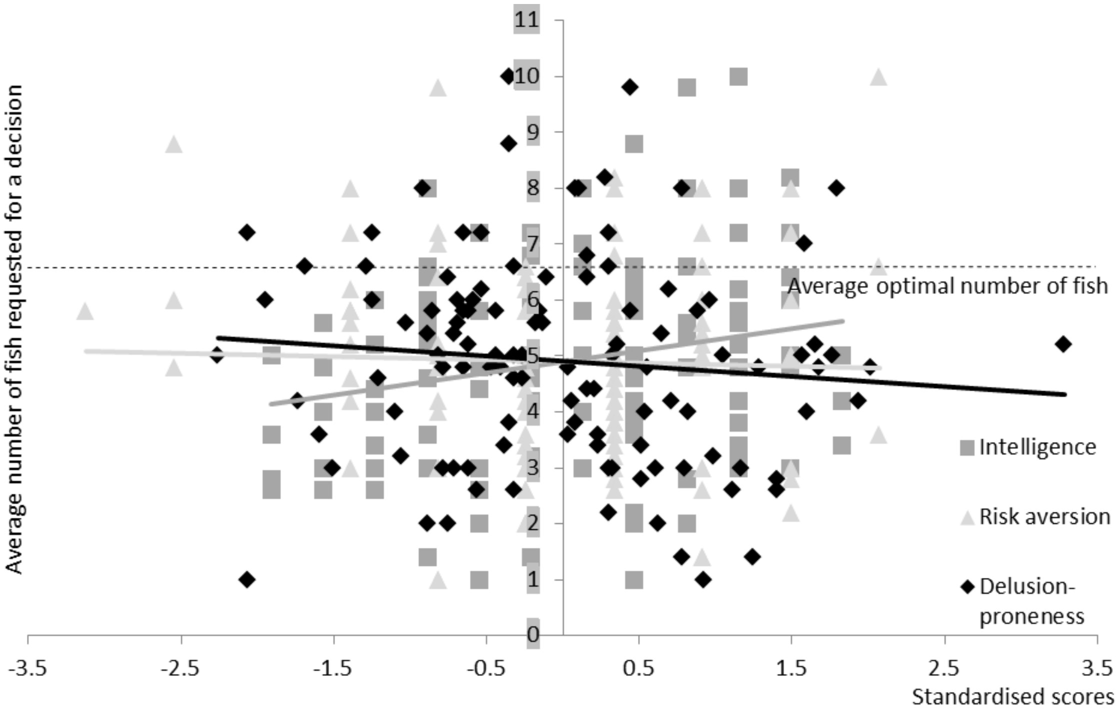


**Figure 2.3** Relationships between the predictors (squares: intelligence as represented by APM-scores; triangles: risk-aversion; diamonds: delusion-proneness as represented by square-root transformed PDI-scores) and the number of fish requested in the static task. The distances from the dotted line (the optimal number of fish to have requested) indicates the deviation from the optimal decision.

As for the dynamic task, one-sample *t*-tests were conducted to investigate absolute JTC in the static task. These tests indicated that the mean draws-to-decision were significantly different from optimal (i.e., 5.4) for the whole sample (mean (SE) difference = -.496 (.173); *t*(111)=2.873, *p*=.005, *d*=0.545, 95%-CI [-0.839, -0.154]). When split by delusion-proneness, only high-delusion-prone participants showed an absolute JTC bias (-.739 (0.257); *t*(55)=2.880, *p*=.006, *d*=0.547, 95%-CI [-1.254, -0.225]). Low-delusion-prone participants requested an optimal number of fish (-.254 (0.229); *t*(55)=1.107, *p*=.273, *d*=0.210, 95%-CI [-0.713, 0.206]).

#### 2.3.4.2 Confidence Levels

As for the dynamic task, confidence levels at the moment of deciding were significantly correlated with the amount of data gathered in the static task (*r*(112)=.341, *p*<.001), so that the more fish were requested for a decision, the more confident participants felt about their decision. A linear regression indicated that delusion-proneness did not significantly predict confidence levels at the moment of deciding (*F*(1,110)=1.099, *p*=.297, $R^2_{\text{ADJUSTED}}$=.001; *b* (SE)=-.490 (.467), 95%-CI of *b* [-1.416, 0.436], β=-.099).

## 2.4 Discussion

In this study, the JTC bias was investigated in incentivised dynamic and static decision-making tasks. The dynamic task involved sequential presentation of information, where after each fish, which could be black or white, the participants could ask to see another fish or could decide on one of two lakes as the source of the fish. In the static task, in contrast, participants had to indicate how many fish they would like to see, all at once, before seeing any fish. The dynamic task is modelled after the classic beads task, which is the most commonly used paradigm to investigate the JTC bias.

In the dynamic task, delusion-proneness predicted the draws taken to reach a decision, both before and after accounting for risk-aversion and intelligence. The higher the scores for delusion-proneness, the fewer fish participants saw before deciding. This was found across trials which included conflicting information within the sequences and also for a sequence of only white fish. This provides support for a relative JTC bias. Our entire sample also showed an absolute JTC bias in the dynamic task: both low-delusion-prone and high-delusion-prone participants saw less evidence than would have been optimal. Delusion-proneness and confidence ratings were not robustly associated in the dynamic task.

In the static task, delusion-proneness also predicted the amount of evidence requested for a decision, both before and after accounting for risk-aversion and intelligence. As with the dynamic task, the higher the scores for delusion-proneness, the fewer fish participants requested to see to base their decision on. This also supports the relative JTC bias finding. In this task, only high-delusion-prone participants showed an absolute JTC bias, as they requested significantly fewer fish than would have been optimal. Low-delusion-prone participants requested an optimal number of fish. Delusion-proneness and confidence ratings were not associated in the static task.

The standard *relative* JTC-finding was found in *both* tasks: the more delusion-prone participants were, the less evidence they used as a basis for a decision. Interestingly, an *absolute* JTC bias was found in the dynamic task for the majority of participants, as *both* low-delusion-prone and high-delusion-prone participants tended to decide faster than they should have (in order to maximise their expected payoff). In the static task, only high-delusion-prone participants were found to decide on the basis of less information than optimal, while low-delusion-prone participants decided on the basis of an optimal amount of information. This suggests that delusion-proneness is also associated with an absolute JTC bias.

It must be noted that it appears that the absolute JTC bias could be influenced by the task environment to a certain extent, as it was found for everyone in the dynamic task, but only for high-delusion-prone participants in the static task. This suggests the absolute JTC bias could be exacerbated by potentially impaired cognitive capacity. If updating beliefs as one gathers information in the dynamic task is too taxing, early decisions might be made to avoid having to maintain and process information in working memory. A lack of motivation might also exacerbate the JTC bias on the dynamic task compared to the static task, as seeing more fish on the dynamic task would prolong the task and require more responses, whereas the amount of time and effort that had to be invested on the static task was the same regardless of the number of fish seen. These effects might be more pronounced for participants higher in delusion-proneness.

The liberal acceptance account of the JTC bias suggests that delusional or delusion-prone participants have a lower decision threshold and thus do not require as high a level of probability of being correct as healthy controls before deciding (e.g., Moritz et al., 2007). This, in turn, would suggest that confidence levels at the moment of deciding would be lower for participants who are more delusion-prone. However, McKay et al. (2006) and Warman (2008) found that participants who were more delusion-prone reported greater confidence, a finding in the opposite direction to that suggested by the liberal acceptance account. We did not find confidence levels to be associated with delusion-proneness. Langdon, Ward, and Coltheart (2010) and Langdon, Still, Connors, Ward, and Catts (2014) did not find a difference in confidence ratings between delusional patients and controls either. This might be due to confidence ratings' reduced sensitivity to the JTC bias compared to a draws-to-decision measure (Bentall, Corcoran, Howard, Blackwood, & Kinderman, 2001; Fine et al., 2007).

Fine et al. (2007) delineate several accounts of the JTC bias. The first two accounts are based on disturbed information integration. First, the account based on work by Menon et al. (2006) stipulates that evidence in the beads task is assigned extra weight, because stimuli acquire extra salience in patients with schizophrenia (Kapur, 2003). This would lead to early decisions and increased confidence in the hypothesis being held for any given piece of information. Although we did find a reduced number of draws to decision, we did not find a difference in confidence at the moment of deciding. It must be noted that confidence at the moment of deciding could be unaffected under this account. Both low-delusion-prone and high-delusion-prone participants might decide when they are 80% confident that the lake they choose is the correct lake; the latter group might simply reach this level of confidence earlier due to increased confidence in each piece of information leading up to the required level. Due to rapidly-decreasing sample sizes, we could not compare confidence levels for each fish in the sequence, but only for the first fish, for which no differences in confidence levels were found. Our results cannot shed light on potentially increased confidence for any accumulating information, which needs to be integrated with previous evidence, for high-delusion-prone participants. Also note that, in the static task, high-delusion-prone participants requested less information than would be optimal, perhaps because evidence in itself, even in an abstract, hypothetical form, is already assigned extra weight. Overall, our results fit the information-integration account, although the results with regards to confidence levels could be explored in more detail in future research. Another account suggests that the JTC bias is a consequence of difficulties with processing sequential information (Young & Bentall, 1995). This would lead to a reduced number of draws to decision. Our results from the dynamic task are in line with this account. However, on the static task, where information was presented instantaneously, rather than sequentially, an effect of delusion-proneness was found for the number of fish requested to see, which would not be expected on the basis of this account.

Fine et al. (2007) also outline motivational accounts. One account posits that delusional participants have a high need for closure, which leads them to decide early, and leads to more certainty. Our findings are not in line with this account, as in the static task, closure could have been obtained equally fast by using more evidence, but yet the JTC bias was still found. Our findings also negate an alternate explanation of the JTC bias: fewer draws might be due to delusion-prone and delusional participants being less motivated to persevere in a seemingly worthless task and in more of a "rush" to finish the experiment (L. O. White & Mansell, 2009). In other words, sampling more information could be perceived as more costly for delusional and delusion-prone participants. Moutoussis et al. (2011) did not find support for this idea when comparing decision models, including a costed-Bayesian model, for data on a non-incentivised beads task. In the dynamic and static tasks described here, various motivational confounds were minimised or virtually eliminated (e.g., providing monetary incentives in both tasks, avoiding additional time and effort associated with gathering more evidence in the static task, minimising cognitive load by not requiring decision updating in the static task) and the standard relative finding, where delusion-proneness was positively associated with JTC, was replicated. This undermines the rushing account.

Another account states that the JTC bias is part of a confirmatory reasoning style, where there is a reduced number of draws due to a limited motivation to search for disconfirming evidence (Dudley & Over, 2003). This confirmatory reasoning style is commonly found for threat-related material, even by healthy controls, when it is important to find supporting, rather than falsifying, evidence for a claim (e.g., "if there is smoke, then there is fire"; Dudley & Over, 2003). Our results might speak to this account. If a smaller sample is considered, perhaps the chance of encountering disconfirming evidence is minimised (i.e., “if I do not look, it is not there”; much like people afraid of heights would close their eyes when on a cliff, for example). A desire to avoid disconfirming evidence might then lead delusion-prone participants to consider fewer fish. This account does not hypothesise an effect on confidence levels.

Overall, the findings of the present study are in line with the conclusion by Fine et al. (2007), who state that only the information-integration account and the confirmatory reasoning style are supported by the findings of their review.

### 2.4.1 Potential Limitations

The beads task was incentivised to generate optimal decision points, but the incentives also minimised the confound where a low motivation might lead delusion-prone participants to decide quickly in order to finish the experiment earlier. If this confound is operating, however, they may also try to finish the PDI more quickly and answer “no” for endorsement of questions, to avoid the three follow-up questions of conviction, preoccupation, and distress. There is evidence that individuals filling out the PDI are more likely to endorse earlier items than later items, especially when the scale is included in a larger battery of tests (R. Ross, personal communication, Aug. 31, 2013). If delusion-prone participants are especially inclined to want to finish tasks and questionnaires more quickly, the PDI might systematically underestimate their delusion-proneness (undermining the validity of the PDI). If this is the case, delusion-prone participants should decide too quickly on the beads task *and* have low scores on the PDI. In other words, a positive association between the decision point and the PDI-scores would be expected. The association found here, however, was negative, so that decisions on the beads tasks were earlier as PDI-scores were higher. Therefore, the motivation to finish tasks and questionnaires quickly does not seem to impact our results, which followed the hypothesised direction. Furthermore, higher endorsement scores, sub-scale scores, and total scores on the 21-PDI all distinguish patients from a healthy sample (Peters et al., 2004). If individuals high in delusional thinking would be less likely to endorse items, such criterion validity would not be found. This again speaks against this potential limitation.

A second potential limitation concerns the static task, in which participants were presented with ten listed options. Although the mean rational number of fish to request was 5.4, participants may have requested to see five fish because they considered the centre of the scale of options an easy, “neutral” point, much like the neutral response alternative on Likert scales (Nowlis, Kahn, & Dhar, 2002). Indeed, across all five trials in the static task, the most frequently requested number of fish was five fish. However, this bias on the centre of the scale was not different for low and high delusion-prone participants (2 (delusion-proneness) × 10 (number of fish requested) Fisher's exact tests for each trial, all *p*s>.172), so our finding that high-delusion-prone participants requested fewer data than low-delusion-prone participants cannot be due to anchoring. Nevertheless, in future studies it might be beneficial to ask participants to enter the number of fish they request, rather than presenting a list of ten options with a clear centre option.

## 2.5 Conclusion

Studies in experimental psychology have often claimed that delusional and delusion-prone participants "jump to conclusions" on probabilistic reasoning tasks. However, this term suggests that premature decisions are made, but such a notion is only meaningful when there is in fact an optimal point at which a rational individual should decide. No previous studies have included both rewards for correct decisions and costs for gathering information, making the claim that delusional and delusion-prone participants "jump to conclusions" unwarranted. In this study, stringent experimental economics methods were adopted to minimise effects of potential confounds (e.g., miscomprehension, limited motivation, impaired memory) and optimal decision points were created through rewards and costs. We found that, in a dynamic task, all participants tended to decide before the optimal decision points, but in the static task only high-delusion-prone participants did this. Furthermore, we replicated the relative JTC bias (i.e., high-delusion-prone participants saw fewer fish before deciding than low-delusion-prone participants) in both tasks, also when accounting for risk-aversion and intelligence. In conclusion, our findings support the claim that delusional ideation is associated with a tendency to "jump to conclusions", in both a relative and an absolute sense.

# 3 Delusion-Proneness and Probability Estimates[10]

## 3.1 Background

In Chapter 2, evidence for the jumping-to-conclusions (JTC) bias was found in two data gathering versions of the beads task. Participants saw two lakes filled with fish of different colours, in opposite ratios across the lakes and a fisherman who was fishing from one of the two lakes. Participants could choose how many fish they wanted to see from the fisherman's catch before deciding which of the two lakes he was fishing from. We employed both a dynamic draws-to-decision version of the task, in which fish were presented sequentially, and a static one-shot version of the task, in which participants decided how many fish they wanted to see before they saw any fish. In contrast to these data-gathering variants of the beads task, in this chapter we used an adaptation of the probability-estimates version of the beads task to investigate the beliefs that underpin data-gathering decisions. In standard probability-estimates versions, participants are shown a sequence of beads and after each bead they provide the probabilities that the sequence of beads is coming from either of two jars (Garety & Freeman, 1999). In such versions, the JTC bias manifests in higher probabilities being provided by deluded or delusion-prone participants than by healthy controls (relative JTC bias) or than Bayes' theorem (see below and Chapter 1) would warrant (absolute JTC bias).

Our aims in the present study were three-fold: 1) To investigate JTC in a variant of the probability-estimates paradigm designed to minimise common confounding factors; 2) to investigate systematically the effect of incentives in a

[10] Part of this chapter has been published in Van der Leer and McKay (2014).

probability-estimates paradigm; and 3) to investigate different aspects of probability reasoning.

As discussed in Chapters 1 and 2, performance on the beads task is vulnerable to several potential confounds. The strain on working memory and a lack of motivation were the most relevant confounds addressed in the study reported in this chapter. Working memory deficits might impair the ability to maintain and manipulate information, which, in turn, may influence the JTC bias (Freeman et al., 2014; Garety et al., 2013). In terms of facilitating the maintenance of information, several studies have included a (visual) memory aid, but this generally did not abolish the bias (e.g., Dudley et al., 1997b; Moritz & Woodward, 2005). The study reported here aimed to facilitate the manipulation of information by presenting only one piece of information in each trial, rather than a sequence. This also prevented the common misunderstanding that the sources of the information (e.g., the lakes or jars) switch throughout the sequence (Balzan, Delfabbro, & Galletly, 2012; Balzan, Delfabbro, Galletly, et al., 2012).

Deviations from optimal performance might be due to a lack of incentives to perform optimally. Only two previous studies have included incentives in the beads task (Lincoln, Ziegler, Mehl, et al., 2010; Woodward et al., 2009), but both adopted the draws-to-decision version. Thus, no previous study has incentivised a probability-estimates version and the present study was the first to do so, with a systematic investigation of the effects of incentives as a second aim of the present study. We compared performance on incentivised and non-incentivised versions of the probability-estimates version of the beads task. Furthermore, a risk-aversion measure was included, because participants might differ on this variable, independently from a JTC bias, as Lincoln, Ziegler, Mehl, et al. (2010) noted, though neither they nor Woodward et al. (2009) included such a measure.

Furthermore, the third aim of the present study was to investigate different aspects of probability reasoning. As indicated in Chapter 1, rational beliefs and decisions are often compared to normative standards, such as a value calculated using Bayes' theorem (Dienes, 2008):

$$P(H|D) = \frac{P(D|H) \times P(H)}{P(D)} = \frac{P(D|H) \times P(H)}{P(D|H) \times P(H) + P(D|\neg H) \times P(\neg H)}$$

Bayes' theorem has previously been used to determine optimal, rational performance on the probability-estimates version of the beads task where participants provide subjective probabilities for how likely a sequence of beads is to come from either of two jars (Huq et al., 1988; Speechley et al., 2010). Applied to the beads task, the priors ($P$(H) and $P$(¬H)) would pertain to how likely each jar is to be selected, which generally should be 50% for each jar as one is selected at random (Garety & Freeman, 1999). The likelihoods ($P$(D|H) and $P$(D|¬H)) are determined by the ratios of the beads in the jars and the colour(s) of the bead(s) shown. The posterior probability ($P$(H|D)) integrates these factors and constitutes what participants are required to report. Whenever the priors of two alternatives hypotheses are equal, as is (at least implicitly) the case in the beads task, the posterior probability is determined by the likelihoods (Dienes, 2008).

As Matthews (2005) noted, differences in subjective posterior beliefs between groups, such as delusional patients and healthy controls, could be due to different priors and to different likelihood ratios ($P$(D|H)/$P$(D|¬H)) being assigned. If one holds an extreme prior belief, such as believing with full conviction that the experimenter will pick the predominantly white jar to draw beads from (i.e., $P$(White jar)=1 and P(Black jar)=0), no number of black beads would be able to change this belief. This resonates with the idea Hemsley and Garety (1986) put forth: the strength with which delusions are held, and their resistance to modification by experience, could be due to very high prior probabilities. No amount of experience might then change the belief.

Alternatively, McKay (2012) postulated that delusional inferences might arise from a bias towards “explanatory adequacy”. On this account, delusional patients would place too much emphasis on the likelihood ratio and would not consider prior beliefs sufficiently. When people have a bias towards explanatory adequacy, they might favour the hypothesis that best *explains* the observed data, without properly considering the prior probability of that hypothesis (Coltheart, Menzies, & Sutton, 2010). For example, consider a woman who experiences pressure on her skull and hears a constant buzzing sound (Maher, 1988, as cited in Langdon & Bayne, 2010). These data could be explained by the hypothesis that she has bees in her head. In fact, the likelihood of her experiences if there are bees in her head would be much higher than if there were no bees in her head. Ignoring the extremely low prior probability of the “bees in head” hypothesis, the woman arrives at the conclusion that she has bees in her head. In reality, this woman experienced pressure and tinnitus from a tumour which softened her skull.

If the “explanatory adequacy” theory holds, an interesting investigation in the beads task would be to vary the prior probabilities of the jars being selected. If delusional or delusion-prone participants indeed have a bias towards explanatory adequacy, they might rely too much on likelihoods (i.e., the colour of the bead and the jar in which that colour is predominant), while not adequately taking into account the prior probability of the jar being chosen. To the best of our knowledge, this variation of the task has not previously been tested and is also explored in the study presented in this chapter.

### 3.1.1 The Present Study

The present study was based on the lakes-and-fish adaptation of the classic beads task (Whitman & Woodward, 2011). Whereas usually there is a single fisherman, in our paradigm there were five fishermen, living at different distances from the two lakes (see Figure 3.1 on page 115). This enabled variation of the prior distribution over the two lakes.

Participants completed 26 trials. On each trial one of the five fishermen displayed a fish he had caught from one of the two lakes. The fisherman's house was depicted in the background, and the location of each fisherman's house in relation to the two lakes determined the probability that he would fish from either lake, thus providing information about the prior distribution over the two lakes. For example, the fisherman second from the left would visit Lake A four out of six times ($P$(Lake A)=.67), as the fisherman was described as fishing six days a week and resting on Sunday. The fish displayed by the fisherman could be black or white. On each trial the two lakes contained black and white fish in one of three complementary ratios: 50:50/50:50, 60:40/40:60, or 85:15/15:85. This provided information about the likelihood of a fish of a certain colour being caught from each of the two lakes. Aside from the colour of the fish displayed by the fisherman, there were thus two key types of information that could vary on each trial: 1) Which of the five fishermen was depicted fishing, and 2) the ratios of black and white fish in the two lakes.

On each trial participants indicated which lake they thought the fish had been caught from; this was done either by betting money on the lake they thought it was from or by providing probability estimates. A single fish was displayed on each trial (see Figure 3.1 on page 115 for an example trial) to minimise working-memory demands and to avoid miscomprehension regarding the swapping of lakes (Balzan, Delfabbro, & Galletly, 2012; Balzan, Delfabbro, Galletly, et al., 2012). As such, each trial in the current paradigm is equivalent to the first fish or bead in the standard probability-estimates paradigm.

The 26 trials encompassed three conditions, based on whether the prior distribution over the lakes was non-uniform (and thus informative) or uniform, and whether the ratio of black to white fish in the lakes was unequal (and thus informative) or equal:

- *Prior* condition (10 trials): non-uniform prior distribution over the lakes; equal ratio of black to white fish in each lake (i.e., 50:50/50:50).
- *Standard (likelihood)* condition (8 trials): uniform prior distribution over the lakes; unequal ratio of black to white fish in each lake (i.e., 60:40/40:60 or 85:15/15:85). This is the standard scenario in beads-task studies.
- *Combined* condition (8 trials): non-uniform prior distribution over the lakes; unequal ratio of black to white fish in each lake. In this condition, both the colour of the caught fish and the location of the fisherman were informative.

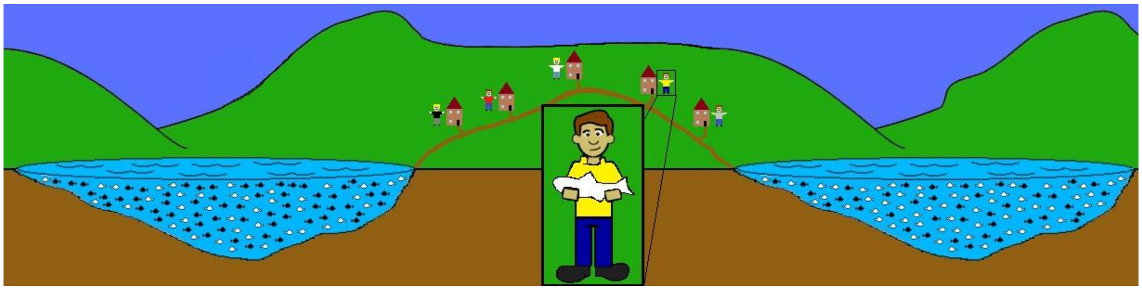


**Figure 3.1** Visual representation of a trial in the *combined* condition. In this case, the fisherman lives closer to Lake B (on the right) and thus visits this lake more often (4 out of 6 times). The displayed fish is white, which is the predominant colour in Lake B (60 white: 40 black), also favouring the hypothesis that the fisherman was fishing from Lake B.

### 3.1.2 Hypotheses

Three main effects and an interaction were predicted.

- First, a main effect of condition was expected. We hypothesised that participants would find it easier to take into account just one type of information (i.e., either priors or likelihoods) than to integrate both types, leading to larger deviations from Bayesian posteriors in the *combined* condition than in the *prior* or *standard* conditions.
- Second, a main effect of group (betting versus control) was expected. We hypothesised that participants in the betting group would be more motivated to make accurate decisions than those in the control group, leading to smaller deviations from Bayesian posteriors in the betting group than in the control group.
- Third, a main effect of delusion-proneness was expected. High-delusion-prone participants were expected to show a JTC-bias and thus to bet more money on or over-estimate the lake with the highest Bayesian posterior, compared to low-delusion-prone participants.
- Finally, an interaction between conditions and delusion-proneness was expected. We predicted that high-delusion-prone participants would use

prior information less when other information was available, compared to low-delusion-prone participants. Thus we predicted that the difference between conditions would be greater for high-delusion-prone participants than for low-delusion-prone participants.

## 3.2 Methods

### 3.2.1 Participants

Participants (n=129, 83 females, 45 males, 1 unknown gender; mean (SD) age = 20.36 (3.1) years) were students from RHUL, recruited using the Online Recruitment System for Economic Experiments (Greiner, 2004). Participants in the betting group received £5 for participation and between £0 and £4 as a bonus, depending on their bets (mean (SD) = £2.50 (£0.87)). Participants in the control group received £8 for participation. The Psychology Department Ethics Committee of RHUL approved this study.

### 3.2.2 Materials

#### *3.2.2.1 Fisherman-Adapted Beads Task*

Before the start of the experiment, participants were provided with written instructions providing information about the ratios of black to white fish in the two lakes (i.e., likelihoods), the fishermen (i.e., priors), and the means of responding (see Appendix C). The response bar ran from "Lake A" on the left to "Lake B" on the right. Participants could indicate how much they wanted to bet on either lake or how likely they thought either lake was by moving a slider within this response bar. Indicator stripes for 25% Lake A, 50%/50%, and 75% Lake A were provided, but the numerical value of the response was not shown. Participants had to answer comprehension questions correctly before starting the experiment (see Appendix C). Answers were checked and participants were referred back to the written instructions when an incorrect answer was given.

In three within-participants conditions differences in priors and likelihoods were represented by which of five fishermen was fishing and by different ratios of black to white fish in the lakes, respectively. Twenty-six trials were created to incorporate the possible combinations of priors and likelihoods per condition, while keeping the number of trials in the different conditions similar; the colour of the caught fish and in which lake that colour was the predominant colour was determined pseudo-randomly with the constraints that each trial could only be presented once. Visual stimuli were based on those used by Speechley et al. (2010). Participants were informed that the distance of a fisherman's house to a lake was directly proportional to the number of times he would go fishing in that lake. Figure 3.1 (on page 115) shows the fourth fisherman from the left, living closer to Lake B, and the colour of the caught fish matches the predominant colour of that lake, so both pieces of information (i.e., prior and likelihood) favour the hypothesis that the fisherman was fishing from Lake B (Bayesian posterior probability that the white fish is from Lake B and not Lake A = .75[11]).

#### 3.2.2.2 *Risk Aversion*

For a description of the computerised risk-aversion measure (Holt & Laury, 2002), see Chapter 2 (2.2.2.2, on page 88).

#### 3.2.2.3 *Delusion Proneness*

For a description of the 21-item Peters et al. Delusions Inventory (PDI; Peters et al., 2004), see Chapter 2 (2.2.2.3, on page 89).

[11] $\frac{prior(Lake\ B) \times likelihood(Lake\ B)}{(prior(Lake\ B) \times likelihood(Lake\ B)) + (prior(Lake\ A) \times likelihood(Lake\ A))} = \frac{{}^{4}/_{6} \times .6}{({}^{4}/_{6} \times .6) + ({}^{2}/_{6} \times .4)} = .75$

### 3.2.3 Procedure

The experiment lasted approximately one hour. Participants were tested in groups ranging from 16-26 people. All sessions were conducted on a local computer network using z-Tree software (Fischbacher, 2007) in the EconLab at RHUL. Before the experiment began, participants provided written informed consent, read instructions, and completed comprehension questions. Participants signed up for a session without any knowledge of the different groups. In the betting group, participants were given £4 to distribute over the two lakes as they wished. They were informed that one randomly chosen trial would be paid out according to their distributions of the money. In the control group, participants simply provided an estimate of how probable each lake was; these decisions were without financial consequence. The response bar was the same for both groups. To the extent that the slider was moved toward "Lake A", Lake A was considered more likely (in the control group) or more money was bet on Lake A (in the betting group); positioning the slider exactly between Lake A and B indicated that the participant did not think either of the lakes was more likely than the other or that they equally split their bets across the two lakes. The 26 trials were presented in random order for each participant. So that participants would not have time to calculate the posterior (e.g., using their phone), each trial had a time limit of 20 seconds. Then, participants completed the risk-aversion measure, the PDI, and some demographic questions (gender, age, subject studied). At the end of the experiment, one of the trials was randomly selected for the betting group. For that trial, the correct lake was drawn based on the Bayesian posterior probability.[12] Participants were paid out the money they bet on the correct lake on that trial as a bonus.

### 3.2.4 Statistical Analyses

The dependent variable on each trial was the deviation of each participant's subjective probability for each lake from the Bayesian posterior probability. To calculate the average deviations, first, for all trials where the Bayesian posterior was below .5 (i.e., Lake A, on the left, was more likely than Lake B, on the right), the Bayesian posteriors and participants' decisions were subtracted from 1, so that they were coded in the same direction as the other trials (i.e., those where Lake B was more likely than Lake A). Then, for each trial, the Bayesian posterior was subtracted from the participants' decision. Deviations were directional, so that negative values indicated underestimation, while positive values indicated overestimation. Next, the average deviations per condition were calculated (i.e., across 10 trials in the prior condition; across 8 trials in the standard condition; across 8 trials in the combined condition). In the betting group, the amounts participants bet on each lake were converted into proportions representing probability estimates. The control group already provided such probability estimates. In principle, the responses in both groups were comparable, since the sliding bar was the same in both groups.

We ran two sets of complementary analyses. In the first type of analysis we used a median split to convert the continuous PDI scores into a categorical between-participants factor (e.g., LaRocco & Warman, 2009; Warman et al., 2007). We

[12] For the randomly chosen trial, Lake A had a posterior probability of .25 and Lake B of .75. A random number between 1 and 100 was generated to determine which lake the fish was from. The numbers 1 to 25 represented the posterior probability of Lake A, while the numbers 26 to 100 represented Lake B. Therefore, the lake with the highest Bayesian posterior was not always the correct lake.

then ran mixed factorial analyses of variance (ANOVA), with condition (*prior*, *standard* or *combined*) as a within-participants factor, and group (betting versus control) and delusion-proneness (low versus high) as between-participants factors. Since the betting group took risks betting their money, ANCOVAs controlling for risk-aversion were also conducted. Greenhouse-Geisser-corrected degrees of freedom are reported when the assumption of sphericity was violated.

The second type of analysis consisted of regression analyses with PDI-scores analysed as a continuous measure to investigate direct links between delusion-proneness and deviations from Bayesian reasoning (McKay et al., 2006). This was done by regressing PDI-scores, group, and the interaction between PDI-scores and group on the deviations from Bayesian posteriors. This was done per condition. Again, we checked whether results remained the same when risk-aversion was controlled for in the betting group.

## 3.3 Results

### 3.3.1 Data Screening

First, the data was inspected for outliers in the planned regression and factorial analyses. For the factorial analyses, outliers were identified through boxplots; for the regression analyses, outliers were determined based on standardised residuals, Cook’s distance, and Mahalanobis distances. Four participants were outliers across the two types of analyses. These were excluded from the analyses, reducing the sample size to 125. Next, outliers on regression analyses were identified based on standardised residual values outside the range from -2 to 2. This inspection indicated that an additional 16 participants were outside this acceptable range; as this is more than the acceptable 5% of the total number of participants (Field, 2005), the 16 participants were excluded from all analyses to enable comparison across analyses, reducing the sample size to 109.

Further statistical assumptions were checked for n=109. For the factorial analyses, the normality of the dependent variable was checked and met (Kolmogorov-Smirnov tests' *p*s>.069). For the regression analyses, absence of multicollinearity was confirmed by the fact that there were no significant, strong correlations between predictors, the tolerance values were >.978, and the VIF values were all <1.022. The predictors were linearly related with the outcome, as inclusion of the squared predictors did not lead to a significantly better model. The residuals were normally distributed. Homoscedasticity was confirmed as the plots of standardised residuals and predicted scores showed that the variance was equal across the range of the predicted scores (i.e., randomly scattered). As all assumptions were met with n=109, analyses were conducted using this sample.

### 3.3.2 Descriptive Statistics

Descriptive statistics for age, PDI, and risk-aversion are provided in Table 3.1 and for the subject studied and gender in Table 3.2. As participants in the betting group essentially performed a different task than participants in the control group, analyses were conducted separately for each group. In order to compare across these two groups, the participants should come from the same underlying population. None of the continuous variables were significantly different in the betting group compared to the control group (age: *t*(107)=1.057, *p*=.293, *d*=0.204, 95%-CI [-0.557, 1.830]; PDI: *t*(107)=0.473, *p*=.637, *d*=0.091, 95%-CI [-10.995, 17.882]; risk-aversion: *t*(107)=0.725, *p*=.725, *d*=0.140, 95%-CI [-0.789, 1.130]). The number of males and females did not differ significantly between the betting and control groups, $\chi^2(2)$=1.491, *p*=.475, $\phi_C$=.117. The breakdown of subjects studied was different across the groups, $\chi^2(3)=10.437$, $p=.015$, $\phi_C=.309$. Separate chi-square tests for the categories Economics, Management, Psychology, and Other (e.g., Zoology, History) were conducted, with a multiple-comparisons-corrected $\alpha$-level of .0125 (=.05/4). None of the tests for the specific major subject groups were significant at the corrected level (Economics: $\chi^2(1)=4.000$, $p=.046$, $\phi_C=.500$; Management: $\chi^2(1)=0.800$, $p=.371$, $\phi_C=.200$; Psychology: $\chi^2(1)=0.059$, $p=.808$, $\phi_C=.059$; Other: $\chi^2(1)=5.786$, $p=.016$, $\phi_C=.321$).

**Table 3.1** Descriptive statistics for age, PDI, and risk-aversion. Values are provided for the total sample and split by betting and control groups.

| Variable | Sample | N | Median | Mean | Standard deviation | Range |
|---|---|---|---|---|---|---|
| **Age** | Total | 109 | 20 | 20.32 | 3.141 | 18-42 |
| | Betting | 52 | 20 | 20.65 | 4.191 | 18-42 |
| | Control | 57 | 20 | 20.02 | 1.685 | 18-25 |
| **PDI** | Total | 109 | 67 | 69.15 | 37.844 | 7-170 |
| | Betting | 52 | 66 | 67.35 | 36.114 | 10-152 |
| | Control | 57 | 67 | 70.79 | 39.604 | 7-170 |
| **Risk-aversion** | Total | 109 | 7 | 7.26 | 2.514 | 1-10 |
| | Betting | 52 | 7 | 7.35 | 2.634 | 1-10 |
| | Control | 57 | 7 | 7.18 | 2.421 | 1-10 |

**Table 3.2** Descriptive statistics for gender and subject studied. Values (n (percentage)) are provided for the total sample and split by betting and control groups.

| Variable | Category | Total sample | Betting | Control |
|---|---|---|---|---|
| **Gender**[A] | Female | 71 (61.5%) | 32 (61.5%) | 39 (68.4%) |
| | Male | 37 (33.9%) | 19 (36.5%) | 18 (31.6%) |
| **Subject studied** | Economics | 16 (14.7%) | 12 (23.1%) | 4 (7.0%) |
| | Management | 20 (18.3%) | 12 (23.1%) | 8 (14.0%) |
| | Other | 56 (51.4%) | 19 (36.5%) | 37 (64.9%) |
| | Psychology | 17 (15.6%) | 9 (17.3%) | 8 (14.0%) |

[A] One participant in the betting group did not indicate their gender.

In the factorial analyses PDI-scores were converted into a categorical between-subjects factor using a median split. Although the median in the current sample (i.e., 67) is lower than that generally reported in the literature (Balzan, Delfabbro, & Galletly, 2012; LaRocco & Warman, 2009; Warman et al., 2007; 75.5, 90, and 97.5, respectively), it is higher than the median score of 49 originally reported by Peters et al. (2004). Participants with a PDI score at the median were excluded from the analysis (LaRocco & Warman, 2009), reducing the sample size to n=107.

### 3.3.3 Factorial Analyses

**Error! Reference source not found.** (on page 125) shows the results of a 2 (group: betting vs. control; between-subjects) × 2 (delusion-proneness: high versus low based on median split; between-subjects) × 3 (condition: prior vs. standard vs. combined; within-subjects) mixed ANOVA.

Condition had a significant main effect ($F(1.848,190.302)=19.532$, $p<.001$, $\eta_p^2=.159$). Planned comparisons, with Bonferroni corrections, indicated that higher estimates were provided in the prior condition (mean (SE) = 1.590 (.515)) than in the standard (-2.974 (.779); $p<.001$, 95%-CI [2.450, 6.679]) and combined (-2.022 (.659); $p<.001$, 95%-CI [1.953, 5.272]) conditions. The estimates in the two latter conditions did not differ ($p=.620$, 95%-CI [-0.871, 2.775]).

The two between-participants factors, delusion-proneness and group, did not have significant main effects ($F(1,103)=.510$, $p=.477$, $\eta_p^2=.005$, 95%-CI [-1.235, 2.626], and $F(1,103)=0.560$, $p=.456$, $\eta_p^2=.005$, 95%-CI [-1.202, 2.660], respectively). Low-delusion-prone (-.788 (.685)) and high-delusion-prone (-1.483 (.692)) participants did not differ in their estimates. Participants in the control (-1.500 (.672)) and the betting (-.771 (.704)) groups did not differ in their estimates.

The interaction between group and condition was significant ($F(1.848,190.302)=3.149$, $p=.049$, $\eta_p^2=.030$). In the control group, estimates in the prior condition (2.209 (.710)) were higher than in the standard (-4.288 (1.075); $p<.001$, 95%-CI [3.578, 9.417]) and than in the combined (-2.420 (.909); $p<.001$, 95%-CI [2.338, 6.920]) conditions, while the latter two did not differ ($p=.221$, 95%-CI [-0.649, 4.385]). In the betting group, estimates for the prior condition (.972 (.745)) were higher than those in the combined condition (-1.625 (.953); *p*=.029, 95%-CI [0.195, 4.998]), but not different to those in the standard condition (-1.660 (1.127); *p*=.116, 95%-CI [-0.429, 5.691]), nor did the estimates in the latter two conditions differ (*p*=1.00, 95%-CI [-2.673, 2.603].

The interaction between delusion-proneness and condition was not significant (*F*(1.848,190.302)=.816, *p*=.435, $\eta_p^2$=.008).

The interaction between group and delusion-proneness was marginally significant (*F*(1,103)=3.652, *p*=.059, $\eta_p^2$=.034). Low-delusion-prone participants in the control group (-2.082 (.950)) showed a trend to provide lower estimates than low-delusion-prone participants in the betting group (.507 (.986); *p*=.062, 95%-CI [-0.128, 5.306]). High-delusion-prone participants in the control group (-.917 (.950)) gave estimates similar to those provided by high-delusion-prone participants in the betting group (-2.049 (1.006); *p*=.415, 95%-CI [-1.613, 3.876]).

The three-way interaction was significant (*F*(1.848,190.302)=4.620, *p*=.013, $\eta_p^2$=.043). In the control group, low-delusion-prone and high-delusion-prone participants did not differ in the prior (2.071 (1.005) vs. 2.347 (1.005), respectively; *p*=.846, 95%-CI [-2.542, 3.094]), standard (-5.664 (1.520) vs. -2.912 (1.520), respectively; *p*=.203, 95%-CI [-1.511, 7.016]), or the combined conditions (-2.653 (1.286) vs. -2.187 (1.286), respectively; *p*=.798, 95%-CI [-3.140, 4.072]). In the betting group, low-delusion-prone and high-delusion-prone participants did not differ in the prior (.943 (1.043) vs. 1.001 (1.063), respectively; *p*=.969, 95%-CI [-2.895, 3.011]) or in the combined condition (-.906 (1.334) vs. -2.343 (1.361), respectively; *p*=.453, 95%-CI [-2.343, 5.216]), but in the standard condition, the estimates from low-delusion-prone participants were higher than those from high-delusion-prone participants (1.485 (1.578) vs. -4.804 (1.609), respectively; *p*=.006, 95%-CI [1.820, 10.758]).

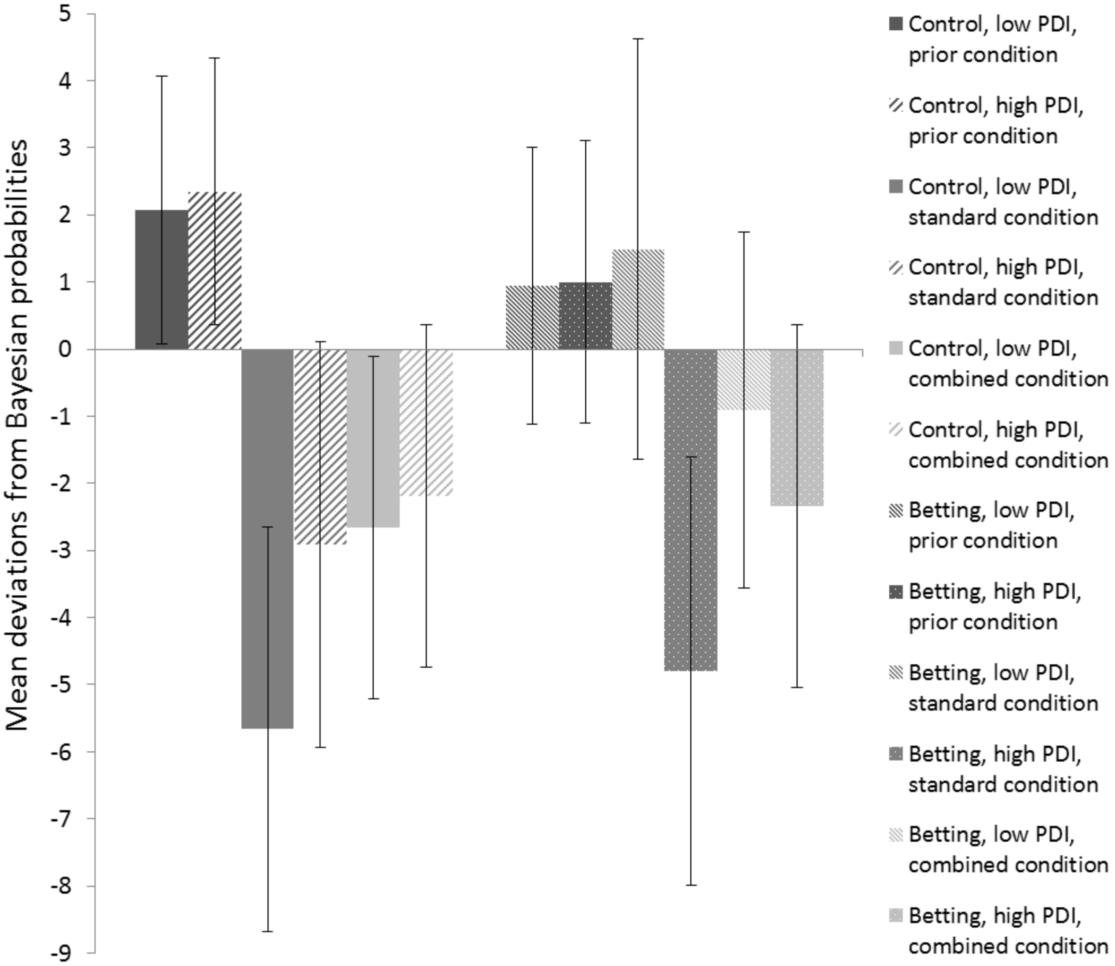


**Figure 3.2** Means (± 95%-CIs) of the average deviations from Bayesian posteriors across the three conditions for the control and betting groups and split by delusion-proneness (low and high PDI; based on a median split).

When risk-aversion was accounted for, a 2 (group) × 2 (delusion-proneness) × 3 (condition) mixed ANCOVA indicated that condition did not had a significant main effect anymore (*F*(1.861,189.773)=.232, *p*=.777, $\eta_p^2$=.002. The two between-participants factors, delusion-proneness and group, did not have significant main effects (*F*(1,102)=.584, *p*=.447, $\eta_p^2$=.006, 95%-CI [-1.187, 2.675], and *F*(1,102)=0.629, *p*=.430, $\eta_p^2$=.006, 95%-CI [-1.159, 2.703], respectively).

The same ANCOVA, accounting for risk-aversion, showed that the interaction between group and condition was significant (*F*(1.861,189.773)=3.387, *p*=.039, $\eta_p^2$=.032). In the control group, estimates in the prior condition (2.218 (.713)) were higher than in the standard condition (-4.341 (1.066); *p*<.001, 95%-CI [3.670, 9.450]) and than in the combined condition (-2.440 (.911); *p*<.001, 95%-CI [2.368, 6.950]), while the latter two did not differ (*p*=.206, 95%-CI [-0.615, 4.417]). In the betting group, estimates for the prior condition (.962 (.748)) were higher than those in the combined condition (-1.604 (.955); *p*=.032, 95%-CI [0.165, 4.967]), but not than those in the standard condition (-1.605 (1.117); *p*=.125, 95%-CI [-0.462, 5.596]), and the estimates in the latter two did not differ either (*p*=1.00, 95%-CI [-2.638, 2.636].

As with the ANOVA reported above, the interaction between delusion-proneness and condition was not significant when accounting for risk-aversion (*F*(1.861,189.773)=.944, *p*=.385, $\eta_p^2$=.009).

After accounting for risk-aversion, the two-way interaction between group and delusion-proneness was significant (*F*(1,102)=4.040, *p*=.047, $\eta_p^2$=.038). Low-delusion-prone participants in the control group (-2.131 (.951)) provided lower estimates than low-delusion-prone participants in the betting group (.605 (.990); *p*=.049, 95%-CI [-5.465, -0.009]). High-delusion-prone participants in the control group (-.911 (.950)) gave estimates similar to those provided by high-delusion-prone participants in the betting group (-2.103 (1.006); *p*=.391, 95%-CI [-1.552, 3.937]).

The three-way interaction remained significant when accounting for risk-aversion (*F*(1.861,189.773)=5.211, *p*=.008, $\eta_p^2$=.049). In the control group, low-delusion-prone and high-delusion-prone participants did not differ in the prior

(2.092 (1.010) vs. 2.345 (1.009); *p*=.860, 95%-CI [-2.579, 3.084]), the standard (-5.787 (1.508) vs. -2.896 (1.506); *p*=.178, 95%-CI [-1.338, 7.119]), or the combined (-2.700 (1.290) vs. -2.181 (1.288); *p*=.777, 95%-CI [3.098, 4.136]) conditions. In the betting group, low-delusion-prone and high-delusion-prone participants did not differ in the prior (.900 (1.051) vs. 1.024 (1.069); *p*=.934, 95%-CI [-2.856, 3.104]) or in the combined condition (-.813 (1.343) vs. -2.394 (1.365); *p*=.412, 95%-CI [-2.225, 5.388]), but in the standard condition, the estimates from low-delusion-prone participants were higher than those from high-delusion-prone participants (1.485 (1.578) vs. -4.804 (1.609), respectively; *p*=.004, 95%-CI [2.218, 11.119]).

In general, accounting for risk-aversion did not change the results, except that the main effect of condition disappeared and that the interaction between group and delusion-proneness became significant, rather than marginally significant. How accounting for risk-aversion led to the absence of an effect of conditions was further investigated by including a between-subjects factor for risk-aversion, consisting of a median split, in the model. This 3 (condition) × 2 (group) × 2 (delusion-proneness) × 2 (risk-aversion: high versus low) mixed ANOVA indicated that risk-aversion interacted significantly with condition ($F$(1.859,158.050)=4.151, *p*=.020, $\eta_p^2$=.047). Highly risk-averse participants showed an effect of condition: deviations in the prior condition (1.944 (.782)) were higher than in the standard (-5.083 (1.139); *p*<.001, 95%-CI [3.844, 10.212]) and combined (-2.481 (1.000); *p*<.001, 95%-CI [1.883, 6.967]) conditions, and the deviations in the latter two conditions did not differ (*p*=.066, 95%-CI [-0.122, 5.327]). For participants low in risk-aversion there was no effect of condition as deviations in the prior condition (1.094 (.862)) did not differ from those in the standard (-.964 (1.256); *p*=.469, 95%-CI [-1.456, 5.570]) or combined (-.761 (1.103); *p*=.330, 95%-CI [-0.950, 4.659]) conditions, and the deviations in the latter two conditions also did not differ (*p*=1.00, 95%-CI [-3.209, 2.803]).

In order to investigate whether probability estimates provided by low-delusion-prone and high-delusion-prone in both the betting and control groups, and in each condition, deviated significantly from the Bayesian posterior probabilities, one-sample *t*-tests were conducted. Table 3.3 reports these results. A Bonferroni correction for multiple comparisons was not applied, as the conservativeness of this test (i.e., .05/12=.004) could mask interesting patterns of results. As such the results should be interpreted with caution, as there is an increased probability of Type I errors. Overall, participants in the control groups tended to overestimate probabilities in the prior condition, while tending to underestimate them in the standard and combined conditions. In general, participants in the betting group were quite accurate across conditions.

**Table 3.3** Results of one-sample *t*-tests, which indicate whether probability estimates were different from the Bayesian posterior probability. This was done per group and per condition. Uncorrected *p*-values are reported.

| Group | Condition | Delusion-proneness | One-sample *t*-test result | 95%-CI |
|---|---|---|---|---|
| **Control** | Prior | Low | *t*(27)=2.795, *p*=.009, *d*=1.076 | [0.551, 3.591] |
| | | High | *t*(27)=2.420, *p*=.023, *d*=0.931 | [0.357, 4.337] |
| | Standard | Low | *t*(27)=-4.493, *p*<.001, *d*=1.729 | [-8.251, -3.077] |
| | | High | *t*(27)=-1.977, *p*=.058, *d*=0.761 | [-5.934, 0.111] |
| | Combined | Low | *t*(27)=-2.476, *p*=.020, *d*=0.953 | [-4.851, -0.454] |
| | | High | *t*(27)=-1.823, *p*=.079, *d*=0.702 | [-4.648, 0.274] |
| **Betting** | Prior | Low | *t*(25)=.843, *p*=.407, *d*=0.337 | [-1.361, 3.246] |
| | | High | *t*(24)=.786, *p*=.439, *d*=0.321 | [-1.626, 3.627] |
| | Standard | Low | *t*(25)=.999, *p*=.327, *d*=0.400 | [-1.577, 4.547] |
| | | High | *t*(24)=-2.409, *p*=.024, *d*=0.983 | [-8.920, -0.689] |
| | Combined | Low | *t*(25)=-.661, *p*=.515, *d*=0.264 | [-3.732, 1.919] |
| | | High | *t*(24)=-1.440, *p*=.163, *d*=0.588 | [-5.702, 1.016] |

### 3.3.4 Regression Analyses

Linear regression analyses showed that the predictors (PDI-scores, group, and the interaction between PDI-scores and group) did not significantly predict deviations in the prior condition ($F$(3,105)=.475, $p$=.701, $R^2_{\text{ADJUSTED}}$=-.015) or in the combined condition ($F$(3,105)=.178, $p$=.911, $R^2_{\text{ADJUSTED}}$=-.023). In the standard condition, however, group was a significant predictor of deviations, as was the interaction between PDI-scores and group (model statistics: $F$(3,105)=3.363, $p$=.021, $R^2_{\text{ADJUSTED}}$=.062; predictor statistics: see Table 3.4, models 1).

A set of hierarchical linear regressions was conducted to account for risk-aversion. The first step, with risk-aversion as the sole predictor, did not lead to a significant model for deviations in the prior ($F$(1,107)=.010, $p$=.920, $R^2_{\text{ADJUSTED}}$= -.009), standard ($F$(1,107)=1.403, $p$=.239, $R^2_{\text{ADJUSTED}}$=.004), or combined ($F$(1,107)=.364, $p$=.548, $R^2_{\text{ADJUSTED}}$=-.006) conditions (see Table 3.4, models 2.1). Although adding PDI-scores, group, and the interaction between PDI-scores and group as additional predictors did not significantly improve the model for the prior condition ($\Delta F$(3,104)=.470, $\Delta p$=.704; $F$(4,104)=.355, $p$=.840, $R^2_{\text{ADJUSTED}}$= -.024) or for the combined condition ($\Delta F$(3,104)=.201, $\Delta p$=.895; $F$(4,104)=.240, $p$=.915, $R^2_{\text{ADJUSTED}}$=-.029), it did for the standard condition ($\Delta F$(3,104)=3.706, $\Delta p$=.014; $F$(4,104)=3.157, $p$=.017, $R^2_{\text{ADJUSTED}}$=.074). After accounting for risk-aversion, in the standard condition, group and the interaction between group and PDI-scores were significant predictors of deviations (see Table 3.4, models 2.2). Overall, the deviations in the betting group increased in value comparing the control group (coded as 0) to the betting group (coded as 1), which, counter-intuitively, means an increase in accuracy, considering the underestimation in the control group. However, comparing participants with high PDI-scores in the control and betting groups, those in the betting group underestimated the probabilities more than those in the control group.

**Table 3.4** B-values, confidence intervals for the b-values, SEs for the b-values, ß-values and significance values for the predictors in the regression analyses, conducted per condition. PDI = delusion-proneness. Group coding: 0 = control, 1 = betting.

| Condition | Model | Predictor | b | 95%-CI of b | SE | β | *p* |
|---|---|---|---|---|---|---|---|
| **Prior** | 1 | PDI | .008 | [-0.028, 0.043] | .018 | .054 | .676 |
| | | Betting group | -.344 | [-4.597, 3.909] | 2.145 | -.033 | .873 |
| | | PDI*Betting group | -.011 | [-0.066, 0.043] | .027 | -.089 | .682 |
| | 2.1 | Risk-aversion | -.021 | [-0.424, 0.383] | .203 | -.010 | .920 |
| | 2.2 | Risk-aversion | -.021 | [-0.430, 0.388] | .206 | -.010 | .919 |
| | | PDI | .008 | [-0.028, 0.044] | .018 | .055 | .674 |
| | | Betting group | -.321 | [-4.619, 3.977] | 2.167 | -.030 | .883 |
| | | PDI*Betting group | -.012 | [-0.066, 0.043] | .028 | -.091 | .678 |
| **Standard** | 1 | PDI | .008 | [-0.046, 0.062] | .027 | .037 | .765 |
| | | Betting group | 8.858 | [2.408, 15.309] | 3.253 | .533 | .008 |
| | | PDI*Betting group | -.089 | [-0.172, -0.007] | .042 | -.448 | .034 |
| | 2.1 | Risk-aversion | -.378 | [-1.009, 0.254] | .319 | -.114 | .239 |
| | 2.2 | Risk-aversion | -.479 | [-1.092, 0.134] | .309 | -.144 | .124 |
| | | PDI | .010 | [-0.043, 0.064] | .027 | .047 | .702 |
| | | Betting group | 9.392 | [2.947, 15.838] | 3.250 | .565 | .005 |
| | | PDI*Betting group | -.096 | [-0.178, -0.013] | .041 | -.481 | .023 |
| **Combined** | 1 | PDI | .001 | [-0.044, 0.047] | .023 | .008 | .949 |
| | | Betting group | 1.512 | [-3.881, 6.906] | 2.720 | .114 | .579 |
| | | PDI*Betting group | -.010 | [-0.079, 0.059] | .035 | -.061 | .779 |
| | 2.1 | Risk-aversion | -.155 | [-0.663, 0.354] | .256 | -.058 | .548 |
| | 2.2 | Risk-aversion | -.171 | [-0.688, 0.347] | .261 | -.064 | .514 |
| | | PDI | .002 | [-0.043, 0.048] | .023 | .013 | .922 |
| | | Betting group | 1.702 | [-3.737, 7.142] | 2.743 | .128 | .536 |
| | | PDI*Betting group | -.012 | [-0.082, 0.057] | .035 | -.076 | .729 |

## 3.4 Discussion

Participants' deviations from Bayesian posteriors were investigated over three within-participants conditions, in which either the prior distribution, the likelihood distribution, or both distributions were non-uniform. This was investigated in a control group, who provided probability estimates for the lakes, and a betting group, who had to bet their endowment on the two lakes in their desired proportion. Analyses focused on averaged deviations from Bayesian posterior probabilities per condition, i.e., the extent to which participants underestimated or overestimated the probability that the presented fish was from the lake with the highest actual Bayesian posterior.

Although an effect of condition was found, it was not in the hypothesised direction. Participants tended to overestimate the posteriors in the prior condition, but underestimated the posteriors in the standard condition and in the combined condition. This effect was moderated by group. The control group gave significantly higher estimates in the prior condition than in the other two. For the betting group, the estimates in the standard condition were between those in the prior and the combined conditions, which were significantly different from each other. Furthermore, risk-aversion influenced the effect of condition, so that an effect of condition was only found for participants who were more risk-averse. Overall, the similarity in response between the standard and combined condition suggests that participants tend to use likelihood information more than prior information when both types of information are available. The hypothesised main effects of delusion-proneness and of group were not found. These two factors did interact, however, as low-delusion-prone participants gave lower estimates in the control group than in the betting group. Yet, high-delusion-prone participants responded similarly in both groups.

The hypothesised interaction between condition and delusion-proneness was not significant, but the analyses indicated a three-way interaction. There was a difference between low-delusion-prone and high-delusion-prone participants in the standard condition in the betting group, which was not found in the control group, and not for either group in the prior or combined conditions. The effect remained when controlling for risk-aversion. In the betting group, low-delusion-prone participants were accurate, while high-delusion-prone participants underestimated the posteriors, when only likelihood information was to be used. This effect was supported by the fact that the interaction between continuous PDI-scores and group significantly predicted average deviations in the standard condition. None of the predictors (PDI-scores, group, or their interaction) significantly predicted deviations in the prior or the combined conditions.

Therefore, the JTC effect was found in the condition that most resembles the standard beads task paradigm. For the control group, the relative JTC bias was found as high-delusion-prone participants provided higher probability estimates than low-delusion-prone participants, although this difference was not statistically significant. In the betting group, low-delusion-prone participants were accurate, while high-delusion-prone participants underestimated.

In the standard beads task, the prior probability of either jar is always 50%/50% and likelihoods are non-uniform, and delusion-prone participants are found to decide on a lake on the basis of less evidence than non-delusion-prone participants (i.e., the JTC-bias). In the present study we investigated whether a non-uniform distribution of *prior* probabilities would also elicit a difference in responses between low-delusion-prone and high-delusion-prone participants.

This expected effect was not found. However, participants did seem to be more accurate (i.e., deviated less from the Bayesian posterior) in the condition with uniform likelihoods but non-uniform prior probabilities, compared to the two conditions with non-uniform likelihoods. A potential explanation is that differences in prior probabilities were visually more salient than the ratios of black to white fish in the lakes.

Interestingly, differences between low-delusion-prone and high-delusion-prone participants were only found in the condition most commensurate with the literature on JTC-effects on the beads task. As most previous studies have not incentivised the task, the control group in our study is similar to the standard probability-estimates version of the beads task. Our finding that low-delusion-prone participants in this group provided numerically lower estimates than high-delusion-prone participants is consistent with the commonly reported finding that delusion-prone participants "jump to conclusions" *relative to controls*.

Yet, this only describes relative "jumping". Although caution with interpretation is warranted due to an increased chance of Type I errors, results from one-sample *t*-tests suggested that absolute "jumping" did not occur, as high-delusion-prone participants were below the Bayesian posterior in both groups. Phillips and Edwards (1966) have found conservatism in a probabilistic reasoning task in a healthy population. They found that conservatism was not affected by different prior distributions, but it was found to be stronger with likelihoods ratios further from 50:50. This might also speak to why the deviations in the prior condition in the present study were different from those involving likelihood information. With likelihood ratios further from 50:50, as in our conditions with 60:40 or 85:15 ratios, more conservatism could be expected on the basis of the findings by Phillips and Edwards (1966), and, indeed, deviations indicated underestimation of the Bayesian posteriors.

The difference between the control and the betting group suggests that incentivising the task can shed light on aspects of the JTC-bias. Our results imply that high-delusion-prone participants performed at a ceiling level, where rewards did not improve performance, as Woodward et al. (2009) suggested for clinical patients. Low-delusion-prone participants were sensitive to the reward and provided estimates that were not significantly different from Bayesian posteriors in the betting group, while they were different in the control group. Phillips and Edwards (1966) found that, for healthy participants, the inclusion of incentives reduced conservatism. This is consistent with our finding that low-delusion-prone participants in the betting group provided estimates not significantly different from the Bayesian posteriors, while low-delusion-prone participants in the control group underestimated them.

### 3.4.1 Potential Limitations

In the present study, participants did not see the numerical value of their decision when they moved the slider of the response bar. This was done to prevent participants from feeling that they should calculate the exact posterior (e.g., by using their phone). However, we may have unwittingly introduced measurement error here. For example, a participant might correctly infer the posterior probability of the two lakes yet misestimate the point on the response bar that corresponds to this correct probability. Future studies could incorporate a few control trials, where participants are asked to place the slider at specified values. The deviation in these trials could then represent measurement error.

Demand characteristics were a potential confound. This study was conducted in a behavioural economics laboratory, where participants may have felt they were supposed to adopt certain economic strategies. One particular confound could be that participants in the betting group may have tried to maximise their expected value and therefore bet all their money on the most likely lake, rather than splitting their reward according to their subjective probabilities of each lake. This would decrease accuracy, in particular leading to overestimation of the lakes, which is also expected when participants jump to conclusions. However, results from the betting group showed that Bayesian posteriors were either accurately estimated or underestimated; there was no evidence of overestimation.

The potential use of an expected-value strategy in the betting group means participants' beliefs about the posteriors of each lake may not have been straightforwardly revealed by their decisions. One potential way to get at these beliefs on any given lakes-and-fish trial would have been to present participants with a table of lotteries (cf. risk-aversion measure by Holt & Laury, 2002), and for each lottery in the table to ask participants to choose whether they would prefer to play that lottery or to play the lottery represented by the current lakes and fish trial (the latter lottery would involve receiving £4 if a given lake was the lake being fished from, and £0 otherwise). One would then select one of these choices at random, and the participant would play the lottery chosen in that case. The benefit of this procedure is that decisions would transparently reveal beliefs (i.e., that reported probabilities equal subjective probabilities; Holt & Smith, 2009), assuming that participants understood the procedure. The major disadvantage of this procedure is the significant risk of confusion and miscomprehension: this elaborate task could confuse many of our participants and so create more problems than it is worth. This concern is especially pertinent given recent evidence that miscomprehension confounds results even in the *standard* beads task (Balzan, Delfabbro, & Galletly, 2012). On balance, we decided to adopt an imperfect, but comprehensible, strategy rather than a perfect strategy that might not be understood.

## 3.5 Conclusion

The study reported in this chapter adopted a probability-estimates version of the beads task. Rigorous methods were developed to minimise the influence of potential confounds concerning working-memory deficits, miscomprehension, and a lack of motivation. From estimates provided for a single fish, rather than a sequence, it became clear that incentives can affect probability reasoning differently in low-delusion-prone and high-delusion-prone participants. However, this was limited to a standard condition with uniform prior probabilities and varying likelihoods. Within this condition, in the control group, the *relative* JTC bias was replicated, in that high-delusion-prone participants provided higher estimates than low-delusion-prone participants. Yet, in the incentivised betting group, this pattern was reversed. Furthermore, no evidence was found for an *absolute* JTC bias, as neither low-delusion-prone nor high-delusion-prone participants overestimated the Bayesian posterior probabilities, in either group, in this condition.

Together with the findings in Chapter 2, these results only partially support the liberal acceptance account, which stipulates unaffected probability reasoning, but a lowered decision-threshold (Moritz et al., 2007). High-delusion-prone participants in the study in Chapter 2 behaved less conservatively than the probability estimates found in this chapter would suggest they should. Of course, the studies in the two chapters used different samples, so a direct comparison of each participant's behaviour and probability reasoning is not available. Nevertheless, taken together, these results would support the liberal acceptance account. Furthermore, we found that the probability reasoning of low-delusion-prone and high-delusion-prone participants was similar in two out of the three within-subject conditions in both control and betting groups. Against the liberal acceptance account, however, we did find a marked difference in probability reasoning for low-delusion-prone and high-delusion-prone participants in the standard condition. Furthermore, the results concerning confidence levels in Chapter 2 were not supportive of the liberal acceptance account either. Future investigations with clinical populations may shed further light on whether results from our incentivised tasks would support the liberal acceptance account.

# 4 Sexual Over-Perception Bias

## 4.1 Background

As described in Chapter 1, the sexual over-perception bias refers to the phenomenon where men perceive more sexual interest from a woman than the woman herself reports feeling and more than female observers perceive (Abbey, 1982; Haselton & Buss, 2000; Lindgren et al., 2008). Error management theory (EMT) has been suggested as an explanation of this bias (Haselton & Buss, 2000). The claim is that for men in the ancestral past, it would have been less costly to mistakenly infer sexual interest and be disappointed (i.e., a false alarm) than it would have been to mistakenly infer the absence of sexual interest and miss a potential reproductive opportunity (i.e., a miss). Evolutionary pressures would thus have selected for this bias, which would persist in modern-day men.

Although several researchers claim their results support the predictions made by EMT with regards to the sexual over-perception bias (Haselton & Buss, 2000; Lindgren et al., 2008), there are some limitations both to the theoretical underpinnings of this bias and to the evidence adduced in support of it. The theory predicts that men supposedly believe women are more interested than women really are. However, it is possible that the relevant bias does not involve biased beliefs, but rather only biased behaviour. McKay and Dennett (2009) have noted that many examples arguably explained by EMT involve behaviour minimising costly errors, but that these behaviours could be created without invoking a bias in beliefs. For example, one may not strongly *believe* there is oncoming traffic, but may check for it regardless when crossing the street, just to be prudent. Similarly, the error costs of misreading sexual interest could be minimised through biased behaviour, without invoking biased beliefs. A man might not strongly *believe* a woman to be sexually interested, but still approach her, thinking his chances are low but that it is still better to at least give it a try (Haselton, Nettle, & Murray, in press). Under uncertainty, nature may have included a policy to behave in a way that minimises costs, without changing beliefs (McKay & Dennett, 2009). In other words, the sexual over-perception bias might be an outcome bias, rather than a cognitive bias (Marshall et al., 2013). Therefore, evidence to support the notion that this bias is also found at a cognitive level is crucial.

Abbey (1982) showed that male observers of a dyadic interaction between a man and a woman showed the bias, while female observers did not. This suggests that the bias is cognitive in nature, as the observers had access to the exact same information. Yet, men and women might have different prior beliefs about general levels of sexual interest, so that a bias in posterior beliefs obtains even when observing the same evidence (McKay & Efferson, 2010). Such a difference between men and women might develop due to different socialisation experiences. Girls are taught to show sexual restraint (Low, 1989), while boys are taught, or encouraged through stereotypes in the media, to show great interest in sex (Haselton & Buss, 2000). As this interest develops, it might become generalised, so that men assume everyone, including women, to have great sexual interest (Abbey, 1982). As such, men and women might have different prior beliefs about others' sexual interest. The paradigms of previous studies (Abbey, 1982; Abbey & Harnish, 1995; Haselton, 2003; Haselton & Buss, 2000) have included only one piece of evidence. Participants' beliefs did not require updating based on several pieces of accumulating evidence. Hence, differences in posterior probabilities provided by men and women might be ascribed to differences in prior beliefs. In other domains of psychology, such as the optimism bias (see Chapter 5), updating paradigms (Sharot et al., 2011) have improved upon "one-point-estimate" paradigms (with beliefs measured on the basis of one piece of evidence) by elucidating the process through which people arrive at biased estimates for future outcomes. In a similar vein, in this study we aimed to investigate the sexual over-perception bias by using an updating paradigm.

### 4.1.1 The Present Study and Hypotheses

In this study, the beads task used to study the jumping-to-conclusions bias (Chapters 2 and 3) was adapted to form a belief-updating paradigm to measure the sexual over-perception bias. We investigated whether men and women differ in the extent to which they incorporate relevant feedback regarding women's sexual interest, and specifically whether men systematically overestimate women's sexual interest. This question was approached from two angles in two related experiments. The first (3a) investigated whether men systematically overestimate the probability that women are sexually interested in men (i.e., are heterosexual). The second (3b) investigated whether men systematically overestimate the probability that women are sexually interested in a given man. Besides the condition hypothesised to lead to biased cognitions, both experiments also involved a neutral condition. The neutral condition was a probability-estimates version of the beads task. Here, black or white beads were shown in succession, and after each bead participants had to give probability estimates for a jar containing mostly black beads and for a jar containing mostly white beads. At the end of the sequence, they had to decide which jar all the beads were from. No gender differences were expected on this task.

For the conditions where a bias was hypothesised (i.e., bias conditions), men were expected to systematically under-weigh evidence that would count against women being sexually interested in men (in general or in a specific man). The bias condition in experiment 3a involved a male character who had gone to either a gay bar or a straight bar ("bars" replacing "jars") and had flirted with women who may or may not have responded positively to his advances ("positive/negative responses" replacing "beads"). After each response, participants had to provide their estimates for either type of bar, based on whether female characters responded to his advances or not. It was predicted that men, compared to women, would overestimate the probability that the man was in the straight bar, as this provided a context where women would be sexually interested, while sexual interest would be low in a gay bar. The bias condition in experiment 3b involved a male character who was either attractive or unattractive to women (replacing jars) and had speed-dated several women who may or may not have wanted to go on a further date with him (replacing beads). Participants had to indicate their estimates for whether the male character was attractive to women or not, based on whether female characters wanted to go on a further date after their speed date. It was predicted that men would provide higher estimates than women for the male character being attractive to women.

## 4.2 Methods

### 4.2.1 Participants

There were 77 participants in experiment 3a and 73 participants in experiment 3b.[13] The majority of participants were students at RHUL. Participants received a show-up fee of £3 and received a bonus between £0 and £2 (mean (SD) reward = £1.83 (0.38)) based on their answers in the experiment. The Psychology Department Ethics Committee of RHUL approved this study.

### 4.2.2 Materials and Procedure

This study comprised two interrelated experiments. In each experiment, there were two within-subject conditions: a neutral condition (beads and jars) and the relevant bias condition (which varied across the two experiments). The two experiments were conducted simultaneously. Participants were not aware which experiment they signed up for when enrolling for the study. Of the eight sessions conducted in total, four investigated general sexual interest (3a: bars), the other four investigated sexual interest in a given man (3b: dates). The order in which within-subject conditions were presented was counterbalanced across sessions for each experiment.

In all sessions participants first read written instructions and had to answer comprehension questions correctly before proceeding to the tasks (see Appendix D).

In the neutral condition (3a and 3b: jars), participants were shown computerised jars filled with beads of two colours in opposite ratios (70% black: 30% white

[13] Due to time constraints imposed by testing in groups, the data of eight additional participants in experiment 3a and four additional participants in 3b was incomplete. These participants were excluded from all analyses.

and vice versa). They were informed that there would be four rounds where series of beads would be drawn from one of the two jars. They had to determine from which jar the beads were coming in each round. For the first nine beads, participants had to indicate the probabilities that the beads were coming from Jar A or Jar B. After ten beads, participants had to decide from which jar all the beads came.

Participants also completed essentially the same task in a bias condition, either before or after the neutral condition, depending on the order of conditions in their session. This bias condition was framed so as to evoke the biased cognition hypothesised by EMT. In the bias condition in experiment 3a, participants were told that a heterosexual man had either gone to a gay bar or to a straight bar (i.e., a non-gay bar). Here, he had flirted with women. Women at a gay bar may not have been interested in flirting with him and so his success rate here was low (30% of flirtations reciprocated, 70% of flirtations ignored). At a "straight" bar, his success rate was higher (70% of flirtations reciprocated, 30% of flirtations ignored). The two bars replaced the two jars and the beads were replaced by successes (i.e., reciprocated flirtations, visualised by green happy faces) and failures (i.e., ignored flirtations, visualised by red sad faces). Participants had to use these pieces of information to determine if the man went to the gay bar or the straight bar. The other three rounds repeated this same scenario with a different order of information and differently named men.

In the bias condition of experiment 3b, participants were told that a man had gone speed-dating and they would see how many women had been willing to go on a subsequent date with him. If women found him attractive, he would have been more successful (on average 70% success and 30% failure); while he would have been less successful if he were not attractive (on average 30%

success and 70% failure). Based on pieces of information that indicated success (i.e., she would go on a subsequent date, visualised by a green check) or a failure (i.e., she would not go on a subsequent date, visualised by a red cross), participants had to decide if the man from the story was attractive or not. Here, the beads from the neutral condition were replaced by information indicating success or failure. The jars were replaced by the question of whether the man was attractive or not. The other three rounds repeated this same scenario with a different order of information with differently named men.

In each of the conditions, participants could win a reward of £1 for a correct decision in one of these rounds. As in Chapter 2, the information in one of the rounds (the first round) was drawn at random, but with probabilities matching the ratio of different types of information (i.e., 70:30) in the randomly selected state of the world. Since this first round, used for pay-out, was randomly determined per session, this round was excluded from analyses. In the other three rounds, the order of information was fixed across participants to facilitate analyses. Table 4.1 shows the sequences of information in these three rounds and the state of the world suggested by the sequence as a whole. The rounds were presented in a fixed order across conditions, to prevent the same rounds appearing sequentially in two different conditions. The sequences were generated with a random number generator: three sets of 10 non-unique numbers ranging from 1 to 100 were generated. All numbers above 30 would represent the suggested source (i.e., that with which the sequence as a whole was most consistent), while the numbers 1 to 30 would represent the opposite source.

**Table 4.1** The sequences of information per condition and per round. In the jars task, w refers to a white bead and b refers to a black bead. In the bars condition, yes refers to a reciprocated flirtation and no refers to an ignored flirtation. In the dates condition, yes refers to a woman wanting to go on a next date after the speed-date and no refers to a woman not wanting to go on a next date. The correctness of a sequence is based on which state of the world was randomly selected, which was the same state suggested by the majority of the information within the sequence.

| Round | Condition | Correct | #1 | #2 | #3 | #4 | #5 | #6 | #7 | #8 | #9 | #10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | All | n/a | Randomly determined per session | | | | | | | | | |
| **2** | Jars | White | w | w | b | w | w | w | w | b | w | b |
| | Bars | Gay bar | no | no | yes | no | no | no | no | yes | no | yes |
| | Dates | Attractive | yes | yes | no | yes | yes | yes | yes | no | yes | no |
| **3** | Jars | White | w | b | b | w | w | w | w | w | w | w |
| | Bars | Gay bar | no | yes | yes | no | no | no | no | no | no | no |
| | Dates | Attractive | yes | no | no | yes | yes | yes | yes | yes | yes | yes |
| **4** | Jars | Black | b | b | w | b | b | b | w | w | b | w |
| | Bars | Straight bar | yes | yes | no | yes | yes | yes | no | no | yes | no |
| | Dates | Unattractive | no | no | yes | no | no | no | yes | yes | no | yes |

After the two conditions, participants answered demographic questions (gender, age, sexuality) and were paid for their participation. Sexuality was assessed through the Kinsey sexuality scale (Kinsey, Pomeroy, & Martin, 1948/1998), which measures sexuality on a continuum.

All sessions were conducted on a local computer network using z-Tree software (Fischbacher, 2007) in the EconLab at RHUL. The experiment lasted approximately 30 minutes. Before the experiment began, participants provided written informed consent.

### 4.2.3 Data Preparation, Data Screening, and Statistical Analyses

The analyses used scores for the state of the world hypothesised to be systematically overestimated by men: the straight bar and being attractive in experiments 3a and 3b, respectively; the jar presented on the same side as the straight bar or attractive label was used for the neutral condition. Note that this may not always have been the state suggested by the evidence across the sequence of information.

Initially, we had intended to investigate probability estimates in a 9 (draw; within-subjects) × 3 (round; within-subjects) × 2 (condition; within-subjects) × 2 (gender; between-subjects) mixed analysis of variance (ANOVA). However, screening of the data, after it had been collected, indicated varying types of extreme deviations from normality per variable at the level of each individual draw (e.g., bimodal distributions, positive skew, negative skew). No straightforward data transformations would aid in normalising all variables, as, for example, a transformation would improve the distribution of one variable, but deteriorate those of other variables. Therefore, analyses could not look at effects of single pieces of information (i.e., each draw). Instead, the average deviation from Bayesian estimates was calculated across the nine draws for each non-random round. This deviation was calculated as the participants' provided probability estimate minus the Bayesian probability. Positive values, therefore, indicate overestimation of the relevant state of the world; whereas negative values indicate underestimation.

However, this did not eliminate non-normality. As a result, the average deviation across the three non-random rounds in each condition was calculated. This also accounted for the fact that the rounds were presented in a fixed order, which may have led to order effects. Hence, deviations from the Bayesian probability averaged across nine draws of a sequence (i.e., one round) were calculated first. Then, these average deviations per sequence were averaged across the three rounds in each condition. In experiment 3a, the average Bayesian probabilities that the character was in a straight bar were .12 in round 2, .38 in round 3, and .76 in round 4; averaged across the rounds this amounted to a probability of .42 for the straight-bar state of the world (i.e., slightly in favour of the opposite gay-bar state of the world). The mostly-black jar within experiment 3a had the same probabilities and was thus the jar-equivalent of the straight bar. In experiment 3b, the Bayesian probabilities for the state of the world where the character was attractive in each round were .88, .71, and .15, with an average Bayesian probability across the rounds amounting to .58. In experiment 3b, the mostly-white jar was the jar-equivalent of the attractive state of the world, based on Bayesian probabilities.

Data screening of these average deviations across the nine draws, averaged across the three rounds, still indicated severe non-normality as indicated by Kolmogorov-Smirnov tests (*p*<.001 for the majority of the dependent variables) and visual inspection of data plots (e.g., extreme kurtosis). Removing outliers (based on standardised scores larger than an absolute 3.29, as no scores in a standardised normal distribution should be larger than this; Field, 2013) or data transformations did not correct for non-normality (e.g., *p*=.039 for remaining groups).

Finally, therefore, robust analyses for mixed designs using M-estimators and bootstrapping with 2000 bootstrapped samples were conducted on the full sample (Field, Miles, & Field, 2012; Wilcox, 2005), in a 2 (condition; within-subjects) × 2 (gender; between-subjects) design. In order to investigate whether men systematically overestimated the Bayesian probabilities compared to women, in the bias conditions, but not in the neutral condition, analyses focused on interactions between gender and condition. Furthermore, one-sample Wilcoxon signed-rank tests were conducted to investigate whether men's and women's deviations were significantly different from zero in either the neutral or the bias condition (with an $\alpha$-level of .05/4=.0125 to correct for multiple comparisons).

## 4.3 Results

### 4.3.1 Descriptive Statistics

Table 4.2 shows the descriptive statistics for the order in which the conditions were presented, gender, and sexuality, for experiments 3a and 3b. The samples of the two experiments did not differ in terms of age ($t(148)=.969$, $p=.334$, $d=0.159$, 95%-CI [-0.400, 1.170]; mean (SD) years = 20.03 (2.406) for experiment 3a and 20.41 (2.460) for experiment 3b), gender ($\chi^2(1)=.279$, $p=.625$, $\phi_C=.043$; see Table 4.2), or sexuality (Fisher's exact test: 5.349, $p=.671$; $\chi^2(7)=5.417$, $p=.670$, $\phi_C=.190$; see Table 4.2).

**Table 4.2** Sample characteristics with regards to the order in which conditions were presented, gender, and sexuality.

| | | **Exp. 3a (n=77)** | **Exp. 3b (n=73)** |
|---|---|---|---|
| **Order** | Neutral (jars) —Bias (bars/dates) | 35 | 35 |
| | Bias (bars/dates) – Neutral (jars) | 42 | 38 |
| **Gender** | Male | 36 | 31 |
| | Female | 41 | 42 |
| **Sexuality** | Exclusively heterosexual | 58 | 50 |
| | Predominantly heterosexual, only incidentally homosexual | 9 | 13 |
| | Predominantly heterosexual, but more than incidentally homosexual | 4 | 2 |
| | Bisexual | 2 | 3 |
| | Predominantly homosexual, but more than incidentally heterosexual | 0 | 0 |
| | Predominantly homosexual, only incidentally heterosexual | 1 | 0 |
| | Exclusively homosexual | 1 | 2 |
| | Asexual, non-sexual | 1 | 0 |
| | Don't want to answer | 1 | 3 |

### 4.3.2 Inferential Statistics – Experiment 3a

The robust ANOVA indicated that the main effect of gender was not significant ($\hat{\Psi}$=-1.468, *p*=.053), with men (mean (bootstrapped SE) = 2.443 (.854); bootstrapped 95%-CI$_{\text{MEAN}}$ [0.918, 4.275]) deviating as much as women (2.733 (.540); bootstrapped 95%-CI$_{\text{MEAN}}$ [1.724, 3.813]). The main effect of condition was significant ($\hat{\Psi}$=1.049, *p*=.018), with lower deviations for the bias condition (2.306 (.578); bootstrapped 95%-CI$_{\text{MEAN}}$ [1.249, 3.532]) than for the neutral condition (2.889 (.549); bootstrapped 95%-CI$_{\text{MEAN}}$ [1.846, 4.012]). The interaction was not significant ($\hat{\Psi}$=1.417, *p*=.144). Figure 4.1 shows the means and bootstrapped 95%-CIs.

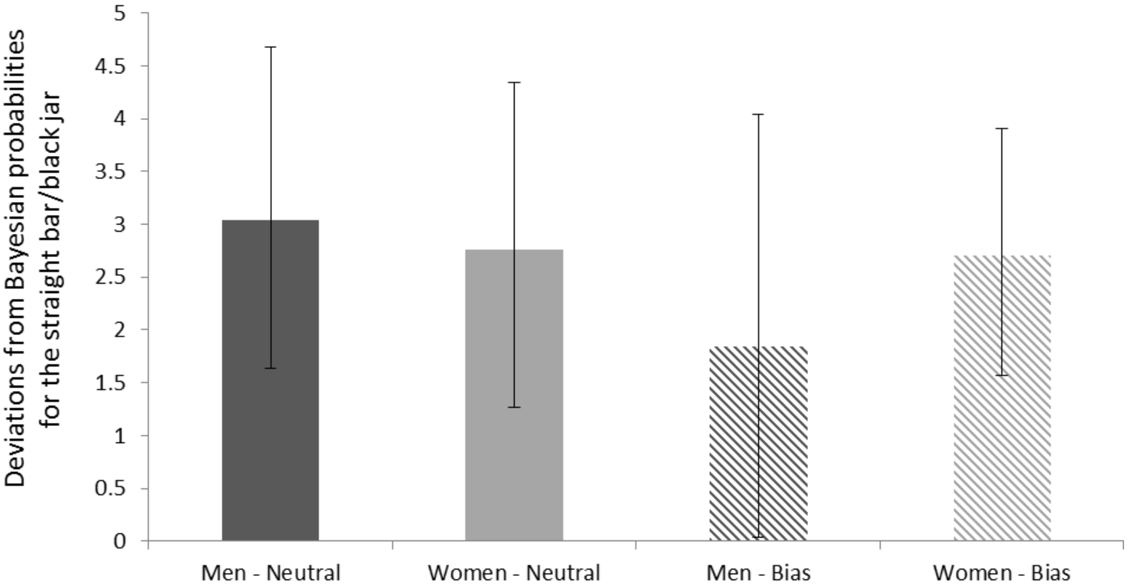


**Figure 4.1** The mean (±bootstrapped 95%-CI) deviations from Bayesian probabilities (i.e., 42%) for the black jar or straight bar in each condition (neutral versus bias) by men and women. Note that at this Bayesian probability, over-estimation indicates conservatism towards the prior probability of 50% (a deviation of 8% from the Bayesian posterior probability).

Women deviated significantly from the Bayesian probability (i.e., 42%) in both the neutral (*T*=684.0, *p*=.001, *r*=.513) and the bias condition (*T*=746.0, *p*<.001, *r*=.638). Men deviated significantly in the neutral condition (*T*=555.0, *p*<.001, *r*=.581), but not in the bias condition (*T*=435.0, *p*=.109, *r*=.267). Note that in experiment 3a, the estimates were compared to a Bayesian probability of 42%. As such, these over-estimations actually reflect conservatism, as the provided estimates are closer to the prior probability of 50% than Bayes' theorem prescribes (Phillips & Edwards, 1966).

### 4.3.3 Inferential Statistics – Experiment 3b

The robust ANOVA indicated that the main effect of gender was not significant ($\hat{\Psi}$=0.198, *p*=.790), with men (-1.935 (.532); bootstrapped 95%-CI$_{\text{MEAN}}$ [-2.986, -0.924]) deviating as much as women (-2.025 (.441); bootstrapped 95%-CI$_{\text{MEAN}}$ [-2.905, -1.187]). The main effect of condition was not significant ($\hat{\Psi}$=-0.730, *p*=.209), where deviations in the neutral condition (-2.501 (.357); bootstrapped 95%-CI$_{\text{MEAN}}$ [-3.221, -1.799]) were not different from deviations in the bias condition (-1.472 (.554); bootstrapped 95%-CI$_{\text{MEAN}}$ [-2.634, -0.455]). The interaction was significant ($\hat{\Psi}$=2.235, *p*=.040). Without the availability of robust post-hoc tests, interpretation of this interaction is based on means and one-sample tests (described below), which suggest that men are equally (in)accurate in both conditions, but women become more accurate in the bias condition (see Figure 4.2).

Women deviated significantly from the Bayesian probability (i.e., 58%) in the neutral condition (*T*=91.0, *p*<.001, *r*=-.696), but not in the bias condition (*T*=330.0, *p*=.129, *r*=-.234). Men deviated significantly in both the neutral (*T*=82.0, *p*=.001, *r*=-.584) and the bias condition (*T*=118.0, *p*=.011, *r*=-.457). Note that in experiment 3b, the estimates were compared to a Bayesian probability of 58%. As such, these under-estimations reflect conservatism, as the provided estimates are closer to the prior probability of 50% than Bayes' theorem prescribes (Phillips & Edwards, 1966).

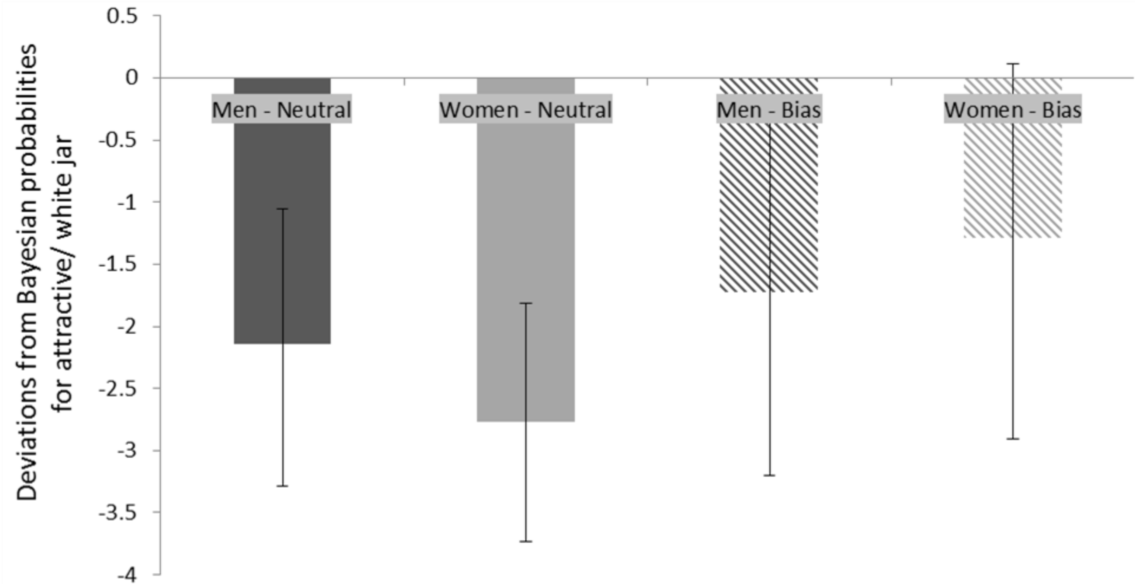


**Figure 4.2** The mean (±bootstrapped 95%-CI) deviations from Bayesian probabilities (i.e., 58%) for the white jar or for the man being attractive in each condition (neutral versus bias) by men and women. Note that at this Bayesian probability, under-estimation indicates conservatism towards the prior probability of 50% (which would be a deviation of -8% from the Bayesian posterior probability).

## 4.3.4 Sensitivity Analyses

### *4.3.4.1 Analyses with Only Heterosexual Participants*

As the evolutionary theory in this study is focused on mating prospects, it could be argued that it would only hold for heterosexual participants, whose reasoning about mating prospects could be more driven by the possibility to produce offspring. Furthermore, the scenarios in the study assumed heterosexual coupling of the characters. Therefore, follow-up analyses (2 (gender) × 2 (condition: within-subjects) were conducted including only participants who indicated they were exclusively heterosexual (n=58 in experiment 3a; n=50 in experiment 3b).

#### *4.3.4.1.1 Analyses with Only Heterosexual Participants – Experiment 3a*

The robust ANOVA indicated that the main effect of gender was not significant ($\widehat{\Psi}$=-1.220, *p*=.111), so men (2.929 (.949); bootstrapped 95%-CI$_{\text{MEAN}}$ [1.344, 5.098]) did not deviate more or less than women (2.393 (.568); bootstrapped 95%-CI$_{\text{MEAN}}$ [1.246, 3.480]). The main effect of condition was not significant ($\widehat{\Psi}$=0.907, *p*=.053), with equal deviations in the neutral condition (2.839 (.570); bootstrapped 95%-CI$_{\text{MEAN}}$ [1.747, 3.999]) as in the bias condition (2.483 (.672); 95%-CI$_{\text{MEAN}}$ [1.248, 3.943]). The interaction was not significant ($\widehat{\Psi}$=1.355, *p*=.183), so that men's deviations in the neutral condition (3.635 (.867); bootstrapped 95%-CI$_{\text{MEAN}}$ [2.030, 5.408]) and in the bias condition (2.222 (1.188); bootstrapped 95%-CI$_{\text{MEAN}}$ [0.321, 4.924]) were not different from women's deviations in the neutral condition (2.0435 (.748); 95%-CI$_{\text{MEAN}}$ [0.614, 3.530]) and in the bias condition (2.743 (.679); bootstrapped 95%-CI$_{\text{MEAN}}$ [1.490, 4.103]).

#### *4.3.4.1.2 Analyses with Only Heterosexual Participants – Experiment 3b*

The robust ANOVA indicated that the main effect of gender was not significant ($\widehat{\Psi}$=0.604, *p*=.484); men (-1.989 (.596); bootstrapped 95%-CI$_{\text{MEAN}}$ [-3.197, -0.899]) and women (-2.228 (.649); bootstrapped 95%-CI$_{\text{MEAN}}$ [-3.559, -1.062]) deviated equally. The main effect of condition was not significant ($\widehat{\Psi}$=-0.405, *p*=.495), with equal deviations in the neutral condition (-2.380 (.451); bootstrapped 95%-CI$_{\text{MEAN}}$ [-3.302, -1.532]) and in the bias condition (-1.837 (.718); bootstrapped 95%-CI$_{\text{MEAN}}$ [-3.326, -0.520]). The interaction was not significant ($\widehat{\Psi}$=1.787, *p*=.157), so that men's deviations in the neutral condition (-2.117 (.684); bootstrapped 95%-CI$_{\text{MEAN}}$ [-3.484, -0.826]) and in the bias condition (-1.860 (.730); bootstrapped 95%-CI$_{\text{MEAN}}$ [-3.426, 0.517]) were not different from women's deviations in the neutral condition (-2.642 (.594); 95%-CI$_{\text{MEAN}}$ [-3.834, -1.504]) and in the bias condition (-1.814 (.1.194); bootstrapped 95%-CI$_{\text{MEAN}}$ [-4.377, 0.320]).

#### *4.3.4.2 Condition as a Between-Subjects Factor*

Feedback from a pilot study suggested that the within-subject condition manipulation was quite evident, which may have obscured a difference between conditions on a within-subject level. This manipulation was kept within subjects for financial considerations. However, it is possible to analyse the difference between the neutral and bias conditions differently. For this alternate approach, analyses only considered the first condition participants encountered, and conditions are compared on a between-subjects level. Robust factorial ANOVAs could not be conducted as the sample sizes in subgroups were not equal and no appropriate non-parametric test was available. Instead, 2 (condition: between-subjects) × 2 (gender) ANOVAs, with 2000 bootstrapped samples, were conducted.

##### *4.3.4.2.1 Condition as a Between-Subjects Factor – Experiment 3a*

There was no significant main effect of gender ($F(1,73)=.190$, $p=.664$, $\eta_p^2=.003$; mean (SE) for men: 2.345 (.782) vs. women: 2.812 (.731); bootstrapped 95%-CI [-1.631, 2.530]). There also was no significant main effect of condition ($F(1,73)=1.374$, $p=.245$, $\eta_p^2=.018$; neutral: 3.206 (.792) vs. bias: 1.951 (.721); bootstrapped 95%-CI [-0.828, 3.366]). Lastly, the interaction was not significant ($F(1,73)=3.220$, $p=.077$, $\eta_p^2=.042$; men (3.933 (1.166)) vs. women (2.479 (1.070)) in the neutral condition, bootstrapped 95%-CI [-2.058, 5.088]; men (.757 (1.043)) vs. women (3.145 (.995)) in the bias condition, bootstrapped 95%-CI [-0.045, 4.861]).

##### *4.3.4.2.2 Condition as a Between-Subjects Factor – Experiment 3b*

There was no significant main effect of gender ($F(1,69)=.176$, $p=.676$, $\eta_p^2=.003$; men: -1.221 (.865) vs. women: -1.697 (.734); bootstrapped 95%-CI [-1.538, 2.553]). There also was no significant main effect of condition ($F(1,69)=1.659$, $p=.202$, $\eta_p^2=.023$; neutral: -2.189 (.831) vs. bias: -.729 (.772); bootstrapped 95%-CI [-0.741, 3.486]). Lastly, the interaction was not significant (*F*(1,69)=.874, *p*=.353, $\eta_p^2$=.013; men (-1.421 (1.318)) vs. women (-2.958 (1.013)) in the neutral condition, bootstrapped 95%-CI [-0.323, 3.423]; men (-1.021 (1.120)) vs. women (-.437 (1.062)) in the bias condition, bootstrapped 95%-CI [-3.182, 4.007]).

## 4.4 Discussion

The experiments described in this chapter sought evidence for the hypothesised sexual over-perception bias using a belief-updating paradigm. One experiment (3a) examined whether men overestimate female sexual interest in men in general, by providing higher-than-warranted ratings for the male character being in a straight bar. The other experiment (3b) examined whether men overestimate how sexually interested women are in a given man, by providing higher-than-warranted ratings for the male character being attractive to women. However, across the experiments no evidence of gender differences was found. Although the interaction between neutral and bias conditions and gender was found to be significant in experiment 3b, this was not a robust effect as it was not found in sensitivity analyses (e.g., analyses involving only heterosexual participants). Moreover, this effect was not in the predicted direction: men gave lower ratings for the male character being attractive than did women. Another non-robust finding was a difference between neutral and bias conditions in experiment 3a, so that the full sample deviated less when estimating the probability that the male character was in a straight bar than when estimating whether beads came from the black jar, but again this result was not found in sensitivity analyses. Overall, we found no evidence of the sexual over-perception bias using this belief-updating paradigm. We did replicate the overall conservatism previously found in this paradigm (Phillips & Edwards, 1966), as the difference between participants' estimates of posterior probabilities and the objective prior probabilities (50% for each state of the world) was generally less than that prescribed by Bayes' theorem.

There are a few aspects of the design that might have contributed to these null findings. In the present design, participants judged the situation of a male character, rather than their own situation or their own interaction. Indeed, ethical constraints precluded asking participants to rate their own attractiveness based on, albeit pre-determined, feedback pertaining to this. Therefore, the study investigated how attractive men thought women would find men in general. It is possible that if males do display a sexual over-perception bias it may be more specific, involving an overestimation of their *own* appeal (and perhaps underestimation of other men's appeal to downplay the competition). However, Abbey and Harnish (1995) found the male sexual over-perception bias for vignettes describing two characters, rather than any task involving the participants themselves, suggesting that this distance factor should not pose a problem.

Another aspect of the design possibly accounting for the null findings was the within-subjects manipulation of conditions. As noted earlier, piloting feedback suggested that the within-subjects manipulation was evident to some participants; but owing to financial considerations we kept this manipulation within subjects. However, a between-subject analysis was available, by comparing the conditions participants encountered first. These analyses did not find any differences between conditions or between genders. Admittedly, this might be because splitting the sample led to underpowered analyses, as Farris et al. (2008b) note that the gender differences should be visible with sample sizes of at least n=45 for each gender, which is slightly higher than the resulting sample sizes when splitting the groups by order in these experiments (i.e., sample sizes ranging from 35 – 42).

Nevertheless, another possible explanation of the null findings relates to the limitations of EMT outlined in the introduction. First, any relevant adaptive bias might be behavioural in nature, rather than cognitive (Marshall et al., 2013; McKay & Dennett, 2009). If so, this bias would not be revealed by our paradigm. One possibility is that the incoming information is integrated with prior probabilities equally by men and women (i.e., there is no cognitive bias in the use of logical inference rules for belief updating), but the value of the incoming information is perceived differently. Another, not mutually-exclusive, possibility (mentioned earlier in 4.1 on page 139) is that men and women have different prior beliefs about women's sexual interest, which may have explained the bias in previous "one-point-estimate" paradigms. Such different prior beliefs could arise from different socialisation of men and women, where men are taught to show great sexual interest and might develop the notion that everyone else also has such great sexual interest (Abbey, 1982; Haselton & Buss, 2000). Within our paradigm, however, the presentation of multiple pieces of information with a given likelihood of occurring would eventually lead everyone to arrive at the same posterior beliefs, despite holding dissimilar prior beliefs (Matthews, 2005). A potential limitation of the present study is that we did not measure subjective prior beliefs, and hence we could not test this suggestion directly.

It is possible that the value of incoming information is perceived differently by men and women, as long as the information is ambiguous. Imagine, for example, that a woman looks away when a man tries to make eye contact. A woman might do this because she is not sexually interested in him. However, it is also possible that she does this in order to play “hard to get” (Jonason & Li, 2013), a behaviour many women are encouraged to display (at least initially) to appear coy (Abbey, 1982). When men assume that women are adopting this strategy, the same behaviour (i.e., looking away) might be considered quite likely in both states of the world (i.e., whether she is sexually interested or not). Such varying interpretations of the likelihood ratio of incoming evidence are only possible for ambiguous cues (e.g., looking away, a smile) and indeed, the sexual over-perception bias is generally only found with ambiguous cues (Buss, 2013; Lindgren et al., 2008). The likelihood ratio of the information presented in our experiments was explicitly stated, so there was no ambiguity regarding the value of the incoming information. This then would avoid arrival at different posterior probabilities due to different subjective likelihood ratios. Indeed, we did not find differences between men and women’s estimates of posterior probabilities.

Furthermore, it has been suggested that the bias might not consist of men *over-perceiving* sexual interest, but actually of women *underreporting* their sexual interest (Perilloux et al., 2012), to avoid being considered promiscuous and thus attempt to protect their reputation (Farris et al., 2008b; Haselton & Buss, 2000). As participants’ self-reports might be biased, our use of vignettes and distant characters in the present design might be a virtue. If reporting one’s own interest is required, accurate reporting could be incentivised, perhaps by increasing the stakes of accurately reporting interest, not through self-report measures on questionnaires, but rather by whether participants’ actual phone numbers would be exchanged, for example, as Perilloux et al. (2012) have suggested.

## 4.5 Conclusion

This study represents the first investigation of how information pertaining to female sexual interest is integrated into posterior probabilities by men in a belief-updating paradigm. In contrast to previous “one-point-estimate” paradigms, no evidence of sexual over-perception was found. This suggests that if men overestimate their sexual prospects, this does not involve irrational belief updating. Furthermore, our study is consistent with the notion that discrepancies between men and women’s reports of sexual interest found in previous studies may not lie in men’s biased perception, but rather in other factors (e.g., women underreporting sexual interest), that we accounted for in the present study. In the next chapter, more general beliefs about future prospects, rather than sexual prospects specifically, are explored.

# 5 Self-Deceptive Optimism

## 5.1 Background

As described in Chapter 1, unrealistic optimism was initially defined as the phenomenon where desirable future outcomes are expected to be more likely, and undesirable future outcomes less likely, than indicated by an objective standard (Segerstrom, 2007; Shepperd et al., 2013). However, following important work by Sharot et al. (2011), this definition has been refined such that "unrealistic optimism" denotes a bias in which beliefs are updated more in response to desirable information than in response to undesirable information. This new definition is based on findings from a paradigm (described in Chapter 1, on page 54, and summarised later in this chapter on page 160) robust to statistical artefacts which may have influenced findings in earlier studies (Shah, 2012). In the present study, unrealistic optimism is investigated to shed light on processes underlying self-deception.

The classic conception of self-deception, or "real" self-deception (Mijovic-Prelec & Prelec, 2010), is analogous to interpersonal deception: one part of the self actively deceives another part (Gur & Sackeim, 1979). The implication is that self-deceived individuals carry two conflicting representations of reality. Proponents of an alternative, "deflationary" account claim that this is paradoxical, and argue that knowledge regarding the use of a deception process should undermine its success (Mele, 1997). Instead, on the deflationary account, cases of putative "self-deception" are thought simply to reflect distortions in the processing of relevant information (Mele, 1997).

To illustrate, consider a standard case of optimistic belief. On the classic conception of self-deception, a heavy smoker who believes her future health prospects are good may also represent a more accurate, and less rosy, state of affairs. In contrast, proponents of the deflationary view might argue that there is no need to suppose that she carries two conflicting representations. She may be processing evidence about the health implications of smoking in a biased fashion (Sharot, 2011) to arrive at one false representation.

The present study combined the optimistic belief-updating paradigm and the crowd-within paradigm as a potential means of testing the "real" self-deception account. To briefly reiterate, in the optimistic belief-updating paradigm, participants provide an initial estimate of their chances of experiencing a negative event, are presented with the base rate of that negative event happening to their demographic, and are then asked to provide a second estimate of their personal chances of experiencing the event (Sharot et al., 2011). Beliefs are updated more when base rates represent desirable information (i.e., the initial estimate was an overestimate) than when they represent undesirable information. Participants in crowd-within experiments provide first and second estimates for *neutral* questions (e.g., "What percentage of the world's airports are in the United States of America?"), without intervening *directional* feedback (Herzog & Hertwig, 2009; Vul & Pashler, 2008). The crowd-within effect refers to the fact that the average of the two estimates has a smaller error than the errors of the individual estimates on average.

To ensure optimism would not constitute a reporting bias (e.g., signalling to oneself or the experimenter that one is healthy), we incentivised accuracy of the answers, so that true beliefs were expected to be reported (Schotter & Trevino, 2014; Simmons & Massey, 2012). However, such incentives introduce a problem for the use of neutral questions in the optimistic belief-updating paradigm.

Participants would provide their first estimate (e.g., the percentage of airports they think are in the United States of America) and then see the correct answer to the question (e.g., 30.3%). With incentives for accuracy, participants' second estimates should not deviate from the correct answer provided. For undesirable events, participants might argue that their own risks are different from the base rate based on individuating information (e.g., no family history of cancer), and the incentivised accuracy might thus not pertain to the presented base rate and they could still deviate while expecting to maximise their payoff. As such, a bias for undesirable events but not neutral events could be due to people aiming to maximise their payoff, rather than due to a cognitive, self-deceptive bias. This confound is avoided in the crowd-within paradigm, where participants are only instructed to assume their first estimate was wrong, but are not informed whether it was too high or too low. Here, systematic, directional biases would suggest self-deception.

The crowd-within effect is thought to occur because, rather than being best guesses, the different estimates are randomly sampled from the same internal distribution of potential estimates, with a mean centred around the true value. All estimates have different random errors, which cancel out when the estimates are averaged (Vul & Pashler, 2008). Moreover, when some estimates are from the lower end of the distribution, and others are from the upper end of the distribution, the mean true value is more likely to be bracketed, which would reduce error even more when averaging (Herzog & Hertwig, 2009).

One possibility is that when asked to supply multiple estimates of their probability of experiencing undesirable outcomes, people sample randomly from an internal probability distribution. If so, the second estimate is just as likely to be more optimistic than the first as it is to be less optimistic than the first, irrespective of the underlying distribution's shape (indeed, this is the basis for the distribution-free Wilcoxon signed-rank test; Howell, 2010).

A second possibility, in line with the “real” self-deception account, imputes more intentionality to the optimist, who samples *selectively* from the optimistic end of an internal distribution.[14] In this case, the two estimates might vary systematically. On the one hand, participants might sample less selectively second time around, providing a less optimistic estimate and producing an enhanced crowd-within effect through reduction in random and systematic error. On the other hand, they might sample even *more* optimistically second time around, perhaps as a kind of defensive manoeuvre (e.g., P. R. Harris & Napper, 2005; Weinstein, 1980). Gal and Rucker (2010) found that individuals induced to experience doubt about their beliefs became stronger advocates of those beliefs than did individuals induced to feel confident in their beliefs, especially when the beliefs were viewed as particularly important. In their experiments, confidence in beliefs was not shaken by presenting evidence that contradicted those beliefs, but via more subtle means (e.g., asking participants to write about their beliefs using their non-dominant hand). In our study, a non-specific prompt for an alternate estimate to one already provided might shake confidence in the initial estimate provided, especially for undesirable questions that might be considered important. This might lead to attempts to bolster one’s position by selecting even more optimistic estimates. In view of Gal and Rucker’s (2010) research, this selective sampling option may be quite likely.

[14] The “real” self-deception account predicts selective sampling from an internal distribution, but is agnostic as to whether that distribution is itself biased (e.g., an outcome of biased information encoding; Sharot, 2011a; Sharot et al., 2011). Selective sampling and biased information processing could work in tandem to produce optimistic estimates.

### 5.1.1 The Present Study and Hypotheses

In the present study, participants provided repeated estimates for neutral and for undesirable questions. With this paradigm, we investigated several questions:

- First, could we replicate the crowd-within effect for neutral questions? We hypothesised that the averaged estimate for neutral questions would have a lower absolute error than the first or second estimate on average.
- Second, would the crowd-within effect obtain for undesirable questions? If so, would this effect be of the same size or larger than that for neutral questions?
- Third, would unrealistic optimism be found in this paradigm? We hypothesised that errors for undesirable questions would indicate underestimation, while no systematic deviation from the true value would be found for neutral questions.
- Fourth, and most importantly, would second estimates for undesirable questions be more optimistic than first estimates for these questions, instead of less optimistic or equivalent?

## 5.2 Methods

### 5.2.1 Participants

The participants were 104 students from RHUL (mean (SD) age = 20.38 (1.90) years; 41 male, 63 female). Participants received a show-up fee of £3 and a decision-based bonus of between £0 and £2 (mean (SD) = £1.83 (£0.38)). The Psychology Department Ethics Committee of RHUL approved this study.

## 5.2.2 Materials

The study included two question-type conditions: neutral and undesirable. The required responses to all questions were percentages. Neutral questions were the eight used in the original crowd-within study (Vul & Pashler, 2008), see Table 5.1. Although Vul and Pashler (2008) provided answers to an accuracy of one decimal place, in the present study participants were asked for integer responses; therefore, participants' estimates were compared to the rounded answers from Vul and Pashler (2008).

### *5.2.2.1 Selection of Undesirable Questions*

Undesirable questions were a selection of the eighty items used by Sharot et al. (2011; obtained through personal communication with C. Korn, 14 February, 2013). We presented the eighty items to thirteen independent raters, who provided estimates of the probability of the events happening to them. The mean estimate was then compared to the "true" values provided by Korn, who derived and calculated these values from PubMed and the Office for National Statistics. Items for which raters provided much lower estimates than the "true" values were considered particularly prone to the optimism bias and, as such, potential candidates for our undesirable questions.

Our final selection of eight undesirable items was made on the basis of several additional considerations. First, raters' comments about the clarity of items were considered and unclear items were excluded (e.g., it was deemed unclear whether "chance of having back pain" referred to chronic or occasional back pain). Second, items that might not be relevant for all participants were removed (e.g., "theft from vehicle" implies the possession of a vehicle). Finally, we selected items involving events that were unlikely or impossible to have happened to the participants already at their current age, so that participants would not give high ratings on the grounds that they were currently experiencing or had previously experienced the event in question. However, as described below, we also explicitly checked whether participants had prior or current experience of the events.

Table 5.1 reports the final questions. The true answers to the undesirable questions (mean (SD) = 35.88 (20.55)) were not significantly different from the answers to the neutral questions (32.50 (23.46)), *t*(14)=.306, *p*=.764, *d*=.164, 95%-CI [-27.021, 20.271]), so this could not explain lower estimates for undesirable questions.

**Table 5.1** The questions used in the task. The eight neutral questions were taken from Vul and Pashler (2008); the eight undesirable questions are a selection from Sharot et al. (2011). Participants' estimates were compared to the (rounded) statistic from the literature.

| Question-type | Question | Literature statistic |
|---|---|---|
| **Neutral** | The area of the United States of America is what percentage of the area of the Pacific Ocean? | 6.3 (6) |
| | What percentage of the world's population lives in China, India, or the European Union? | 44.4 (44) |
| | What percentage of the world's airports are in the United States of America? | 30.3 (30) |
| | What percentage of the world's roads are in India? | 10.5 (11) |
| | What percentage of the world's countries have a higher fertility rate than the United States of America? | 58 (58) |
| | What percentage of the world's telephone lines are in China, the United States of America, or the European Union? | 72.4 (72) |
| | Saudi Arabia consumes what percentage of the oil it produces? | 18.9 (19) |
| | What percentage of the world's countries have a higher life expectancy than the United States of America? | 20.3 (20) |

**Table 5.1 continued**

| **Question-type** | **Question** | **Literature statistic** |
|---|---|---|
| **Undesirable** | What is the chance that you will have gallbladder stones? | 16 |
| | What is the chance that you will have a limb amputated? | 11 |
| | What is the chance that you will die before 90? | 68 |
| | What is the chance that you will have serious hearing problems? | 22 |
| | What is the chance that you will have irritable bowel syndrome (disorder of the gut)? | 30 |
| | What is the chance that you will have hepatitis A or B (inflammation of the liver)? | 36 |
| | What is the chance that you will have an eye cataract (clouding of the lens of the eye)? | 61 |
| | What is the chance that your arteries will harden (narrowing of blood vessels)? | 43 |

### 5.2.3 Procedure

The experiment was conducted over five sessions, all conducted on a local computer network using z-Tree software (Fischbacher, 2007) in the EconLab at RHUL. The experiment lasted approximately 40 minutes. For purposes of non-deceptive payment (see 5.2.3.1 on page 168), in addition to asking participants to provide their *own* estimates for each of the questions, we asked them to estimate the average answers given by *other* participants in the session, in separate rounds.

After the first estimates, participants were shown their initial estimate and given “dialectical bootstrapping instructions” (Herzog & Hertwig, 2009), where participants were asked to consider reasons for why their initial estimate might be incorrect and to take on a new perspective for their second estimate (see Appendix E). Second estimates could not be equal to the first estimate.

A fixed order of rounds was used where participants provided different estimates in each: in round one their own estimate (e.g., "What percentage of the world's countries have a higher fertility rate than the United States of America?"); in round two what they thought the other participants' average estimate for each question was (e.g., "What is the average estimate for the following question: What percentage of the world's countries have a higher fertility rate than the United States of America); in round three an alternative, second own estimate for the exact same questions as in round one; and in round four an alternative, second estimate of the others' average estimate for the exact same questions as in round two. This order was chosen for three reasons. First, to calculate others' average estimate, everyone's own estimates had to precede the average rounds. Second, if participants had been unexpectedly asked for the second own estimate prior to giving their first estimate of the others' average estimate, they might have expected to have to do the same when then asked for others' average. Instead of providing their best estimate on the first guess, they might therefore have chosen to give the lower and higher bounds of their best guess to maximise payment (e.g., if they thought the correct answer was 30, they might first estimate 25 and then 35, as 30 falls into the paid range for both estimates; see below). Therefore, we elicited both sets of first estimates before eliciting the unexpected second sets of estimates. Finally, we thought the interval between the first and second guess should be equal for own and average estimates. The order in which questions were presented within each round was randomised for each participant; this order remained the same across the four rounds.

Participants first provided written informed consent, then read instructions and continued to the task when everyone had finished reading the instructions (see

Appendix E). After completing the four rounds, participants answered demographic questions (gender, age, nationality), check questions (i.e., whether they had experienced or were currently experiencing any of the possible undesirable events), and were paid for their participation.

#### 5.2.3.1 *Incentives*

We incentivised accurate responding for both question types. As participants might have had individuating information regarding their own personal vulnerability to certain future misfortunes (e.g., family history of an illness), we paid them for their accuracy in estimating the average of others' estimates for undesirable questions. These averages, computed in-session, enabled us to pay participants based on their accuracy in estimating objectively correct responses for both neutral *and* undesirable questions. We do not report analyses of these others-estimates, as they were included for logistical reasons and do not bear on our research questions. Although we did not deceive our participants, we did not make it explicit that payment was not based on their *own* estimates for undesirable questions. As far as participants were concerned, they were paid based on their accuracy in estimating both question types (which was true), but did not know upon which particular questions payment was based.

In each round, four of the sixteen questions were rewarded for accuracy. In the rounds concerning own estimates, only the neutral questions were rewarded; half in the first round and the other half in the third round. In the rounds concerning others' average estimate, the undesirable questions were rewarded; half in the second round and the other half in the fourth round.

Across the experiment, all questions were rewarded as follows: 10 points for an estimate that was the true value; 4 points for estimates off by up to 2% in either direction; 2 points for estimates off by up to 5% in either direction, 1 point for estimates off by up to 10% in either direction; and 0 points for estimates that were off by more than 10% in either direction. Each point was worth £0.20.

For the four questions rewarded in the first estimate rounds, the first estimate was used for payment. In the second-estimate rounds, the four previously unrewarded questions were rewarded. For these four questions, the better of the two estimates, whether the first or the second estimate, was used for payment. Participants were informed of this payment scheme, although they were not informed at the outset about having to provide a second estimate, so as to keep participants from providing lower and upper bounds of their estimates.

### 5.2.4 Analytic Strategy

The crowd-within effect was investigated by comparing the absolute error of the first estimate ($A_1$), the absolute error of the second estimate ($A_2$), and the absolute error of the two estimates averaged ($A_{avg}$). Because squared errors penalise large errors more heavily than smaller errors, and therefore favour average errors over either of the errors chosen at random (Soll & Larrick, 2009), we did not measure accuracy through squared errors as Vul and Pashler (2008) did. Instead, errors were calculated as per C. M. White and Antonakis (2013), with the slight alteration of using the mean rather than the median to align with more recent work (Herzog & Hertwig, 2014; personal communication with S. Herzog, 21 January, 2014):

$$A_1 = Mean_{i=1}^{i=8}\left(|R_{1,i} - T_i|\right)$$

$$A_2 = Mean_{i=1}^{i=8}\left(|R_{2,i} - T_i|\right)$$

$$A_{avg} = Mean_{i=1}^{i=8}\,(|\bar{R}_i - T_i|)$$

Here $A_1$ is the mean absolute difference between participants' first responses, $R_1$, and the true values, $T$, across the eight questions ($i$); and $A_2$ is the mean absolute difference between participants' second responses, $R_2$, and the true values, $T$, across the eight questions ($i$). These are compared to $A_{avg}$, which is the mean absolute difference between participants' averaged first and second responses, $\bar{R}$, and the true values. Together, these values constituted the three within-subjects levels of estimate-type. These values were calculated separately for both question-type conditions: neutral and undesirable.

To investigate whether we replicate the crowd-within effect for neutral questions, and if the same effect can be found, to a similar or different degree, for undesirable questions, a 3 (estimate-type: first estimate versus second estimate versus the average of the two estimates) × 2 (question-type: neutral versus undesirable) repeated-measures analysis of variance (RM ANOVA) was conducted on the means of absolute errors of own estimates.

Furthermore, the crowd-within effect is thought to partially stem from bracketing of the true value, where one estimate is an underestimate of the true value and the other an overestimate (Herzog & Hertwig, 2009). We conducted a paired-samples *t*-test on the mean number of questions where the two estimates bracketed the true value for each question type to investigate whether bracketing rates were equal.

Finally, to investigate optimism, log-transformed (see under data screening) signed (i.e., non-absolute) errors were analysed through a 2 (estimate-type: first estimate versus second estimate) × 2 (question-type: neutral versus undesirable) RM ANOVA and through one-sample *t*-tests.

An alpha level of .05 was adopted for these analyses. When the assumption of sphericity was violated, Greenhouse-Geisser corrections or multivariate tests (if ε<.7 in Mauchly's test) were used.

## 5.3 Results

### 5.3.1 Data Screening

Kolmogorov-Smirnov tests were used to check for normality, but as these tests tend to detect even trivial deviations in larger samples (Field, 2013), an $\alpha$-level of .01 was adopted. The tests indicated that signed errors for undesirable questions were not normally distributed for the first estimate (*p*=.004) and for the second estimate (*p*=.005). To correct for positive skew, all signed errors were transformed by first adding a constant to make all values positive and then taking the logarithm (i.e., log($x$+31), where $x$ was the original score) as advised by Field (2013). After this transformation, all variables were normally distributed. Analyses included all participants (n=104). However, seventeen participants endorsed at least one check question. Therefore, we also conducted analyses with a subsample of n=87, consisting only of participants who had not experienced (or were not currently experiencing) any of the undesirable events. These analyses did not lead to different results than those obtained with n=104, and thus only the latter are reported. Reported means and standard errors (and graphical displays) are based on non-transformed data to facilitate interpretation.

### 5.3.2 Crowd-Within Effects

There was a main effect of estimate type, as shown by a 3 (estimate-type) × 2 (question-type) RM ANOVA on absolute errors (Wilks' Lambda=.446, $F(2,102)=63.395$, $p<.001$, $\eta_P{}^2=.554$). Planned comparisons with a Bonferroni correction indicated that the absolute errors of the two estimates averaged (mean (SE) = 17.143 (.346)) were lower than errors of first estimates (18.368 (.367); $p<.001$, 95%-CI [0.833, 1.617]) and errors of second estimates (17.889 (.377); $p<.001$; 95%-CI [0.289, 1.203]). The errors of the first and second estimates did not differ ($p=.334$, 95%-CI [-0.247, 1.206]). There was no main effect of question-type, with equal errors for undesirable questions (18.374 (.425)) and neutral questions (17.225 (.481); $F(1,103)=3.636$, $p=.059$, $\eta_P{}^2=.034$, 95%-CI [-0.046, 2.343]). Estimate-type and question-type did not interact (Wilks' lambda=.991, $F(2,102)=.464$, $p=.630$, $\eta_P{}^2=.009$). In summary, we found an overall crowd-within effect, which was not moderated by question type. Figure 5.1 shows these results.

There was no significant difference in the number of questions for which first and second estimates bracketed the true value between neutral questions and undesirable questions ($t(103)=1.611$, $p=.110$, $d=.317$, 95%-CI [-0.064, 0.622]). Of the eight neutral questions, 1.90 (.138) questions were bracketed on average. Of the eight undesirable questions, 1.63 (.134) questions were bracketed on average.

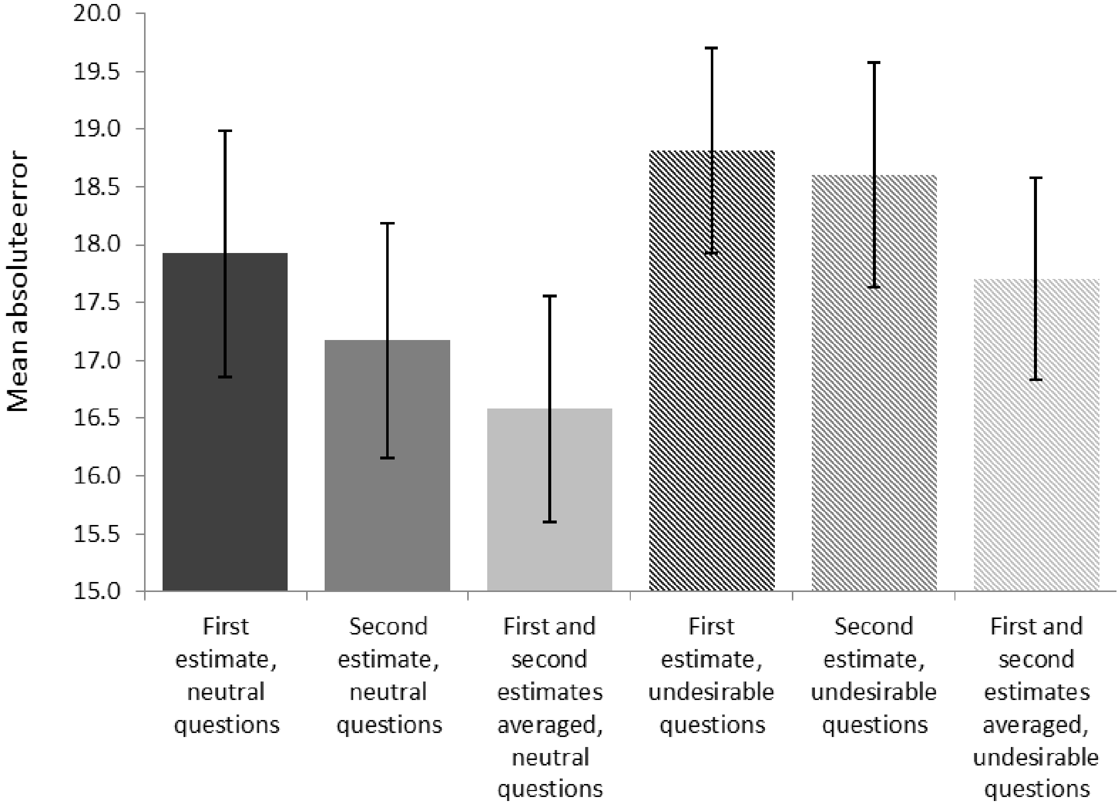


**Figure 5.1** The mean absolute error (±95%-CI) of first estimates, second estimates, and the first and second estimates averaged, for both neutral and undesirable questions.

### 5.3.3 Optimism Effects

There was a main effect of estimate-type, such that second estimates (-1.952 (.817)) were lower than first estimates (-.290 (.810)), as revealed by a 2 (estimate-type) × 2 (question-type) RM ANOVA on the log-transformed signed errors ($F$(1,103)=11.468, $p$=.001, $\eta_p^2$=.100, 95%-CI [0.654, 2.671]). There was also a main effect of question-type, such that estimates for undesirable questions (-4.770 (.993)) were lower than estimates for neutral questions (2.528 (.953); $F$(1,103)=43.993, $p$<.001, $\eta_p^2$=.299, 95%-CI [4.950, 9.645]). Estimate-type and question-type interacted as well ($F$(1,103)=8.604, $p$=.004, $\eta_p^2$=.077). Planned contrasts with a Bonferroni correction showed that there was no difference between the first and second estimates for neutral questions (3.020 (1.028) and

2.036 (.952), respectively; *p*=.165; 95%-CI [-0.088, 2.056]). For undesirable questions, second estimates (-5.940 (1.069)) were lower than first estimates (-3.599 (1.033); *p*=.001, 95%-CI [0.972, 3.709]). Figure 5.2 shows these results.

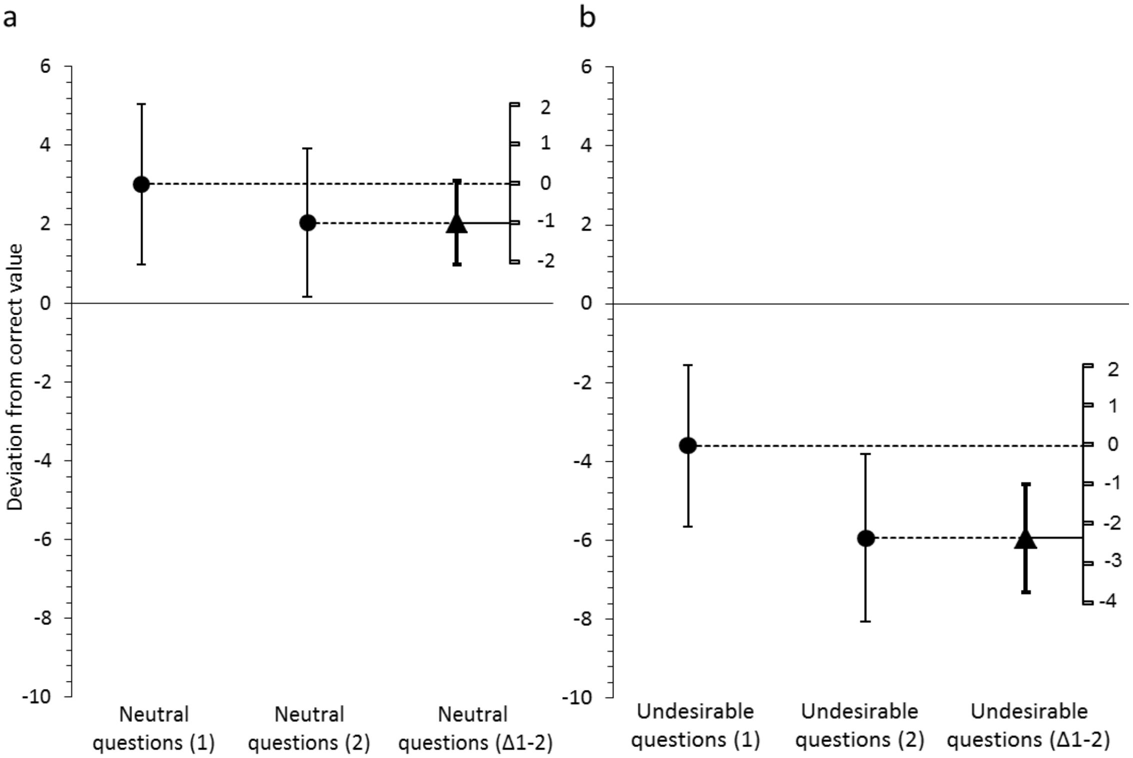


**Figure 5.2** The mean biases of the first (1) and second (1) estimates (circles and left vertical axis in each panel) and the differences (Δ1-2) between these estimates (triangles and right vertical axis in each panel), with 95% confidence intervals, for neutral questions (panel a) and undesirable questions (panel b).

As the above analysis showed they were not different, the log-transformed signed errors of first and second estimates for neutral questions were collapsed and a one-sample *t*-test indicated that they were not different from equivalently-transformed zero (i.e., log(0+31); *t*(103)=1.183, *p*=.720 (Bonferroni-corrected), mean difference=.015, 95%-CI [-0.010, 0.041]). As the planned contrasts indicated that log-transformed errors of first and second estimates for undesirable questions were not equal, they were analysed through separate one-sample *t*-

tests. In both cases the log-transformed errors were lower than transformed zero, i.e., optimistically biased (first estimate: *t*(103)=-4.965, *p*<.001 (Bonferroni-corrected), mean difference=-.086, 95%-CI [-0.120, -0.052]; second estimate: *t*(103)=-6.323, *p*<.001 (Bonferroni-corrected), mean difference=-.140, 95%-CI [-0.183, -0.096]).

In summary, one's estimates for undesirable (but not neutral) questions were optimistically biased, the second estimate more so than the first (see Figure 5.2).

## 5.4 Discussion

The crowd-within effect describes the phenomenon where the average of two estimates from the same person has a lower error than either of the individual estimates from that person on average (Vul & Pashler, 2008). In the present study we found this crowd-within effect both for neutral questions, thereby replicating results from previous studies (e.g., Herzog & Hertwig, 2009; Vul & Pashler, 2008), and for undesirable questions, equally.

Furthermore, consistent with the optimism bias (Sharot, 2011a; Weinstein, 1989), we found that participants consistently and significantly underestimated the answers to undesirable, but not neutral, questions. Moreover, and as the most important aim of the present study, we investigated self-deception. Comparing signed errors of first and second estimates for both types of questions suggested the second estimates for undesirable questions were significantly rosier than the first estimates, while no difference was found for first and second estimates regarding neutral questions.

The significance of this latter result is that it indicates participants were sampling *selectively* from an internal probability distribution for undesirable questions. Whatever the shape of the underlying distribution, if they had been sampling *randomly* (as per Vul & Pashler, 2008), participants would have been just as likely to provide a more optimistic second estimate as a less optimistic second estimate.[15] As such, no *systematic* difference in bias would have emerged across estimates. Our results imply participants carried a more accurate, and less rosy, representation of their future prospects than their individual estimates (at least their *second* estimates) for undesirable questions conveyed. Our results are consistent with the "real" self-deception account (Gur & Sackeim, 1979; Mijovic-Prelec & Prelec, 2010). Just as someone might sample selective information to give another person a desirable impression of themselves (e.g., showcasing specific, rather than random, examples of previous employment in a job interview), people might mislead *themselves* by sampling selective examples which would convey desirable information regarding their future prospects. Likewise, just as people exaggerate their prospects when their claims are challenged (Gal & Rucker, 2010), and potentially "protest too much", our findings suggest this same defensive strategy may operate intrapersonally (McKay, Mijović-Prelec, & Prelec, 2011).

The results of the present study raise several questions. For example, one might wonder how the crowd-within effect could obtain for undesirable questions, given that second estimates were more biased, on average, than first estimates for such questions. Note that although the signed errors were different for the first and second estimates of undesirable questions, the *absolute* errors, upon

[15] Note that Lench, Smallman, Darbor, and Bench (2014) have recently found that people perceive greater variance for given probabilities or given ranges of probabilities of desirable outcomes for the self, but not for others. As such, the internal probability distribution for the self might have a higher variance than the distribution for others, so that any estimate away from the (accurate) mean, including any optimistic estimate, is more likely to be given for the self than for others. However, this does not make it more likely for the second estimate to be systematically more optimistic than the first estimate.

which the crowd-within effect is based, were not. First estimates to some questions could have been overestimations, while all second estimates might have been underestimations. For the absolute errors, both these types of errors could add up to an equal value for first estimates as for second estimates when averaged across questions. However, for signed errors, some overestimations and underestimations could have cancelled out, leading to a smaller averaged signed error for first estimates compared to second estimates. With some overestimations in the first estimates, but none in the second estimates, one would expect some bracketing of true values, which we indeed found. As Vul (n.d.) noted, even low rates of bracketing can correct for higher absolute errors for second estimates and still lead to the crowd-within effect.

Second, one might wonder if the optimistic estimates participants provided did not convey their actual beliefs, but were distorted for impression formation purposes (e.g., to deceive the experimenters of participants' low chance of misfortune). Against this possibility, we note that our participants provided estimates under conditions of strict anonymity and they stood to gain financially by providing accurate estimates. Mijovic-Prelec and Prelec (2010) and Simmons and Massey (2012) have shown that participants supply optimistic estimates even in the face of substantial incentives to be accurate.

One might wonder why the data pertaining to others' estimates was not analysed to investigate comparative optimism (i.e., comparing the self to others), in addition to the absolute optimism we did investigate. Here, one crucial methodological aspect should be noted. In comparative optimism, participants are asked to estimate the likelihood that others will experience relevant outcomes (Garrett & Sharot, 2014; Shepperd et al., 2013). In the present study, however, the "other estimates" involved asking participants what they thought others had estimated for themselves. This phrasing could lead to different results compared to estimating another's likelihood if people are aware of the optimism bias. Given the coverage of the optimism bias in the media (e.g., Cadwalladr, 2012), a popular psychology book (Sharot, 2011b), and a TED talk on the topic (Sharot, 2012), some of our participants may have had a certain degree of awareness of the optimism bias. With this knowledge, participants might assume others would underestimate their risk, and they would adjust their others-estimate accordingly. This would then confound the interpretation of any analyses between own-estimates and others-estimates in this study. Hence, we decided not to analyse these estimates, although they maintained their practical purpose for non-deceptive payment.

Finally, following previous investigations of the optimism bias (Sharot, 2011; Sharot et al., 2011; Weinstein, 1989), we deliberately chose to use negative items as they carry more relevance in terms of taking action to reduce risks (Weinstein & Klein, 1995). Future studies could include desirable events, or could ask participants to estimate their chance of *not* experiencing the undesirable events in question (Sharot et al., 2011), and investigate if second estimates are higher than first estimates for such items. If so, this would strongly support the notion of self-deceptive selective sampling we propose here.

The negating phrasing of "not experiencing" undesirable events was avoided, due to concern that if the undesirable events described would be more salient than the negating word, the events might still lead to negative affective states. This, in turn, could potentially influence the optimism bias (Helweg-Larsen & Shepperd, 2001). Admittedly, this concern has not been empirically tested. Therefore, future studies could include a condition with negating questions to investigate if second estimates for negated undesirable questions are higher than the first of such estimates, while also including measures of mood. It must be noted that more than the current eight items might need to be included for reliable measurement if the number of items is to be split across conditions.

## 5.5 Conclusion

In the present study, we combined the optimistic belief-updating and crowd-within paradigms to investigate whether "self-deceptive" processes underlie the optimism bias. First, we found the crowd-within effect (i.e., lower errors for averaged estimates than for either estimate alone) for neutral and undesirable questions. Second, we found optimism as the true values for undesirable questions were underestimated, but neutral questions were accurately estimated. Finally, and most importantly, we found self-deception as the second estimate for undesirable questions was systematically lower than the first, suggesting systematic, biased sampling from an internal probability distribution. First and second estimates for neutral questions were not systematically different, suggesting random sampling from an internal probability distribution in those cases. This systematically biased sampling supports the "real" self-deception account.

Overall, the present study's findings are not especially optimistic about the possibility of correcting the optimism bias, when it would be desirable to do so, such as when forecasting changes in stock markets. The results indicate that second guessing oneself and taking the average of the two guesses may improve accuracy somewhat compared to taking either of the two guesses at random. However, this seems to only minimise random error.

Unrealistic optimists seem less willing to have their beliefs tested when this could require updating of those beliefs. This might be because they have rosier beliefs than warranted, but also because finding out bad news might be more detrimental to people with optimistic beliefs. This last possibility is investigated in Chapter 6, which investigates the phenomenon of betrayal aversion and so the thesis moves from optimism about one's future prospects to pessimism about others' trustworthiness.

# 6 Betrayal Aversion

## 6.1 Background

As discussed in Chapter 1, in the binary trust game, a sender first decides whether to opt in or opt out of a trust game. If the sender opts out, the sender and (generally) the trustee receive a small reward (e.g., 10|10 for the sender and trustee, respectively). If the sender opts in, the trustee can then decide to reciprocate trust and pick a fair outcome (e.g., 15|15) or to betray trust and choose an unfair outcome (e.g., 8|22; Berg et al., 1995; Camerer & Weigelt, 1988). According to the neo-classical economic model, trustees should maximise their material return and thus choose the unfair outcome, and anticipating this, senders should opt out. Yet, people are found to trust and to be trustworthy in the trust game (Camerer, 2003), which is irrational in light of the neo-classical economic model as it does not maximise material self-interest (Glimcher, Fehr, Camerer, & Poldrack, 2008; Manapat et al., 2013). Besides these behaviours, betrayal aversion has been observed in the trust game, which forms another irrational behaviour from a neo-classical economic point of view (Fehr, 2009).

Briefly, betrayal aversion is shown when participants indicate they need more certainty of receiving the good outcome when another player selects that outcome compared to when the outcome is selected by a random process (i.e., a computerised lottery). For example, if a trustee determines the outcome, a sender might require the probability that the trustee will select the good outcome to be at least 70% in order to opt in. That same sender might require a lottery's probability of the good outcome only to be at least 60%.

As described in Chapter 1, Aimone and Houser (2012) argue that betrayal aversion occurs because people want to avoid the emotional costs of finding out they have been personally betrayed after trusting someone. These authors have now reported support for this suggestion from an imaging study, which found more insula activity when opting into the trust game compared to making risky decisions in a computerised lottery (Aimone, Houser, & Weber, 2014), with insula activity suggested to signal aversive emotions.

However, some studies have not replicated the betrayal-aversion phenomenon. Bohnet and Zeckhauser (2004) first reported this phenomenon and replicated it across several countries (Bohnet et al., 2008), while Fetchenhauer and Dunning (2012) did not find evidence for betrayal aversion in a slightly different paradigm (described below). As mentioned in Chapter 1, differences in methodologies may have accounted for the presence versus absence of betrayal aversion.

Bohnet and Zeckhauser (2004) asked participants what the minimum probability of receiving the good outcome had to be in order for them to opt into a game, rather than opt out. Three different games were used, which varied in whether a computerised lottery or another person selected an outcome and in whether another person’s payoffs depended on the sender’s decision. Participants reported a higher required minimal probability of the good outcome if another person were to select the outcome than if a computerised lottery were to select the outcome.

Fetchenhauer and Dunning (2012) took a different approach and asked senders if they wanted to opt out (5|0 for the sender and trustee, respectively) or opt in to the trust game. If opting in, the trustee would pick between the good outcome

(10|10) and the bad outcome (0|20). In an additional task, these authors asked participants whether they wanted to keep 5, or opt into a lottery where they could win 0 or 10. One group of participants was informed that the probability of the good outcome, in both the trust game and in the lottery, was 46%; another group was informed it was 80%. When the chance of the good outcome was high, senders opted into the trust game and the lottery at equal rates. However, when the chance of the good outcome was low, 28.6% of the senders opted into the lottery, while 54.3% opted into the trust game. This finding does not support the notion of betrayal aversion, as participants were *more* willing to accept the risk of the bad outcome through another player's choice than through a computerised lottery.

One methodological difference between Bohnet and Zeckhauser's (2004) and Fetchenhauer and Dunning's (2012) studies is inequality of payments to the sender and trustee if the sender decides to opt out. Social motives such as altruism, efficiency motives (i.e., larger total payoffs if opting in), or inequality aversion (Fehr & Schmidt, 1999), could have led participants in Fetchenhauer and Dunning's (2012) study to opt into the trust game. However, opting in could have resulted in inequality for the sender, especially if the probability of the good outcome is low. Hence, inequality-averse senders might opt out in this scenario. Yet, high levels of opting in were found. Therefore, inequality aversion cannot fully explain the differences between the studies. Nevertheless, in the present study, we ensured that both players in a dyad would receive a positive sum of money, irrespective of which outcome was selected and how it was selected, and that the outcomes had equal efficiency.

A second methodological difference between the two studies lies in how participants construe the trust scenario. In Bohnet and Zeckhauser's (2004) study, participants gave a conditional, minimal acceptable probability of the good outcome in the lottery or trust game to signal the minimal level of trust needed, without specifically signalling distrust to the trustee. In Fetchenhauer and Dunning's (2012) study, participants could opt in or out, signalling distrust when doing the latter, as trustees learned that the sender opted out. Signalling distrust could be of great influence on behaviours in economic games. Much as people are willing to pay a cost to avoid displaying unfairness in the dictator game (see Dana, Cain, & Dawes, 2006)[16], some people might be willing to forego larger rewards in the trust game if this allows them to avoid signalling negative qualities about themselves (e.g., distrust of others). We avoided this confound by asking participants for their minimal acceptable probabilities of the good outcome, so that they would not signal distrust directly if preferring the computerised lottery over the trust game.

We developed a theory to explain betrayal aversion, which may also account for differences in the aforementioned studies with contradictory findings. We build on Aimone and Houser's (2012) notion that betrayal aversion might be influenced by emotional costs. In the present study, we hypothesised that prior beliefs about others' trustworthiness could predict betrayal aversion. Intuitively, one might expect that if people believe others are untrustworthy, they would not trust another with their payment. Thus, one might expect trustworthiness-

[16] This cost was not a fair share of the endowment, but Dana et al. (2006) asked dictators to split $10 between themselves and the recipient. After dictators had made their decision, they were presented with the option to exit the dictator game and receive $9, in which case the recipient would get nothing but would also not be informed about the fact that a dictator game had been played. This exit option is costly to the dictator as they could have kept $10 ($1 more than the exit option), without any change in outcome for the recipient, or they could have kept $9 (equal to the exit option), with a positive reward for the recipient ($1 more than the exit option). Yet, the exit option, taken by 33% of the dictators, provides a way of being selfish without showing such unfairness to another player.

beliefs to predict more opting in. Yet, this would not explain betrayal aversion. Instead, we suggest that betrayal aversion might be explained by different utility functions, which factor in prior beliefs about others' trustworthiness. When one decides to trust, and subsequently is betrayed, not only does the bad outcome lead to disutility, but there are emotional costs from knowing that one has been betrayed (Aimone & Houser, 2012) and additional disutility from knowing that the betrayer received extra money with their unkind action. The emotional costs and additional disutility from knowing the betrayer received more money is perhaps less for someone who initially believes they are likely to be betrayed than for someone who is surprised by this. When one decides to trust, and gets the good outcome, the good outcome leads to an increase in utility, and additional utility may accrue from knowing that the person was trustworthy (i.e., honor benefits; Bohnet et al., 2008). This additional utility might be higher for someone who believed he or she was likely to be betrayed than for someone who was expecting this to happen.

The differences in the additional (dis)utility could be underpinned by dopamine prediction errors. Rewarding sensations are reflected in prediction errors coded through increased dopamine release for outcomes better than expected and decreased dopamine release for outcomes worse than expected (Schultz, 1998). If one's prediction is that one will be betrayed (i.e., a pessimistic belief), and this consequently occurs, the outcome is as expected. Similarly, if one does not expect to be betrayed (i.e., an optimistic belief), and trust is reciprocated, the outcome is as expected. However, if one believes that one will be betrayed and consequently receives the good outcome, the outcome is better than predicted, a positive prediction error. In contrast, if one believes one's trust will be rewarded, and consequently gets betrayed, this results in a negative prediction error. Therefore, when one is pessimistic, yet opts into the trust game, the dopamine release upon seeing the outcome will be as expected or better; a positive, rewarding sensation. Hence, participants with a pessimistic belief should be relatively more willing to opt into a trust game compared to a lottery or should at least be indifferent about playing the trust game versus a lottery. On the other hand, optimistic participants receive, at best, the expected dopamine release, or, if betrayed, less than expected; a negative sensation. Hence, participants with an optimistic belief should be relatively less willing to opt into a trust game and have their beliefs tested compared to a lottery, or should be indifferent between the two methods of determining an outcome. As an analogy, consider someone with an optimism bias (see Chapter 5) who thinks they are healthy (i.e., a positive belief): they might not be particularly eager to have that belief tested by subjecting themselves to medical tests and potentially facing evidence that challenges the positive belief. Someone who thinks they are ill (i.e., a negative belief) might be more eager to undergo medical tests, in case they obtain evidence that would speak against the negative belief (or, of course, obtain confirmation of the illness which they could then obtain treatment for).

This could explain why, in a different study, Fetchenhauer and Dunning (2009) found that participants simultaneously believed others' trustworthiness to be lower than it actually was, and yet trusted others more than they should have based on their misguided beliefs. This seems to go against the intuitive notion of betrayal aversion, as one might expect those who anticipate betrayal to avoid this possibility by opting out. Yet, the findings are in line with the theory presented in the present study.

Aimone and Houser (2012) argue that participants trust more when they can avoid information about betrayal. However, they overlooked the fact that the computer’s selection of a bad outcome still provides information about the trustworthiness of people in general, as the selection procedure is based on the proportion of unfair trustees. Therefore, people are still exposed to information that might challenge their beliefs, not about their specific counterpart per se, but rather about the general population’s trustworthiness. This is presumably the belief being used in the trust game, given that the identity of the trustee is never revealed and the trustee thus constitutes a random member of the general population. Furthermore, Aimone and Houser (2012) did not measure beliefs and hence their study cannot shed light on the association between beliefs about others’ trustworthiness and betrayal aversion. The present study measured such beliefs and investigated whether they can predict levels of betrayal aversion.

### 6.1.1 The Present Study and Hypotheses

In the present study we wanted to test the association between beliefs about others’ trustworthiness and the choice between a trust game and a lottery, in order to illuminate the notion of betrayal aversion. Participants in the role of the sender provided an incentivised estimate of their belief of others’ trustworthiness (i.e., trustworthiness-beliefs) and indicated at what probability of the good outcome they would prefer playing a lottery rather than have a trustee decide on an outcome. They set this probability by requesting a minimal number of white beads (representing the good outcome) in an urn of 1000 beads. If the urn held at least the requested number of white beads, the computer drew an outcome from the urn. Otherwise, the decision of the trustee was implemented. Figure 6.1 (on page 192) illustrates this task.

In our design, senders determined the level at which another person’s decision versus the computer’s selection would determine the outcome. This avoided having the minimal acceptable probability represent opting in and out of the risky option, and thereby avoids confounding loss aversion and betrayal aversion (Aimone & Houser, 2012). This also avoided any influence of efficiency motives, where people prefer options with the highest *overall* payoffs. In the standard binary trust game, the overall payoffs of opting out are smaller than the overall payoffs of opting in, which might lead people to opt in to increase efficiency. In the present study, trusting a trustee or a computer had the same overall payoffs (i.e., the good outcome in the trust game and the lottery was 15|15, while the bad outcome in both scenarios was 8|22), and participants' decisions should thus not have been influenced by efficiency motives.

If utility depended only on the outcome obtained, a rational act would have been to request as many white beads as represented the believed number of fair trustees. If a sender believed seven out of eight trustees selected the good outcome (7/8=87.5%), this would suggest that with 875 white beads in the urn, the sender would be impartial between the two ways of implementing an outcome. The number of white beads that matched the trustworthiness-belief, and indicated the point at which participants would be indifferent between the two methods of implementing the outcome, represented the belief-equivalent number of beads. Betrayal aversion was expressed by requesting a lower minimal number of white beads than the belief-equivalent number suggested. This translated into a negative deviation of the number of requested white beads from the belief-equivalent number. A negative deviation indicated that participants accepted a higher risk of receiving the bad outcome from the computer than run the risk of having another player betray them.

We predicted that participants' trustworthiness-beliefs would influence their level of betrayal aversion. Specifically, we predicted that optimists would show more betrayal aversion and pessimists would show less betrayal aversion. In this study, optimism was defined as believing others are trustworthy (i.e., holding high trustworthiness-beliefs). The higher one's trustworthiness-beliefs, the more disutility would be obtained from being betrayed, and the more one might avoid exposure to potential betrayal. Therefore, we predicted a negative association: the higher the number of trustworthy people was believed to be, the more negative the deviation from the belief-equivalent number of beads.

In order to test the robustness of the predicted association, several potentially influential factors were measured and accounted for in analyses. First, as evidence regarding the similarity between the willingness to trust and the willingness to take risks is mixed (Bohnet & Zeckhauser, 2004), risk-aversion was measured in order to account for the willingness to take risks. Second, beliefs about others' trustworthiness might be influenced by paranoia. The more paranoid thoughts one has, the lower one's trust in others. This paranoia might extend to distrust in the experimenters' (true) claim of a random computerised draw from the urn. Thus, paranoia would not only affect the beliefs about the number of trustworthy trustees, but also potentially the beliefs about the fairness of the computer lottery. In order to account for possible influences of paranoia, a questionnaire measuring paranoia was included. Third, reciprocity, which "is a behavio[u]ral response to perceived kindness and unkindness" (Falk & Fischbacher, 2006, p. 294), might affect expectancies of others' trustworthiness and levels of opting into the trust game (Naef & Schunk, 2009). Someone who feels strongly about reciprocity might experience more disutility from learning that a trustee received a higher payoff by betraying a sender. In contrast, someone who does not feel strongly about reciprocity would mainly be interested in his own payoffs, and not experience additional (dis)utility from the payoff a trustee receives. As such, our hypothesised mechanism might only apply to participants with stronger reciprocity norms. Therefore, we measured reciprocity norms as well.

## 6.2 Methods

### 6.2.1 Participants

Participants were 208 students from RHUL (mean (SD) age = 20.75 years (2.65 years); 77 male, 131 female). Participants received a decision-based payment between £4 and £14.85 (mean (SD) = £7.85 (£2.30)), which combined the results from the trust game or lottery, the trustworthiness-belief question if answered correctly (only for senders), and the risk-aversion measure. Half of the participants were randomly assigned to the role of a sender, while the other half were assigned to the role of a trustee. The analyses reported here focused on senders (n=104; mean (SD) age = 21.09 years (3.25 years); 36 male, 68 female). The Psychology Department Ethics Committee of RHUL approved this study.

### 6.2.2 Materials

#### *6.2.2.1 Trust Game*

Although descriptions throughout this chapter use terms such as “trust game”, “trustee”, and “good outcome”, neutral language was used throughout the experiment (i.e., “Player X” was the sender, “Player Y” was the trustee, “Outcome A” was the good outcome, and “Outcome B” was the bad outcome) to minimise demand effects (Aimone & Houser, 2012). Furthermore, in order to avoid demand effects with regards to directly translating trustworthiness-beliefs into actions, the use of the term “probability” was also avoided. The game had two outcomes: the good outcome with 15|15 points for the sender and the trustee, respectively; and the bad outcome with 8|22 points. All participants were informed that one of the two outcomes would be arrived at in one of two possible ways: the computer would decide or Player Y (i.e., the trustee) would decide.

Trustees were always asked to select an outcome, in case their decision would be implemented. They were informed that if the computer selected an outcome, their choice would not be relevant for payment and player X (i.e., the sender) would not learn about their choice.

Senders were informed that they could influence how the outcome would be determined. An urn was filled with 1000 neutral (grey) beads, which would be replaced by $x$ white beads and 1000 – $x$ black beads. Note that white represents the colour for the good outcome in this chapter, but whether white or black beads represented the good outcome was counterbalanced across participants in the experiment. Senders were shown the urn with 1000 neutral (grey) beads and had to indicate how many white beads, at minimum, they wanted to be in the urn for the computer to draw a bead from the urn in the lottery ($y$). If there were fewer white beads ($x$) than the requested number of white beads ($y$), the outcome would be decided by the trustee's decision. If there were more white beads than (or as many as) requested (i.e., $x \geq y$), the computer randomly selected an outcome from the urn with the ratio of $x$ white beads and 1000 – $x$ black beads. If the bead was white, the good outcome was selected; if it was black, the bad outcome was selected. Figure 6.1 shows how the outcome was determined, based on the sender's requested minimal number of white beads ($y$).

Several examples and comprehension checks were included before senders indicated how many white beads they wanted at minimum for the computer to draw an outcome (*y*). The minimal number of white beads requested (*y*) is the equivalent of Bohnet and Zeckhauser's (2004) minimal acceptable probability of the good outcome in the lottery or trust game.

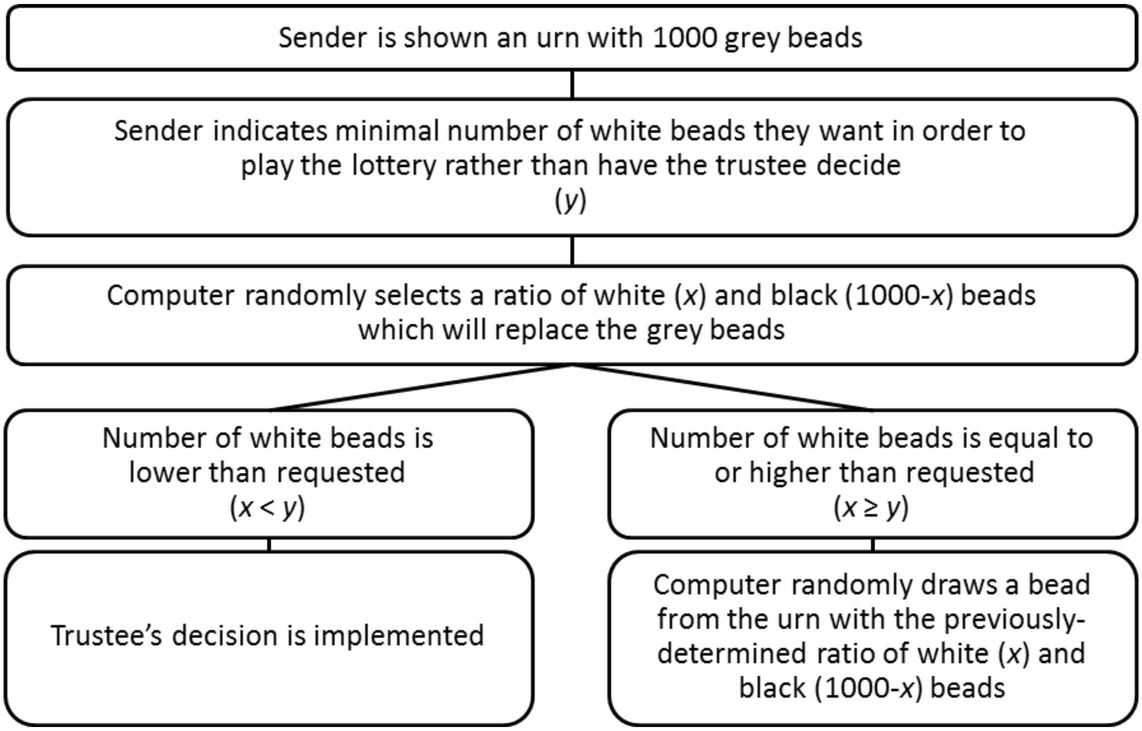

Sender is shown an urn with 1000 grey beads
Sender indicates minimal number of white beads they want in order to play the lottery rather than have the trustee decide (y)
Computer randomly selects a ratio of white (x) and black (1000-x) beads which will replace the grey beads
Number of white beads is lower than requested (x < y)
Number of white beads is equal to or higher than requested (x ≥ y)
Trustee's decision is implemented
Computer randomly draws a bead from the urn with the previously-determined ratio of white (x) and black (1000-x) beads

**Figure 6.1** The task procedure, depicting how the outcome was determined. Participants set their minimal level of probability of a good outcome at which they would prefer a lottery over the trust game. Random draws from the computer determined if this level was met; and, if so, which outcome was selected in the lottery. Grey beads are neutral and are replaced by white beads (representing the good outcome) and black beads (representing the bad outcome) when the computer randomly draws the number of white beads (*x*).

Senders also indicated how many of the trustees they thought would choose the good outcome and how many would choose the bad outcome. This question was incentivised to encourage truthful reporting of participants' trustworthiness-belief (Schotter & Trevino, 2014). If the answer was exactly correct, senders gained an additional 10 points at the end of the experiment; if it was incorrect, they did not receive any points and were not informed about the correct answer.

Senders were also asked to indicate their confidence that their belief was close to the correct answer (off by one trustee at most), on a four-point Likert scale: 0 = not at all confident; 1 = somewhat confident; 2 = quite confident; and 3 = very confident. Furthermore, senders were asked which option they thought was more likely if their belief was incorrect: that there were more trustees who chose the good outcome, that there were fewer trustees who chose the good outcome, or that these two possibilities were equally likely. All these questions were asked before senders learned which outcome was received and how it was selected.

#### *6.2.2.2 Risk Aversion*

For a description of the computerised risk-aversion measure (Holt & Laury, 2002), see Chapter 2 (2.2.2.2 on page 88). As an improvement to previous studies in this thesis, the risk-aversion measure in this study was genuinely paid for one randomly-selected participant in the session, regardless of their role. For this participant, a decision from this measure was randomly selected and their selected lottery was played and the bonus added to their payment.

#### *6.2.2.3 Questionnaires*

##### *6.2.2.3.1 Paranoia*

The Paranoia/Suspiciousness Questionnaire (PSQ; Rawlings & Freeman, 1996) consists of 47 questions, such as “When people are especially nice, do you wonder what they want?” or “Do you feel at times that you’ve got a raw deal out of life?”, with yes/no responses. This questionnaire is designed to measure paranoia and suspiciousness in the general population. The summed total score can range from 0 to 47.

#### *6.2.2.3.2 Reciprocity*

A selection of six questions from the personal norm of reciprocity questionnaire (PNR; Perugini, Gallucci, Presaghi, & Ercolani, 2003) was used. The original PNR measures three aspects of reciprocity (beliefs in reciprocity; positive reciprocity, which is the behaviour in response to kind actions; and negative reciprocity, which is the behaviour in response to unkind actions), all measured with nine items each. We used a selection of six items that has been used before (Dohmen, Falk, Huffman, & Sunde, 2009) and consists of items with the highest factor loadings (all factor loadings >.70; Perugini et al., 2003): three for positive reciprocity (e.g., “If someone does a favour for me, I am ready to return it”) and three for negative reciprocity (e.g., “If somebody puts me in a difficult position, I will do the same to him/her”), and none for beliefs in reciprocity. Each item was scored on a seven-point Likert scale ranging from strongly disagree (1) to strongly agree (7). The positive and negative reciprocity items were then combined, so that the summed total score can range from 6 to 42.

### 6.2.3 Procedure

The experiment comprised 13 sessions, all conducted on a local computer network using z-Tree software (Fischbacher, 2007) in the EconLab at RHUL. The experiment lasted approximately 40 minutes. All sessions had sixteen participants (i.e., eight senders and eight trustees), so that the belief question for trustees was equally sensitive across all sessions (i.e., nine options ranging from 0 to all 8 trustees choosing the good outcome).

Before being assigned their roles, all participants read general instructions regarding the possible outcomes and the possible ways of arriving at one of the outcomes (i.e., computer's or trustee's selection). After correctly answering comprehension questions about these general instructions, participants were assigned their roles and their one specific partner, whose identity they would never learn.

Trustees were asked to select one of the two outcomes, in case the outcome would be determined through their decision. They completed the risk-aversion measure after this. Next, they completed tasks for an unrelated, non-incentivised, pilot experiment.

Senders were provided with instructions about how to influence how the outcome would be determined (i.e., setting a threshold for the number of white beads). After correctly answering the comprehension question about this aspect of the task, they indicated the minimal number of white beads they wanted to be in the container. Senders also provided their belief regarding the number of trustees who would pick the good and the bad outcomes, and indicated their confidence in this belief. Whether the belief question or the task of setting the minimal number of white beads was presented first was counterbalanced across participants. Next, senders also completed the risk-aversion measure, the PSQ, and the reciprocity questionnaire.

Finally, all participants answered demographic questions (gender, age, nationality). Then, they were informed about the outcome and how it was selected, and were paid for their participation, with experimental points being converted to British currency using the exchange rate of 1 point = £0.50. At this feedback stage, senders were also informed if their beliefs were correct or not (and thus if they received an additional 10 points or not).

### 6.2.4 Analytic Strategy

If utility were to depend only on the outcome, a sender would place the number of white beads at the percentage representing the number of trustees the sender believed would choose fairly (i.e., the belief-equivalent number of beads). The deviation from this belief-equivalent number was calculated by subtracting the belief-equivalent number of beads from the requested minimal number of white beads. Betrayal aversion was then defined as negative deviations from the belief-equivalent number of white beads. To test if people were betrayal averse, we used a one-sample *t*-test assessing whether the deviations from the belief-equivalent number of beads were significantly below zero.

Furthermore, we investigated whether beliefs about how many trustees would choose fairly (i.e., trustworthiness-beliefs) could predict the number of white beads requested, and more importantly, whether beliefs could predict deviations from belief-equivalent numbers. To this end, linear regressions with trustworthiness-beliefs predicting the minimal number of white beads requested and with trustworthiness-beliefs predicting the deviation from the belief-equivalent number of white beads requested were conducted. The robustness of any such associations was checked through hierarchical linear regressions. In the first model, risk-aversion, reciprocity norms, and paranoia were accounted for; in the second model, trustworthiness-beliefs were added as an additional predictor to see if they could explain unique variance not already accounted for by the factors included in the first model.

One potential limitation with the analyses listed above is that we might find a negative relationship between trustworthiness-beliefs and deviations from belief-equivalent numbers of beads even if these two are not related. This is because people with low trustworthiness-beliefs, and thus a low belief-equivalent number of beads, will have more space to deviate in the positive direction than in the negative direction. Conversely, people with high trustworthiness-beliefs have more space to deviate in the negative direction than in the positive direction. This, then, could lead to a negative association between beliefs and deviations.

Therefore, as a test for robustness, we also conducted regressions on only those data points that could deviate from the belief-equivalent number of beads equally in both directions. This means that these follow-up analyses included only those participants with trustworthiness-beliefs between one and seven trustworthy trustees. Senders who believed no one or everyone (i.e., 0 or 8 trustees) was trustworthy could only deviate from their belief-equivalent number (i.e., 0 or 1000, respectively) in one direction: the hypothesised one. Therefore, senders with these extreme trustworthiness-beliefs were excluded in this robustness check. Furthermore, of the included trustworthiness-belief levels, only data points of requested numbers of beads which fell within an equidistant range of the belief-equivalent number were included so that deviations in both directions were equally possible. For example, for senders who believed 3 trustees were trustworthy, the belief-equivalent number of beads would be 375 (3/8=.375), and as long as their requested number of white beads was between 0 and 750, their data were included. All other data points were excluded. As such, this robustness test was conducted on a subgroup of participants who could deviate in both directions of their belief-equivalent number of beads equally. This group will simply be referred to as “subgroup” in the results.

With this more stringent selection of data points, the regression analyses might be underpowered to detect an effect. Therefore, group comparisons based on a median split of beliefs were also conducted as a robustness check. Any participants falling at the median of the beliefs of number of trustworthy players were excluded from these analyses. Pessimists, those with trustworthiness-beliefs lower than the median, were compared to optimists, those with trustworthiness-beliefs higher than the median, in analyses of variance (ANOVAs) and analyses of covariance (ANCOVAs), accounting for risk-aversion, reciprocity norms, and paranoia.

## 6.3 Results

### 6.3.1 Data Screening

Figure 6.2 gives an overview of the data, the belief-equivalent numbers of beads at each level of trustworthiness-beliefs, and the range of data included in the subgroup analyses (n=35).

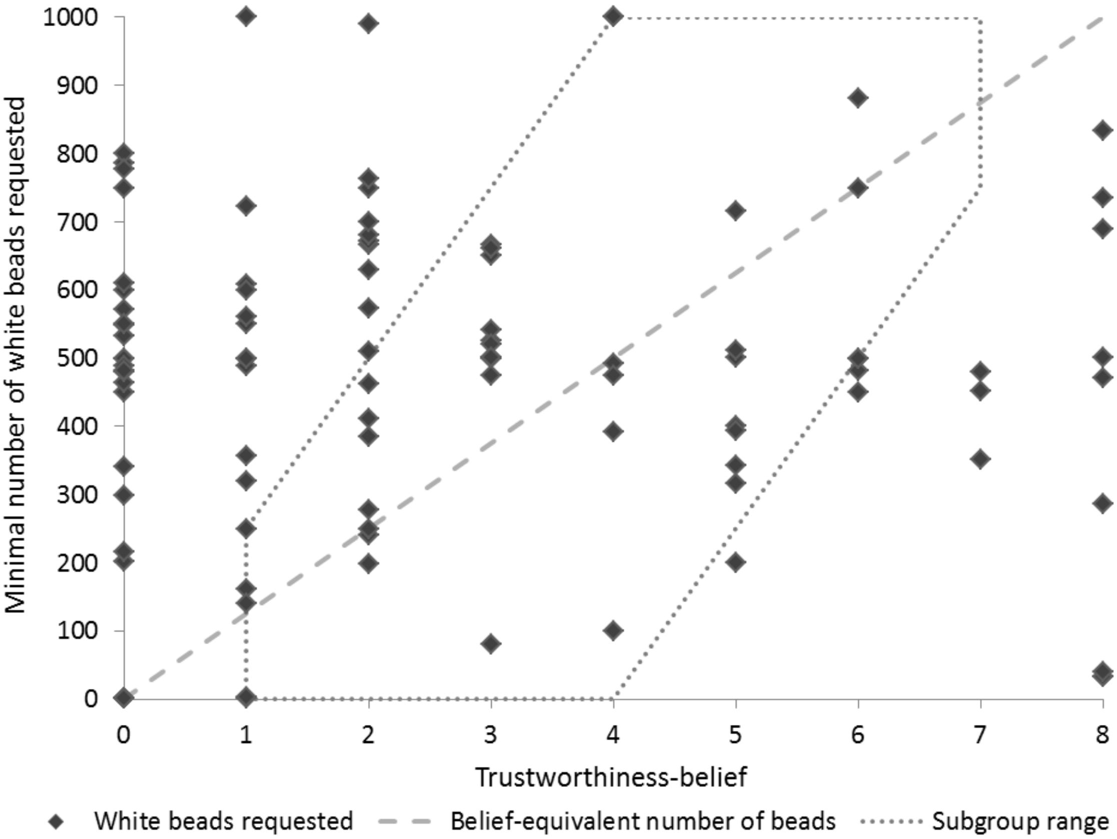


**Figure 6.2** The minimal number of white beads requested graphed against the trustworthiness-beliefs (i.e., the number of trustees expected to pick the good outcome). The dashed line indicates the belief-equivalent numbers of white beads. The deviations analysed are the distances from each number of beads requested (diamonds) to the dashed line. The dotted line indicates the range of data included in the subgroup analyses (n=35).

Assumptions of linearity, of absence of multicollinearity, of normality of the residuals, and of homoscedasticity were checked, both for total beads requested and for the deviations from belief-equivalent numbers of beads. There was no clear non-linear trend between trustworthiness-beliefs and either the total beads requested or the deviations from belief-equivalent numbers of beads; hence the assumption of linearity was not violated. Absence of multicollinearity was confirmed by the facts that none of the predictors (beliefs, risk-aversion, paranoia, and reciprocity norms) were very strongly correlated, the tolerance values were >.866, and the VIF values were all <1.155. The standardised residuals were normally distributed. Homoscedasticity was confirmed as the plots of standardised residuals and predicted scores showed that the variance was equal across the range of the predicted scores. Less than 5% of the participants had standardised residuals >|2|, which is an acceptable level for regression analyses (Field, 2013). As none of the assumptions were violated, regression analyses were conducted using the data from the full sample (n=104) and the subgroup (n=35).

The median trustworthiness-belief of the full sample was 2 trustworthy players; the median belief of the subgroup was 3 trustworthy players. Participants with the median belief in the respective analysis-group were excluded, leaving n=87 (43 pessimists, 44 optimists) for the full sample and n=24 (11 pessimists, 13 optimists) for the subgroup.

For the group comparisons, normality of the dependent variables was checked for pessimists and optimists, separately. All distributions were normal in the subgroup (*p*s>.115). The distributions of deviations in the full sample were normal (*p*s>.106), but the distributions of the total beads requested showed a slight positive skew and Kolmogorov-Smirnov tests indicated that they were non-normal (both *p*s=.006). Transformations (e.g. square-root transformations) did not correct this (Kolmogorov-Smirnov tests: *p*s<.001). Given the robustness of ANOVAs and ANCOVAs to non-extreme deviations from normality (Field, 2013), analyses were continued with untransformed variables. Non-parametric Mann-Whitney tests were also conducted, with results reported in footnotes. However, there is no non-parametric equivalent of an ANCOVA and hence it was not possible to check for robustness after accounting for potentially influential factors in non-parametric analyses.

### 6.3.2 Descriptive Statistics

Table 6.1 and Table 6.2 present the descriptive statistics for the study. Overall, senders were quite accurate in their trustworthiness-beliefs, as the median sender expectation was that 2 trustees (25% of the trustees in a session) would pick the good outcome, when 27.9% of the trustees picked the good outcome in reality. However, due to variation in the number of trustees who picked the good outcome across the different sessions (ranging from 1 to 8), only 9 senders were rewarded for their trustworthiness-belief, which was accurate within their specific session.

**Table 6.1** Descriptive statistics for the continuous variables.

| | Mean | Median | SD | Range |
|---|---|---|---|---|
| **Trustworthiness-belief** | 2.68 | 2 | 2.57 | 0-8 |
| **Risk-aversion** | 7.22 | 7 | 2.08 | 0-11 |
| **Reciprocity norms** | 26.47 | 26 | 5.09 | 6-39 |
| **Paranoia** | 18.12 | 18 | 6.56 | 4-39 |

**Table 6.2** Descriptive statistics for the categorical variables.

| | N (%) |
|---|---|
| **Trustees who picked the good outcome** | 29 (27.9%) |
| **Senders who had the exactly correct trustworthiness-belief** | 9 (8.7%) |
| **Confidence that trustworthiness-belief was close to correct answer** | |
| Not at all confident | 10 (9.6%) |
| Somewhat confident | 39 (37.5%) |
| Quite confident | 37 (35.6%) |
| Very confident | 18 (17.3%) |
| **Guess what would be more likely if belief was wrong** | |
| More trustworthy players than thought | 43 (41.3%) |
| Fewer trustworthy players than thought | 47 (45.2%) |
| Equally likely for there to be more or fewer trustworthy players than thought | 14 (13.5%) |

### 6.3.3 Betrayal Aversion

In the overall sample, a higher minimal number of white beads was requested than prescribed by beliefs (mean (SE) deviation from belief-equivalent numbers of beads = 129.06 (38.454); *t*(103)=3.356, *p*=.001, *d*=.661, 95%-CI [52.794, 205.321]). This means that participants, on average, were more willing to play against another player than against the computer compared to what would be expected on the basis of their trustworthiness-beliefs. In the subgroup, the requested minimal number of white beads was not significantly different from that prescribed by beliefs (8.25 (29.541); *t*(103)=.279, *p*=.782, *d*=.055, 95%-CI [-51.721, 68.221]). As such, betrayal aversion was not found at a general level.

### 6.3.4 Regression Analyses

The results of a linear regression using the whole sample showed that trustworthiness-beliefs could not significantly predict the minimal number of white beads requested (*F*(1,102)=.375, *p*=.542, $R^2_{\text{ADJUSTED}}$=-.006; see Table 6.3, model 1).

However, a linear regression using the subgroup showed that trustworthiness-beliefs significantly predicted the total number of white beads requested (*F*(1,34)=18.215, *p*<.001, $R^2_{\text{ADJUSTED}}$=.330; see Table 6.3, model 2). This was a robust association as a hierarchical linear regression indicated that adding beliefs as a fourth predictor significantly improved a model with risk-aversion, reciprocity norms and paranoia (Δ*F*(1,31)=12.085, *p*=.002, Δ$R^2$=.241; *F*(4,31)=4.796, *p*=.004, $R^2_{\text{ADJUSTED}}$=.303; see Table 6.3, model 3). This robust association within the subgroup is not surprising, as only participants whose number of requested white beads fell (widely) along the diagonal imposed by the belief-equivalent numbers of beads were included.

**Table 6.3** B-values, 95%-confidence intervals (95%-CI) of the b-values, standard errors (SE), β-values, and *p*-values for each of the predictors in the linear regression models for the minimal number of white beads requested. Model 1 for the full sample; models 2 and 3, for the subgroup.

| **Model** | **Predictor** | **b** | **95%-CI of b** | **SE** | **β** | ***p*** |
|---|---|---|---|---|---|---|
| **1** | Trustworthiness-beliefs | 5.795 | [-12.975, 24.565] | 9.463 | .061 | .542 |
| **2** | Trustworthiness-beliefs | 86.876 | [45.509, 128.244] | 20.355 | .591 | <.001 |
| **3.1** | Risk-aversion | 26.408 | [-14.271, 67.087] | 19.971 | .224 | .195 |
| | Reciprocity | -10.538 | [-24.367, 3.292] | 6.789 | -.263 | .130 |
| | Paranoia | 0.394 | [-10.589, 11.376] | 5.392 | .012 | .942 |
| **3.2** | Risk-aversion | 20.866 | [-14.385, 56.118] | 17.284 | .177 | .236 |
| | Reciprocity | 1.326 | [-12.489, 15.141] | 6.774 | .033 | .846 |
| | Paranoia | -1.217 | [-10.741, 8.307] | 4.670 | -.038 | .796 |
| | Trustworthiness-beliefs | 84.795 | [35.048, 134.543] | 24.392 | .576 | .002 |

Another set of linear regressions investigated whether beliefs could (robustly) predict how much the requested number of white beads deviated from belief-equivalent numbers.

Results for the full sample indicated that trustworthiness-beliefs could significantly predict deviations from the belief-equivalent number of white beads ($F(1,102)=158.679$, $p<.001$, $R^2_{ADJUSTED}=.605$; see Table 6.4, model 1). The association between beliefs and deviations was robust, because adding beliefs as a predictor in addition to risk-aversion, reciprocity norms, and paranoia significantly improved the model ($\Delta F(1,99)=154.246$, $p<.001$, $\Delta R^2=.600$; $F(4,99)=39.470$, $p<.001$, $R^2_{ADJUSTED}=.615$; see Table 6.4, model 2). Note that the b-values of beliefs are close to -125, and their 95%-CIs include -125. This is the expected value of deviations based on their calculation, if trustworthiness-beliefs and the number of white beads requested were statistically independent[17]. Statistical independence between the number of white beads requested and trustworthiness-beliefs was also suggested in the analyses above.

**Table 6.4** B-values, 95%-confidence intervals (95%-CI) of the b-values, standard errors (SE), β-values, and *p*-values for each of the predictors in the linear regression models for the deviation from the belief-equivalent number of white beads. Models 1 and 2 for the full sample; models 3 and 4 for the subgroup.

| Model | Predictor | b | 95%-CI of b | SE | β | *p* |
|---|---|---|---|---|---|---|
| **1** | Trustworthiness-beliefs | -119.205 | [-137.975, -100.435] | 9.463 | -.780 | <.001 |
| **2.1** | Risk-aversion | 6.054 | [-31.642, 43.749] | 19.000 | .032 | .751 |
| | Reciprocity | 5.403 | [-10.885, 21.691] | 8.210 | .070 | .512 |
| | Paranoia | -7.073 | [-19.542, 5.395] | 6.285 | -.118 | .263 |
| **2.2** | Risk-aversion | 4.917 | [-28.673, 18.838] | 11.972 | -.026 | .682 |
| | Reciprocity | 3.310 | [-6.932, 13.552] | 5.162 | .043 | .523 |
| | Paranoia | -4.333 | [-12.181, 3.516] | 3.955 | -.072 | .276 |
| | Trustworthiness-beliefs | -118.903 | [-137.899, -99.906] | 9.574 | -.778 | <.001 |
| **3** | Trustworthiness-beliefs | -38.124 | [-79.491, 3.244] | 20.355 | -.306 | .070 |
| **4.1** | Risk-aversion | 18.239 | [-17.746, 54.224] | 17.666 | .183 | .310 |
| | Reciprocity | 6.951 | [-5.283, 19.185] | 6.006 | .205 | .256 |
| | Paranoia | -1.981 | [-11.697, 7.734] | 4.770 | -.073 | .681 |
| **4.2** | Risk-aversion | 20.866 | [-14.385, 56.118] | 17.284 | .209 | .236 |
| | Reciprocity | 1.326 | [-12.489, 15.141] | 6.774 | .039 | .846 |
| | Paranoia | -1.217 | [-10.741, 8.307] | 4.670 | -.045 | .796 |
| | Trustworthiness-beliefs | -40.205 | [-89.952, 9.543] | 24.392 | -.323 | .109 |

For the subgroup, trustworthiness-beliefs showed a trend towards significantly predicting deviations from belief-equivalent numbers of beads (*F*(1,34)=3.508,

[17] If trustworthiness-beliefs and the number of white beads requested are statistically independent, the number of white beads requested is a random draw between 0 and 1000, with an expected average of 500. If trustworthiness-beliefs are 0/8, the belief-equivalent number of white beads is 0 and so deviations are between 0 and 1000 (expected average 500); if trustworthiness-beliefs are 1/8, the belief-equivalent number of white beads is 125, and deviations are between -125 and 875 (expected average of 375); and so on. Hence, with a randomly selected number of white beads requested, the deviations should decrease in steps of -125 with each increasing step of trustworthiness-beliefs.

$p$=.070, $R^2_{ADJUSTED}$=.067; see Table 6.4, model 3). In this subgroup, however, this association was not robust, as it did not significantly improve a predictive model accounting for risk-aversion, reciprocity norms, and paranoia ($\Delta$F(1,31)=2.717, $p$=.109, $\Delta R^2$=.075; $F$(4,31)=1.262, $p$=.306, $R^2_{ADJUSTED}$=.029; see Table 6.4, model 4). Note that this may be due to a lack of power with analyses based on a sample of n=35.

### 6.3.5 Factorial Analyses

In the full sample, there were no significant differences between pessimists (mean (SE) = 411.953 (37.705)) and optimists (486.955 (37.274)) in the minimal number of white beads they requested ($F$(1,85)=2.001, $p$=.161, $\eta_p^2$=.023, 95%-CI [-30.414, 180.416]). They also did not differ when risk-aversion, reciprocity norms, and paranoia were accounted for in an ANCOVA ($F$(1,85)=1.970, $p$=.164, $\eta_p^2$=.023, 95%-CI [-31.548, 182.685]). However, within the subgroup, there was a significant difference ($F$(1,22)=11.382, $p$=.003, $\eta_p^2$=.341, 95%-CI [97.174, 407.259]). As expected, participants with pessimistic beliefs (230.091 (55.022)) requested fewer white beads than participants with optimistic beliefs (482.308 (50.613)). This difference was reduced to marginally significant when accounting for risk-aversion, reciprocity norms, and paranoia ($F$(1,19)=3.922, $p$=.062, $\eta_p^2$=.171, 95%-CI [-10.049, 363.433]).

When assessing deviations from belief-equivalent numbers of beads, there was a difference between pessimists and optimists, both in the full sample and in the subgroup.

In the full sample ($F$(1,85)=61.267, $p$<.001, $\eta_p^2$=.419, 95%-CI [-663.444, -394.667]), pessimists requested significantly more white beads than their belief-equivalent numbers would suggest (365.422 (46.086); $t$(42)=8.906, $p$<.001, $d$=2.748, 95%-CI

[282.629, 448.255]). Optimists requested significantly fewer white beads than they should based on their trustworthiness-beliefs (-163.614 (47.518); *t*(43)= -3.061, *p*=.004, *d*=.934, 95%-CI [-271.391, -55.836]). This difference was robust as it was also found after accounting for risk-aversion, reciprocity norms, and paranoia (*F*(1,82)=58.777, *p*<.001, $\eta_p^2$=.418, 95%-CI [-663.097, -389.875]).

This effect of trustworthiness-beliefs on deviations from belief-equivalent numbers of beads was also robustly found in the subgroup, where pessimists and optimists differed significantly (*F*(1,22)=6.879, *p*=.016, $\eta_p^2$=.238, 95%-CI [-266.830, -31.184]). Pessimists requested slightly more beads than their belief-equivalent number of beads would suggest (25.545 (41.813)), although a one-sample *t*-test indicated this was not significantly different from the belief-equivalent number (*t*(10)=.784, *p*=.451, *d*=.496, 95%-CI [-47.077, 98.168). In contrast, optimists requested significantly fewer white beads than they should based on their trustworthiness-beliefs (-123.462 (38.463); *t*(12)=-2.787, *p*=.016, *d*=1.609, 95%-CI [-219.998, -26.925]). This difference was also found after accounting for risk-aversion, reciprocity norms, and paranoia (*F*(1,19)=6.769, *p*=.018, $\eta_p^2$=.263, 95%-CI [-331.973, -35.970]).[18]

## 6.4 Discussion

The first aim of the study presented here was to seek evidence of betrayal aversion in a paradigm designed to be free of methodological confounds. Our general sample was not found to be more willing to take the risk of a bad

[18] Non-parametric Mann-Whitney tests converged on the same conclusions: optimists and pessimists did not differ in their total number of beads requested in the total sample (*U*=827.0, *p*=.315, *r*=-.108), but they did in the subgroup (*U*=21.0, *p*=.002, *r*=-.597). Moreover, optimists and pessimists differed in their deviations from belief-equivalent numbers of beads in the total sample (*U*=222.5, *p*<.001, *r*=-.688) and in the subgroup (*U*=31.5, *p*=.019, *r*=-.473).

outcome when determined by a computer than when determined by another player, which would be a classic indication of betrayal aversion. However, these results changed drastically, as discussed below, when using beliefs regarding others' trustworthiness to predict the levels of betrayal aversion. Including trustworthiness-beliefs in the analysis, to test a novel theoretical explanation of betrayal aversion, was the second aim of this study.

Our hypothesis was that beliefs about others' trustworthiness would predict how betrayal averse participants would be. This hypothesis was tested in two steps. As a first step, the hypothesis would entail a positive correlation between trustworthiness-beliefs and number of white beads requested. More optimistic participants (i.e., those who believe more trustees are trustworthy) would require a higher probability of a good outcome from the computer's draw to prefer having the computer draw an outcome compared to having another, believed-to-be-trustworthy trustee draw an outcome.

### 6.4.1 Trustworthiness-Beliefs and Numbers of White Beads Requested

For the whole sample, the beliefs about others' trustworthiness did not predict the minimal number of white beads senders requested. For the subgroup, the trustworthiness-beliefs did predict the number of white beads requested, but this might be related to inclusion criteria for this subgroup: data points had to be (widely) along the line imposed by belief-equivalent numbers (see Figure 6.2). A lack of a correlation between the trustworthiness-beliefs and the total white beads requested in the total sample could be explained by a number of factors.

First, our assumption that behaviour in economic games should be in line with beliefs might have been wrong. Fetchenhauer and Dunning (2009) had already hinted at this when they found that participants were sceptical about others' trustworthiness in the trust game, yet opted into the trust game. Furthermore, a disconnect between beliefs and behaviour was recently shown in a signal-sender-receiver game (Sheremeta & Shields, 2013). In this game, there are two states of the world: A or B. One player, the signal-receiver, is endowed an initial 10 points that can be invested, which leads to 18 points in state A, but to 0 points in state B. The other player, the signal-sender, receives 13 points if the signal-receiver invests, but 0 points if the signal-receiver does not invest. The signal-sender, who knows the true state of the world, sends the signal-receiver a signal regarding the state of the world; this can be honest or deceptive. It would be in the signal-sender's interest to signal state A, regardless of the true state of the world, as this would lead signal-receivers to invest, which, in turn, leads to a reward for the signal-sender. Therefore, in this game, a signal A could represent either a true state A (honest signal) or a true state B (deceptive signal). Signal-receivers are aware of this, as Sheremeta and Shields (2013) found that a majority of signal-receivers believed that signal-senders might send a deceptive signal A. This belief was incentivised for correctness and so was expected to reflect a true belief. From a rational point of view, if signal-receivers distrust the authenticity of signal A, they should ignore this signal. Instead, they should use the probabilities of the states of the world (here: $p(\text{A})=p(\text{B})=.5$) to calculate their expected payoff to make the rational decision not to invest. Yet, 67% of the participants receiving signal A decided to invest (Sheremeta & Shields, 2013). This suggests that participants often do not act in accordance with their incentivised, stated beliefs in economic games. As such, this could be a reason for the lack of a correlation between beliefs and (rational) behaviour. However, in Fetchenhauer and Dunning's (2009) and Sheremeta and Schields's (2013) studies, participants may have opted in or invested out of altruism. If they had not invested, which is what their beliefs dictated they should do in order to maximise their own expected payoff, the other player would not have received any payoff. Anticipating this, we did not include an option to opt out, so trustees would always receive a reward. In fact, in our study, outcomes would be equal for the senders and trustees, or unequal in the trustees' favour, which would have led to reduced willingness to trust the trustee, if anything.

Second, although a pilot study gave no indication of miscomprehension and any clarification questions participants asked during the experiment proper pertained to the risk-aversion measure, rather than to the betrayal-aversion task, it is possible that participants did not fully comprehend the task. In particular, participants may not have understood that if their minimal number of white beads requested was not met, the other player's decision would be implemented. Such miscomprehension could account for the participants who, despite having pessimistic beliefs about the trustworthiness of others, still required that the urn contained at least 900 white beads, which would occur with a probability of only 10%. Therefore, they took a 90% chance that their outcome would be determined by another player, who they believed was more likely to be untrustworthy than to be trustworthy. Against this notion of miscomprehension, we did include detailed instructions and comprehension checks, which presumably would have avoided miscomprehension. It is, however, still possible that participants kept selecting answers for the multiple-choice questions until they selected the correct answer (as they could not continue to the task otherwise). The correct answer would then not be based on full understanding of the instructions. We did not include open-ended questions, such as “If there were 3 players (out of 8) in role Y choosing outcome A (15/15), how many white beads would the container (with 1000 beads) have to hold so that the probability of receiving outcome A is the same in both situations?” or “If I wanted 1000 beads representing outcome A in the container, what is the probability that the player in role Y determines the outcome?”. We avoided such questions because the use of the term “probability” may have led to a suggestive mechanism to match trustworthiness-beliefs and numbers of white beads requested, potentially creating demand effects. Future research could incentivise (additional) comprehension questions so that each incorrect answer leads to a penalty taken from the show-up fee.

If miscomprehension did occur, this may have been due to the abstract language used in the experiment. As noted above, in order to avoid demand effects, suggestive terms such as “trustee”, “good outcome”, and “probability” were avoided. Instead, abstract terms (e.g., “Player Y” and “Outcome A”) were used. A suggestion that people struggle arriving at logical decisions when dealing with abstract problems, but not when dealing with more contextualised problems, comes from Wason’s Selection Task (Wason, 1968). In the abstract form of this task, participants are presented with four cards, all of which have a letter on one side and a number on the other side. Participants see the letters (P and Q) of two cards, and the numbers (1 and 2) of the other cards. Participants are then asked which card or which cards they would need to turn over to see if the rule “If P, then 1” is true, without checking unnecessary cards. Many participants do not make the correct selection (i.e., the card that says P, to check if it has 1 on the other side, *and* the card that says 2, to check that it does *not* have P on the other side). Yet, when framed in more concrete, social contexts (in particular, a context of cheater detection), participants perform well on this task

(Cosmides, 1989). For example, if the cards show the age of people (minor or not) on one side, and the type of beverage (alcoholic or not) on the other, participants check both the card of the minor and the card with an alcoholic beverage when asked to check if there is no underage drinking (Dudley & Over, 2003). Perhaps the abstract terms used in the present study led to illogical responses from participants.

Notwithstanding the above considerations, we doubt that miscomprehension was responsible for the lack of an association between trustworthiness-beliefs and numbers of white beads requested. Instead, our hypothesised mechanism explaining betrayal aversion may well have masked the correlation between beads requested and beliefs. This relates to the second step of our hypothesis, which focused on deviations from belief-equivalent numbers of white beads.

### 6.4.2 Trustworthiness-Beliefs and Deviations from Belief-Equivalent Numbers of White Beads Requested

Our hypothesis suggests that, besides the lower monetary payoff from receiving the bad outcome (and disadvantageous inequality of payoffs), being betrayed would lead to additional disutility. This additional disutility would stem from knowing the untrustworthy trustee received more money from their unfair selection than if the trustee had selected the good outcome (i.e., the trustee was rewarded for being unkind), and its level would depend on trustworthiness-beliefs. In particular, the additional disutility would be higher for optimists than for pessimists, because the betrayal is unexpected, leading to a negative prediction error, for the former. As such, participants were expected to deviate from the number of white beads that their trustworthiness-beliefs would prescribe, so that the higher the beliefs about trustworthiness, the more negative this deviation would be. This would attenuate the assumed underlying positive correlation between beliefs and number of beads requested in the first step of the test of our hypothesis.

We found evidence for this negative association between trustworthiness-beliefs and deviations from the belief-equivalent number of white beads. Through regression analyses, we found a strong negative correlation for the whole sample and a trend-level negative association for the subgroup. The subgroup provided a more robust test of hypotheses regarding deviations, as only participants who could deviate from their belief-equivalent number of beads equally in both directions were included. This avoided a potential correlation being driven by statistical independence between trustworthiness-beliefs and number of white beads requested, which were suggested to be independent, as described earlier for the first step of our hypothesis testing. However, the inclusion criteria for the subgroup led to a small sample size, which may have made regression analyses underpowered. In group comparisons, which might be more appropriate than the regression given the small sample size (Field, 2013), a robust effect of trustworthiness-beliefs was found. Pessimists requested slightly more beads than their beliefs dictated, while optimists requested fewer beads than their beliefs dictated. In other words, optimists showed betrayal aversion and were willing to accept a higher risk of the bad outcome in the lottery than what their trustworthiness-beliefs indicated they thought the risk would have been in the trust game. Pessimists, on the other hand, were more willing to play the trust game than their trustworthiness-beliefs suggested.

This provides tentative support for the notion that betrayal aversion might be associated with the disutility stemming from the untrustworthy person being rewarded. Furthermore, this additional disutility factors into the sender's utility function above and beyond the disutility from the actual outcome received and depends on a sender’s expectations about the trustee’s behaviour. As such, our hypothesised mechanism of trustworthiness-beliefs affecting betrayal aversion might cloud the association between trustworthiness-beliefs and the number of white beads requested.

The association between levels of betrayal aversion and trustworthiness-beliefs may influence whether betrayal aversion is found or not. Fetchenhauer and Dunning’s (2012) sample may have included many participants with relatively pessimistic trustworthiness-beliefs. These participants may have considered the trust game an opportunity for a pleasant surprise by other’s kind actions, and therefore would have been willing to opt into the trust game. Furthermore, at the low probability of receiving the bad outcome (i.e., a *manipulated* low trustworthiness-belief), receiving the bad outcome after having opted into the trust game would be as expected, but receiving the good outcome would lead to additional utility as this outcome would be better than expected.

In closing, it must be noted that this study can only point to an association between betrayal aversion and beliefs. In order to investigate whether our found association is causal, beliefs would have to be *manipulated*. This could be done in a future study where senders are informed they will be placed in a group with four other players, one of whom will be their assigned trustee, and they are informed about how many of the four players are trustworthy. Senders then have to provide their minimal acceptable probability of the good outcome at which they would switch from the trust game to a computerised lottery. This could be done using a strategy method (N. D. Johnson & Mislin, 2011), where senders provide answers for several possible group compositions, ranging from all four trustees being trustworthy to all four being untrustworthy. By informing senders how many trustees are trustworthy, one can manipulate the beliefs. Then, participants' minimal acceptable probabilities of a good outcome in the lottery might show differences within the uncertain groups (i.e., 1, 2, or 3 out of 4 trustees are trustworthy) that are above the trust game's equivalent probability of a good outcome in the relatively untrustworthy group, but below it in the relatively trustworthy group. For example, in groups with one trustworthy trustee the probability of a good outcome in the trust game is .25. The minimal acceptable probability of the good outcome in the lottery in untrustworthy groups might be higher than .25 (e.g., .33). In contrast, in groups with three trustworthy trustees, the probability of a good outcome in the trust game is .75. The minimal acceptable probability of the good outcome in the lottery in trustworthy groups might be lower than .75 (e.g., 64). Such results, if found, would provide further support for the theory presented in this chapter.

## 6.5 Conclusion

The present study shows that beliefs about others' trustworthiness are associated with the level of betrayal aversion people display. People with pessimistic beliefs about the trustworthiness of others show a slightly stronger preference for the trust game than their beliefs would dictate. We suggest this is because they stand nothing to lose by trusting another: they either find out that others are indeed not trustworthy or get pleasantly surprised that others are more trustworthy than expected. As such, they may as well play the trust game or request a high probability to win the good outcome in a computerised lottery. People with optimistic beliefs about others' trustworthiness, however, show a stronger preference for the computerised lottery than their beliefs would dictate. They can either have their beliefs confirmed or be *un*pleasantly surprised by others' lack of trustworthiness, and might thus prefer not to have their beliefs tested. Instead they prefer to play a computerised lottery, even when it has a lower probability of the good outcome than their beliefs would indicate the probability of that outcome to be in the trust game.

# 7 Discussion and Conclusion

In this thesis I have investigated a range of “irrational beliefs”, defining irrational beliefs as deviations from the probability that one *should* assign to given propositions. Irrationality is critically important on a personal and on a social scale as it can carry large costs. Biased beliefs about the self and about the future might lead to smoking, unsafe sex, financial recessions, or even wars (D. D. P. Johnson & Fowler, 2011; Sharot, 2011a). Men’s biased beliefs about women’s sexual interest may lead to unwanted sexual advances or even sexual assault (Farris et al., 2008b). Less common, but extreme irrational beliefs, such as delusions, affect society due to loss of functioning and high mental health care costs (Coltheart et al., 2011).

The vast number of findings where people systematically provide answers that do not follow formal logic rules, has led some to argue that humans are simply not rational (Kahneman, 2003; Tversky & Kahneman, 1983, 1986). Defenders of bounded rationality consider this notion too harsh and emphasise that ecological influence can lead to decisions deemed irrational when compared to normative rules, which assume unlimited time and cognitive resources and the application of the correct logic rules (Cosmides & Tooby, 1996; Gigerenzer & Goldstein, 1996; Gigerenzer & Sturm, 2012; Stanovich & West, 2000). Furthermore, often biased beliefs are inferred from biased behaviour (e.g., Haselton & Buss, 2000), an inference some (e.g., Bortolotti, 2009) might consider valid as they claim that rational beliefs should have matching behaviour. However, others (e.g., Marshall et al., 2013; McKay & Dennett, 2009) argue that biased behaviours can occur without biased beliefs and vice versa. Therefore, the inference of a biased belief from a biased behaviour might not be valid.

In this thesis, I mainly focused on biased beliefs, rather than on biased behaviours. In each case I introduced novel elements to the investigation of the respective biases, so as to clarify or resolve a series of compelling theoretical issues. In addition, I took great care to minimise or eliminate relevant confounds, as imperfect methodologies have been at the heart of the irrationality debate (Stanovich & West, 2000). The studies reported in Chapters 2 and 3 investigated the data-gathering and probability-reasoning components of the jumping-to-conclusions (JTC) bias associated with delusion-proneness, using newly incentivised versions of the beads-task paradigm. The study reported in Chapter 4 investigated the sexual over-perception bias in a belief-updating paradigm, so as to assess whether this phenomenon reflects a bias in the integration of relevant information and is not fully attributable to different socialisation of men and women. The study reported in Chapter 5 investigated the "real" self-deception account in the domain of the optimism bias, to assess whether people hold both a realistic and a more desirable representation of their future prospects simultaneously. The study reported in Chapter 6 investigated whether beliefs about others' trustworthiness could influence the level of betrayal aversion participants displayed in the trust game.

In this final chapter, I first summarise the findings of these empirical studies and consider the theoretical implications of these findings for each specific psychological phenomenon. Then, I discuss overall limitations of the research, as limitations specific to the researched biases have already been discussed in the empirical Chapters 2 to 6. Next, I consider the implications of the findings at a more general level of rationality research. Throughout, I provide future suggestions. Finally, I conclude the thesis.

## 7.1 Summary and Theoretical Implications of the Specific Biases

### 7.1.1 Jumping-to-Conclusions Bias

That delusional and delusion-prone individuals "jump to conclusions" (JTC) on probabilistic reasoning tasks is perhaps the most important and influential claim in the entire literature on cognitive theories of delusions. However, previous investigations of this bias have suffered from conceptual and methodological limitations. First, although the notion of "jumping to conclusions" implies that delusion-prone individuals gather insufficient evidence and reach premature decisions, no previous study has actually investigated whether the evidence gathering of such individuals is, in fact, suboptimal. The standard "jumping to conclusions" effect is a *relative* effect, but using relative comparisons to substantiate absolute claims is problematic, as non-delusion-prone individuals could potentially gather objectively too *much* data and decide too *late*. Second, although some studies have varied the likelihood ratios of relevant evidence, no previous investigation has examined the effect of varying the prior distribution over relevant states. Third, many previous investigations have been vulnerable to a range of confounds, including effects of miscomprehension, working-memory and motivation.

The studies reported in Chapters 2 and 3 were designed to minimise the influence of these factors to assess whether the JTC bias is a genuine cognitive bias, rather than a methodological artefact. In both studies, we used an adapted version of the beads-task paradigm, in which a fisherman presented fish from either a mostly-white lake or a mostly-black lake. In Chapter 2, financial incentives were utilised to generate optimal decision points in two data-gathering versions of this paradigm and thus move beyond the standard, relative finding. In Chapter 3, we systematically investigated the effects of incentives and relevant information parameters (i.e., prior distribution, likelihood ratios) in a probability-estimates version of this paradigm.

In the two tasks in Chapter 2, participants' data gathering was compared to the optimal amount of data to gather, as determined by a maximisation of expected payoff, combining a reward for accuracy and a small cost for additional data gathering. In the first, dynamic decision-making task, participants indicated whether they wanted to see more information or decide on a lake, after each fish. In the second, one-shot decision-making task, participants indicated how many fish they wanted to see to base their decision on, before seeing any fish. Across both these tasks, more delusion-prone participants based their decisions on less data than less delusion-prone participants (i.e., a relative JTC bias). Evidence for an absolute JTC bias was found for high-delusion-prone participants in both tasks, who requested less than the optimal amount of data before deciding. Low-delusion-prone participants requested less data than optimal in the dynamic task, but performed optimally in the static task. No association between delusion-proneness and confidence levels at the moment of deciding was found in either of the two tasks.

In Chapter 3, the presence of incentives was directly investigated between subjects and the effect of varying prior probabilities was investigated within subjects in a probability-estimates version of the fish task. In a condition where the likelihoods of the fish being from either lake were equal, but the prior probability of each lake was different, no effects of incentives or of delusion-proneness were found. In a condition where both prior probabilities and likelihoods of the two lakes were different, again, no effects of incentives or of delusion-proneness were found. Interestingly, in the condition where the prior probabilities of each lake were equal, but only the likelihoods were different, as is the case in the standard beads task, an interaction between incentives and delusion-proneness was found. High-delusion-prone participants underestimated the posterior probabilities in this condition, whether incentivised or not. Non-incentivised low-delusion-prone participants underestimated the posterior probabilities as well, but incentivised low-delusion-prone participants provided higher posterior probabilities close to the Bayesian posterior.

The effects of delusion-proneness were more robust in the evidence-gathering versions of the task used in Chapter 2 than in the probability-estimates version used in Chapter 3. This accords with the conclusions of several reviews, which have noted that the JTC bias is more robust on the former measure compared to the latter (Fine et al., 2007; Garety & Freeman, 1999).

The fact that some differences in probability reasoning were found between low-delusion-prone and high-delusion-prone participants is inconsistent with the liberal acceptance account of the JTC bias. This account states that earlier decisions are based on a lower decision-threshold for delusional or delusion-prone participants, while they hold the same subjective probabilities for the hypotheses as controls (e.g., Moritz et al., 2007). Furthermore, the findings from Chapter 2 only partially fit with this account: although we did find earlier decisions for high-delusion-prone participants, which this account predicts, these decisions were not made at lower decision thresholds, as we found similar confidence levels for high-delusion-prone and low-delusion-prone participants.

As described in Chapter 2, the results of our data-gathering JTC study could fit the information-integration account (Menon et al., 2006), where evidence is assigned extra weight due to dysregulated dopamine mechanisms (Kapur, 2003). As a result, decisions would be made early and hypotheses would be held with more confidence. The data-gathering results of the study reported in Chapter 2 were consistent with this account, but the results regarding confidence levels were somewhat ambiguous. As discussed earlier, we analysed confidence levels at deciding and found no difference between low-delusion-prone and high-delusion-prone groups, but perhaps both groups make their decision at the same confidence level and high-delusion-prone groups arrive at this level faster. Due to power constraints, this could not be analysed. Also ambiguous were the probability-estimates results from the study reported in Chapter 3. In contrast to the information-integration account's predictions, we found probability estimates provided by high-delusion-prone participants to be either similar to or lower than those provided by low-delusion-prone participants. One reason why the predicted direction (i.e., higher probability estimates from high-delusion-prone participants) was not found could be that we presented only one fish, and as such, subsequent information did not need to be integrated with beliefs based on previously-acquired information. Alternatively, as the JTC bias is less robust in probability-estimates versions of the reasoning paradigm (Fine et al., 2007; Garety & Freeman, 1999), a significant difference might only be found when comparing clinical and non-clinical groups, as the probability reasoning of our high-delusion-prone participants may not be affected to the extent that it is consistently different from that of low-delusion-prone participants.

An alternative account might explain our findings better. Dudley and Over (2003) argue that the JTC bias might result from delusional or delusion-prone participants having a confirmation bias combined with an overactive threat-detection system which extends to neutral material. With such a combination, instances that confirm a (perceived-to-be-threatening) hypothesis are sought out. For example, if someone is laughing in the vicinity of a delusional person, he might hypothesise the laughter is an indication of a plot against him. To maximise survival, it might be wise to focus on instances that confirm the hypothesis of a developing plot, and to take appropriate actions, rather than to search for instances which disconfirm this (e.g., evidence that someone was merely laughing at a joke). Doing the latter would result in an inefficient use of time in case people are plotting against him. Green, Freeman, and Kuipers (2011) provide support for this type of confirmatory biased reasoning in a non-clinical sample as they found that individuals higher in paranoid ideation tended to give more paranoid explanations of laughter they overheard between the experimenter and a confederate. Balzan, Delfabbro, Galletly, and Woodward (2013) also found that schizophrenia participants, and delusion-prone participants to a lesser degree, have a preference for a positive-test strategy (i.e., a test that will provide evidence that matches the hypothesis), even if the positive test is non-diagnostic (i.e., that the evidence also matches other hypotheses) and diagnostic negative tests (i.e., tests that disprove the hypothesis) are available. In the (adapted) beads task, such a confirmation bias might lead delusional or delusion-prone participants to accept hypotheses early "to be on the safe side", without necessarily believing that the hypothesis is more probable than controls believe it to be.

It might be difficult to see how confirmation of threat would apply in a neutral beads (or fish) task. However, as Dudley and Over (2003) state, threat might be detected even in neutral material due to an overactive threat-detection system, which, in turn, might result from high levels of arousal and anxiety (Salvatore et al., 2012). Evidence for an association between psychotic symptoms and attentional biases to threat has been mixed (Tone & Davis, 2012), where some (e.g., Colbert, Peters, & Garety, 2010) find a bias *away* from threat, and others finds a bias *towards* threat detection (e.g., Dudley et al., 2014).

Perhaps rather than constituting a threat in itself, a fish of a different colour than that seen so far might induce anxiety concerning the difficulty of integrating disconfirming evidence into a belief (and, as a result, arriving at the wrong conclusion and missing out on a reward). Potentially due to poorer working-memory functioning, delusional or delusion-prone participants might have difficulty accounting for disconfirming evidence, and either discard it, leading to a bias against disconfirmatory evidence (e.g., Speechley et al., 2012; Woodward, Buchy, Moritz, & Liotti, 2007), or overvalue it, leading to *over*-adjustment in probability estimates (e.g., Garety et al., 1991; Speechley et al., 2010), provided the latter is not due to miscomprehension (Balzan, Delfabbro, Galletly, et al., 2012). To avoid the mental complexity of potentially having to integrate disconfirming evidence, delusion-prone individuals might simply stop considering evidence after a few pieces, as the probability of encountering a sequence of solely confirming evidence drops dramatically as more evidence is gathered. This would then result in a low number of draws to decision.

The use of a confirmatory reasoning style in neutral tasks, such as the beads task, might also stem from hyper-salience of evidence-hypothesis matches (Balzan et al., 2013; Speechley et al., 2010). Besides displaying a strategic search for evidence that would confirm the hypothesis, such evidence itself would be more salient and be considered in more detail than evidence that would not match the hypothesis. Speechley et al. (2010) found that schizophrenic participants increased their probability estimates for the lake that matched the evidence more than the clinical and non-clinical control groups. Their probability estimates for the lake that did not match the evidence were the same as those provided by the other groups. Balzan et al. (2013) found that schizophrenic and delusion-prone participants valued and recalled confirmatory evidence better than disconfirmatory evidence. Furthermore, the salience of the initial confirmatory evidence led to reduced adjustment of beliefs in the face of disconfirmatory evidence in these groups compared to the non-delusion-prone group (Balzan et al., 2013). These findings suggest that hyper-salience of evidence-hypothesis matches might underlie the confirmation bias. Hyper-salience of evidence is at the heart of the information-integration account of the JTC bias. As such, the results in this thesis might support a combination of both the confirmation-bias account and the information-integration account of the JTC bias.

Overall, the notion of a confirmatory reasoning style, perhaps combined with hyper-salience of hypothesis-evidence matches affecting information integration, could explain the results of the JTC studies in this thesis. This reasoning style might be the result of an imbalance in the weighting of the costs of errors, with a preference given to detection of a threat that is not there, rather than missing a threat that is there (Dudley & Over, 2003). Such differential weighting of errors is at the heart of error-management theory, which has been applied to the sexual over-perception bias, studied in Chapter 4.

### 7.1.2 Sexual Over-Perception

The study reported in Chapter 4 focused on men's alleged over-perception of women's sexual interest. There are theoretical and methodological limitations to the claim that men over-perceive women's sexual interest. The theoretical limitation stems from the fact that sexual over-perception might be a biased (and potentially adaptive) behaviour, without necessarily requiring biased beliefs. As such, the theoretical underpinning of the sexual over-perception bias as a *cognitive* bias is questionable. The methodological limitation pertains to previous literature's focus on point-estimate paradigms, which do not exclude the possibility of differential socialisation explaining gender differences. In our study, we investigated sexual over-perception as a cognitive bias, and as a novel element, we presented repeated information to avoid potential differences between men and women being due to different prior estimations resulting from different socialisation.

We used a belief-updating paradigm with the beads task as a neutral condition and two adapted bias conditions assessing whether men would overestimate women's sexual interest. Both when estimating how likely women are to be sexually interested in men in general and in a given man, male and female participants provided similar estimates. We found no evidence of a sexual over-perception bias in our belief-updating paradigm as both men and women integrated information similarly into their beliefs. Compared to the rational Bayesian belief-updating norm, both men and women were conservative in their belief-updating, as Phillips and Edwards (1966) have found in the neutral beads task.

One addition to the existing literature (e.g., Abbey, 1982; Haselton & Buss, 2000) is that our stimuli were non-ambiguous (i.e., each piece of information was unambiguously indicative either of sexual interest or disinterest), while previous studies have generally used stimuli which are ambiguous with respect to whether they signal sexual interest or not. Our lack of support for this bias may be due to our use of non-ambiguous stimuli. Buss (2013) and Lindgren et al. (2008) have suggested that the sexual over-perception bias is only found when women display ambiguous behaviours that occur in friendly as well as sexual interactions (e.g., making eye contact), and not when the behaviours are unambiguous (e.g., touching a person's genitals). This then raises the question why a woman would display ambiguous behaviours. One reason might be that she is not yet sure if she is sexually interested, so sends mixed, ambiguous signals, to maintain the possibility to act on potential future sexual interest. Another reason is that she might play "hard to get" (Jonason & Li, 2013), and as such wants to downplay her interest, perhaps to avoid a reputation of being promiscuous. If this playing "hard to get" would be the underlying reason for sending ambiguous signals, it is not unlikely that women would underreport their own sexual interest (Haselton & Buss, 2000). Men may have learned to read "between the lines" of such games over time, and interpret the supposed low sexual interest from ambiguous signals as a concealed genuine sexual interest. These hypothesised processes might underlie the difference between men's perception of women's true intentions and women's biased self-reported intentions. This might lead to a new interpretation of the sexual "over-perception" bias: the difference might not be due to men *over-perceiving* sexual interest, but rather due to women *underreporting* such interest.

This thesis moved from biased beliefs in men about their future sexual prospects to a more general biased belief about future prospects. In the optimism bias, investigated in Chapter 5, future prospects are thought to be better than an objective standard indicates.

### 7.1.3 Self-Deceptive Optimism

The study reported in Chapter 5 investigated possible self-deception underlying the optimism bias. This was accomplished by combining the optimistic belief-updating paradigm (e.g., Sharot et al., 2011) and the crowd-within paradigm (Herzog & Hertwig, 2009; Vul & Pashler, 2008). Participants provided multiple estimated answers to neutral and undesirable questions, with the latter pertaining to participants' chances of experiencing negative events. The crowd-within effect, where the averaged estimate has a lower absolute error than either of the two estimates alone, was found for both neutral and undesirable questions. Signed errors, however, indicated that second estimates for undesirable questions were systematically lower than first estimates for these questions; whereas first and second estimates for neutral questions were equal. These results imply that when providing estimates of their probability of experiencing undesirable future outcomes, participants sampled *selectively* from an internal distribution, producing estimates that were optimistic initially and even more so on a second sampling. Optimism was also found in the absolute sense, where the estimates for undesirable questions underestimated the true value, while the answers to neutral questions were accurate.

These results indicate that people mislead *themselves* by selecting specific estimates of their future prospects, just as they might seek to mislead another person by cherry picking relevant data (e.g., in a job interview, they might showcase select examples of their previous employment performance rather than choosing examples at random). Hence, just as people potentially "protest too much" by exaggerating their prospects or beliefs when they are made to doubt them (Gal & Rucker, 2010; McKay et al., 2011), people might bolster their own beliefs about their prospects through a similar defensive strategy applied intrapersonally.

This might also apply to other putative cases of self-deception, such as the oft-cited better-than-average effect (Alicke, 1985). Perhaps people selectively think of certain examples (e.g., those tests on which they received a particularly good result) to arrive at the conclusion that they are more intelligent than average. This conclusion might then be wrong compared to the true value (e.g., their true IQ score might actually be average). However, with a single estimate, it is not possible to determine if someone samples *selectively* from a desirable end of an internal distribution, or whether they sample randomly from a distribution with a biased mean. Selective sampling, found for unrealistic optimism in this thesis, would support the "real" self-deception account (Gur & Sackeim, 1979; Mijovic-Prelec & Prelec, 2010), as one actively avoids a particular, less rosy, conclusion that they also hold besides their more rosy representation of reality. But besides selective sampling, the mean of the distribution could be biased, perhaps due to biased information processing (Mele, 1997). These two possibilities could even act in tandem (i.e., sampling selectively from a biased distribution). In order to investigate whether people sample selectively, repeated sampling is necessary; single estimates cannot distinguish between the different possibilities.

Overall, people become even more optimistic in response to tests of their optimistic beliefs. An unwillingness to have one's optimistic beliefs about future outcomes tested could extend to an unwillingness to have other optimistic beliefs tested. In Chapter 6, I explored whether people with more optimistic beliefs about others' trustworthiness were more betrayal averse, and as such less willing to have their beliefs tested in a trust game.

### 7.1.4 Betrayal Aversion

The study reported in Chapter 6 explored betrayal aversion in the trust game. Participants indicated their minimal acceptable probability of a computerised lottery selecting the better of two outcomes, at which level they would prefer to play that lottery than have another person select the outcome. This study avoided the choice between opting in or out, so as to avoid confounds of signalling distrust or efficiency maximisation. Furthermore, strongly incentivised beliefs about others' trustworthiness were elicited to determine the point at which participants should be indifferent between having the computer or another player select the outcome, in the absence of betrayal aversion. Importantly, these beliefs were then used to test a novel theory that a person's trustworthiness-belief could predict their level of betrayal aversion. Deviations from belief-equivalent points indicated that betrayal aversion was present and its strength related to trustworthiness-beliefs. People who believed that only a few others would be trustworthy (i.e., pessimistic trustworthiness-beliefs) set the minimal probability of the good outcome in the lottery at a higher level than their beliefs would suggest. Conversely, people who believed many others were trustworthy (i.e., optimistic trustworthiness-beliefs) set the minimal probability of the good outcome in the lottery at a lower level than their beliefs would suggest.

The explanation advanced in this thesis concerns the emotional costs of betrayal. People with pessimistic beliefs about others' trustworthiness have nothing to lose by having another player determine their payoffs: either they will have their pessimistic beliefs confirmed or be pleasantly surprised by others' trustworthiness. As such, they may as well play the trust game or request a high probability to win the good outcome in a computerised lottery. People with optimistic trustworthiness-beliefs, however, can either have their beliefs confirmed or be *un*pleasantly surprised by others' lack of trustworthiness, and might thus prefer not to have their beliefs tested. Instead they prefer to play a computerised lottery, even when it has a lower probability of the good outcome than the probability they would expect in the trust game.

Assuming our findings are not due to alternative explanations (outlined in Chapter 6 and to be confirmed through future research), the association between beliefs about others' trustworthiness and betrayal aversion may explain the mixed evidence for betrayal aversion in previous work. Previous work which did not find betrayal aversion may have had many participants with relatively pessimistic trustworthiness-beliefs; indeed, Fetchenhauer and Dunning (2012) effectively *created* low trustworthiness-beliefs by informing participants that the chance of a good outcome was only 46%. Participants with pessimistic trustworthiness-beliefs may have felt that the trust game offered an opportunity to be pleasantly surprised by others' kind actions, and so may have opted into the trust game at rates equal to or even higher than those of opting into a lottery.

## 7.2 Strengths and Potential Limitations of the Overall Research

The strength of the studies presented in this thesis arises from our unusual methodological rigour. First, we based our sample sizes (all *n*s>103) on those found in the literature and aimed for sufficient power to avoid Type II errors due to small sample sizes. Second, although not completely representative of the general population, the gender balance across our studies (59% female, 41% male) was more representative than that in most psychology studies. For example, across the samples of studies published in the *Journal of Personality and*

*Social Psychology* in 2002, 71% of the participants were female (Gosling, Vazire, Srivastasa, & John, 2004). Finally, in our designs, we aimed to minimise confounds which may have been underlying the psychological phenomena studied in this thesis in previous research (e.g., miscomprehension in the beads task, equality concerns in the trust game). In addition, when a confound could not be minimised within the study design, we collected self-report measures to account for potential influences in the analyses (e.g., risk-aversion).

The adoption of experimental economics methods also forms a strength of this thesis. First, I included detailed written instructions and required participants to answer comprehension questions correctly before starting the experimental task. This reduced the chance of miscomprehension and thus reduced noise, which, in turn, increased power (Hertwig & Ortmann, 2001). Similarly, by incentivising the paradigms, I motivated participants to stay attentive which generally improves performance (e.g., Nicholls et al., 2014). Such improved performance, again, may have reduced noise and increased power. Furthermore, I avoided the use of deception, even when this meant I had to implement complex, convoluted solutions (e.g., estimating others' estimates for undesirable questions to provide an objectively correct answer for such questions). A major concern for many experimental economists is that participants who have experienced deception in one study experience spill-over effects of such deceit into other studies (Barrera & Simpson, 2012). As such, participants would not trust claims made in other experiments. Although circumventing deception may have complicated the studies presented in this thesis, I avoided adding distrust to the subject pool.

However, there are a few potential limitations to the research presented here. Limitations specific to each study have been discussed in Chapters 2 to 6.

Therefore, the focus here is on limitations that apply to the collection of studies reported in this thesis.

### 7.2.1 Sample Considerations

All studies used samples from a healthy student population, which leads to a few considerations regarding the extent to which the findings can be generalised to other populations.

For the JTC bias, our results were found in non-clinical populations. This means that our conclusions concerning clinical populations are tentative. However, many researchers consider delusional ideation to form a continuum with normal experiences (e.g., Peters et al., 1999; van Os et al., 2009), where psychosis would be at the extreme end of the spectrum (e.g., Freeman et al., 2008; Green et al., 2011). People with non-clinical symptoms tend to be at a higher risk of developing clinical, psychotic disorders than those without such symptoms (e.g., Heriot-Maitland et al., 2012; Kelleher et al., 2012; Van Os, Hanssen, Bijl, & Ravelli, 2000). Furthermore, we measured delusion-proneness with the Peters et al. Delusions Inventory (PDI), which has good validity, as, for example, delusional patients score higher than healthy participants (Peters et al., 2004). In addition, Lincoln, Ziegler, Lüllmann, Müller, and Rief (2010) found that patients' self-reported delusional ideation on the PDI corresponded with observer-rated symptom severity. As such, we believe the insights from our study using a non-clinical population are relevant to clinical populations. Research on non-clinical populations might even offer advantages, as confounds arising from medication or hospitalisation are avoided. Nevertheless, replications of our findings using clinical populations in future research would strengthen our conclusions.

Another consideration is that all participants were students or staff members at RHUL, a university in the United Kingdom. They thus formed a sample of a western, educated, industrialised, rich and democratic (WEIRD) society (Henrich, Heine, & Norenzayan, 2010), potentially limiting the generalisability of the findings to other societies. Yet, with regards to the western, industrialised, and democratic characteristics, several of the phenomena investigated in this thesis have previously been investigated in other cultures. For example, Chang, Asakawa, and Sanna (2001) found that European Americans and Japanese participants both showed optimistic biases for negative future events, which formed the stimuli in our self-deceptive optimism study. As another example, Bohnet et al. (2008) found evidence for betrayal aversion across diverse countries (e.g., Oman, Switzerland, and the United States). Therefore, it is unlikely that our findings are due to the specific cultural background of our participants. Nevertheless, as with all findings, replication of the current findings in diverse populations is important to ascertain their robustness and generalisability.

With regards to the characteristic of being educated, one might assume that our participants, being at university, were more intelligent than the average population. This might affect the extent to which they deviate from rationality as defined by formal logic rules. Indeed, in the study reported in Chapter 2, we did find intelligence to significantly predict variation in deviations from the optimal amount of data to gather. Nevertheless, the variable of interest (i.e., delusion-proneness) predicted additional, unique variation in these deviations. Therefore, intelligence was not the only predictor of rationality. Furthermore, Stanovich and West (2008) found that cognitive ability (i.e., intelligence) and rationality are only related when the task clearly outlines the presence of a conflict between a normative and a biased response, in which case more intelligent participants show less biased responding. However, when the more intelligent participants are not aware of the need of a normative response, they are as biased as less intelligent participants. As the studies in this thesis did not present a clear conflict between different types of responses (in the way the Linda problem presents a conflict between a probabilistically-correct and an intuitive, conversational-exchange-correct response), intelligence may not have influenced our findings to a large extent. Ideally, future studies on rationality would include measures of intelligence, but this may not always be possible due to financial or time constraints (as discussed later under 7.2.2 on page 236).

Finally, with regards to the wealth characteristic, the incentives may have influenced our samples less than they would have influenced participants from less affluent backgrounds. The magnitude of our rewards (i.e., ranging from a few pence to a few pounds) may not have motivated our participants enough to report accurate answers. However, for the optimism bias, for example, Simmons and Massey (2012) found equal levels of optimism if accuracy was rewarded with $5 as when it was rewarded with $50. In addition, Woodward et al. (2009) did not find a difference for the JTC bias between a condition with rewards of $0.25 and a condition with rewards of $5. Finally, in a meta-analysis, N. D. Johnson and Mislin (2011) did not find an effect of the amount at stake on either trusting or trustworthy behaviours. They did find that students are less trustworthy (i.e., as the trustee, they return less or choose the unfair outcome more frequently) than non-student populations. Students and non-students were not different in their trusting behaviour, which constitutes the behaviour we considered. Overall, this suggests that the magnitude of the reward might not influence deviations from rationality. Future research would benefit from including a measure of socioeconomic status (SES), so as to account for differences in utility obtained from the experimental rewards, as the small rewards might lead to less utility for participants with a higher SES than for participants with a lower SES. Participants could be asked to report their parents' education level and occupation, from which SES can be estimated (Hollingshead, 1975). It must be noted that several of our manipulations were within-subject. Therefore, influences of the sample population (e.g., sensitivity to reward magnitude), for example, should not have been responsible for differences between conditions.

### 7.2.2 Financial Constraints Imposed by Use of Incentives

In order to encourage arriving at the rational response, defined as the response with the highest expected value, incentives were included for all studies. The importance of this practice was highlighted by the findings in Chapter 3: in the presence of incentives, low-delusion-prone participants provided estimates that were not significantly different from Bayesian probabilities, while these probabilities were underestimated without incentives. However, incorporation of incentives necessarily posed certain financial constraints on the amount of data that could be collected (Baron, 2001). As described in Chapter 1 (on page 76), sample size was prioritised to ensure that individual differences (e.g., delusion-proneness, trustworthiness-beliefs) would be varied enough to investigate their association with other measures. Given the financial constraints, this meant that study duration needed to be kept to a necessary minimum, as studies conducted in the EconLab at RHUL must pay participants £8 to £10 per hour, on average. As a consequence, certain pertinent measures may have been excluded.

In Chapter 3, for example, intelligence was not measured. Including an intelligence measure would have increased the cost of the experiment by at least £300 (at least 100 participants to be paid £3 more for the additional 20 minutes they would need). Given the mixed evidence for the role of intelligence on the JTC bias (Ziegler et al., 2008), this was deemed too expensive for its potential worth. As a result, we cannot establish to what extent intelligence may have affected our findings in this study. However, intelligence was measured in Chapter 2, where delusion-proneness could explain unique variance not accounted for by intelligence, and so we feel confident that results in Chapter 3 are not (fully) attributable to differences in intelligence. Ideally, however, future studies using probability-estimate variants of the beads task would include a measure of intelligence, to further elucidate its role in the JTC bias.

In Chapter 4, feedback from a pilot study indicated that the within-subject manipulation of a neutral versus a bias condition was quite evident. However, given the sample sizes required to detect gender differences for sexual over-perception (at least n=45 for each gender; Farris et al., 2008b), use of a between-subjects manipulation would have required at least 90 additional participants. This was expected to have cost at least an additional £360. For this reason, the manipulation was kept within-subject but counterbalanced across sessions. To account for potential effects due to the salience of the manipulation, an analysis of only the first condition participants encountered was conducted. This analysis also did not provide support for men's sexual over-perception bias. Hence, we feel that the within-subject nature of the design cannot explain the lack of a difference between the two conditions. Nevertheless, future research might benefit from comparing neutral and bias conditions using a between-subjects design.

In Chapter 5, it was noted that including items that negated the chance of a negative event could shed further light on the self-deceptive sampling we hypothesised. Including additional items, which were to be repeated in four rounds, would have made the experiment longer and thus more expensive by approximately £250 (at least 100 participants to be paid £2.50 for the estimated additional 15 minutes they would need). Anecdotally, several participants indicated that the experiment was already quite repetitive. Hence, making the experiment longer by including more items could have led to a drop in attention. Yet, without such items, we cannot establish whether second estimates for desirable (or more accurately: not-undesirable) items would be higher than first estimates. Future research could incentivise the items at a higher rate and decrease the show-up fee to compensate, in order to maintain attention throughout the experiment. This may also reduce the additional costs of including extra items, as participants only need to be paid for estimates close to the correct answer; an increased show-up fee would be paid independently from performance and thus increase study costs more than increasing the rewards for accuracy.

### 7.2.3 Defining a Normative Standard

The incentives were used to create optimal decisions from the *homo economicus* perspective, where people are considered rational and self-interested to the extent that they aim to maximise their expected monetary value (Glimcher et al., 2008). This view has been challenged, the charge being that putatively "irrational" behaviour may simply stem from responses being compared against the wrong norm (Gigerenzer, 1996; Stanovich & West, 2000). Measured against alternative norms (i.e., norms that participants may have assumed they were expected to invoke and employ), responses might not be biased (De Neys &

Bonnefon, 2013). Consider the Linda problem, outlined in Chapter 1 (on page 21): if people realise the normative standard is to use probability theory, they might correctly decide that it is more likely that Linda is a bank teller than that she is a bank teller *and* a feminist. However, some might apply the norm of conversational exchange (i.e., that conversation does not generally include irrelevant details) and assume the descriptive details are important for their answer (Stanovich & West, 2000). Measured against the appropriate norm (i.e., that which people applied, such as the norm of conversational exchange), people's judgement that Linda is more likely to be a bank teller *and* a feminist than just a bank teller might not constitute biased or incorrect responses (De Neys & Bonnefon, 2013).

If people consider a different or an additional norm than the one the experimenters used as their standard, participants might experience conflict between different norms when providing their judgment. Although people's verbal descriptions of their thought processes while making a decision on tasks like the Linda problem generally do not indicate that they are considering different norms or experiencing conflict between different norms, implicit measures do tend to suggest a detection of conflict (De Neys, Cromheeke, & Osman, 2011). De Neys et al. (2011) showed that people were less confident in their decisions when there was conflict between norms, as in the standard Linda problem, compared to when there was no conflict between norms. The conflict was removed by including the likely, rather than the unlikely, characteristic in both answer options: a) "Linda is a feminist", or b) "Linda is feminist and a bank teller" (cf. the standard answers: a) "Linda is a bank teller", or b) "Linda is a bank teller and a feminist"). For the no-conflict condition, the intuitive, stereotypical response is the same as the normative response according to the conjunction rule. As participants were more confident about their responses to no-conflict questions than to conflict questions, this suggests they detected conflict between different norms in the latter case (De Neys et al., 2011).

Given De Neys et al.'s (2011) findings, it would have been interesting to measure confidence about the estimates provided in Chapter 5. If confidence levels for estimated answers to undesirable questions were lower than those for neutral questions, this might point to people's awareness of conflict between their desirable "self-deceptive" and more objective estimates for such undesirable questions. In addition, such data could provide further support for Lench et al.'s (2014) finding that the probability of desirable information is perceived to have a higher variance than the probability of non-desirable information. However, including a measure of confidence would have made the study longer and thus more expensive, which was a strong determinant for excluding this measure given financial constraints (see 7.2.2 on page 237).

In this thesis, optimal decisions were determined with the utmost care, and generally from the *homo economicus* perspective. Deciding what other norm would be correct is a difficult and contentious issue (De Neys & Bonnefon, 2013). In this thesis, responses were generally *not* significantly different from the norm we selected, at least in neutral conditions, for low-delusion-prone participants, or for tasks that minimised methodological confounds. This suggests that our chosen norm was often also the norm selected by participants. However, individual differences in the selection of a norm exist. For example, in Chapter 6 we found that maximising expected value is not the ultimate goal for everyone, as some people were willing to face a larger risk of obtaining only a small reward if that meant they could avoid the emotional costs of betrayal. This suggests that normative answers should focus on expected *utility*, taking into account relevant preferences and prior beliefs and attitudes, rather than just expected value. Although the optimal decisions in Chapter 2 were based on expected value from a risk-neutral perspective, risk-aversion was measured, so that it would be possible to investigate whether deviations from the optimal decision could be explained by risk attitudes. This did not seem to be the case, however. First, risk-aversion was not associated with draws-to-decision. Second, our sample was generally slightly risk-averse, which would suggest they would want *more* certainty of obtaining the reward, and thus decide *later* than the calculated optimal points; yet decisions were made *before* these optimal points. As such, our results of deviations from rationality cannot be ascribed to having assumed a risk-neutral attitude in determining the norm.

## 7.3 General Implications for Rationality Research

As discussed in Chapter 1, psychologists have long debated about human irrationality, with one side arguing that the human mind is simply not rational (Kahneman, 2003; Tversky & Kahneman, 1983, 1986), while the other side suggests that the environment of the tasks used to measure rationality might contribute to putative "irrationality" (Gigerenzer & Goldstein, 1996; Gigerenzer & Sturm, 2012). The results in this thesis lend support to aspects of both arguments, though they seem to be slightly in favour of the bounded rationality view. By adapting the task environments, some deviations from rationality were still found (e.g., optimism bias), suggesting the human mind is limited in its rationality, while other deviations were not replicated (e.g., sexual over-perception), suggesting that previous biases may have been influenced by the task environment. Furthermore, deviations from rationality have been rather small and some of them seem to be associated with signs of psychopathology.

As such, it appears that the healthy human mind may be more rational than has been assumed.

One possibility is that behaviours are more biased than beliefs. As such, previous research, which has used predominantly decision-making tasks, may have rightly found biases, but specifically biased *behaviour*. The results from the predominantly judgment-related tasks in this thesis suggest that biased *beliefs* are less common.

The human mind may not reflect the ideal of *homo economicus*, as natural selection has shaped it for its adaptive abilities in ancestral environments, rather than for theoretically rational abilities or perfect representations of reality. The biological limitations of the human mind may lead to imperfect, noisy (i.e., distorted) cognitive processes, which could lead to systematic biases. Hilbert (2012) has suggested that noise in memory processes underlies at least eight common cognitive biases, including conservatism. Conservatism, where people underestimate high probabilities, but overestimate low probabilities, has been found in probability reasoning tasks (Phillips & Edwards, 1966) and this finding was replicated in Chapters 3 and 4. In binary decisions, the conservatism bias consists of insufficient use of conditional Bayesian likelihoods (Hilbert, 2012), where posterior probabilities are not updated from the prior probability sufficiently (Phillips & Edwards, 1966). Individually different assessments of likelihoods (i.e., noise), such as when a man might think that a woman looking away is quite likely even if she is interested in him, could ultimately lead to different beliefs and decisions, including ones that are conservative compared to the normative standard. In support of the notion that noise could underlie deviations from rationality, Moutoussis et al. (2011) found that computational models of beads-task performance could distinguish between data from patients and data from controls through a noise parameter. Note that noise could make evidence more or less ambiguous than intended in the experimental set-up. As mentioned before, ambiguity regarding the subjective likelihood of observing a smile from a non-sexually-interested woman (i.e., $P$(behaviour|not-interested)) or of observing a sexually-interested woman breaking eye contact (i.e., $P$(behaviour|interested)), could contribute to differences between men and women in research on the sexual over-perception bias. As such, noise in reasoning processes might increase the ambiguity of stimuli, which, in turn, could increase noise in empirically measured cognitions and behaviours.

We undertook stringent procedures to minimise noise in the task environments in the studies reported in this thesis and biases were still found in the majority of the studies. For example, likelihood information was explicitly stated in the relevant studies in this thesis. Furthermore, memory load was minimised by displaying previous draws of the sequence in Chapter 2, or by displaying participants' previously entered answers in Chapter 4. As biases were still found despite these adjustments to the tasks, it would seem unjustified to ascribe all irrationality to noise in cognitive processes.

Finally, people may have different goals within reasoning tasks, leading to different responses. This was alluded to above with the suggested confirmatory reasoning style of delusion-prone individuals (Dudley & Over, 2003). These individuals might not have the goal of arriving at the theoretically correct, and maximally paying, decision; instead, they might prefer to find supportive, confirming evidence for the hypothesis they expect to be correct. With such goals in mind, by considering a minimal amount of (supporting) evidence, people are showing "personal rationality", which is "reasoning or acting in such a way as to achieve one's goals", even if they are not showing "impersonal rationality”, which is “reasoning or acting in conformity with a relevant normative system such as formal logic or probability theory” (Evans & Over, 1996, p. 357). Hence, normatively irrational responses might not solely be due to participants applying a different normative standard to the experimenter, as suggested earlier when lower confidence levels indicate a detection of conflict between which norms to apply (De Neys et al., 2011). Participants might also have (personal) goals different from applying any normative standard, such as wanting to confirm their hypothesis. Mixed methods, which use qualitative methods in addition to the quantitative methods used in this thesis, could shed light on the goals people have within various reasoning tasks. Balzan, Delfabbro, and Galletly (2012) used such mixed methods to assess the underlying reasons for irrational over-adjustment of probability estimates in light of a single piece of disconfirming evidence. From the verbal reports provided by participants while they chose their responses it appeared that the extreme over-adjustment (i.e., choosing the non-favoured jar) was mostly due to participants thinking that jars were swapped throughout the sequence. Without these verbal reports of the reasons behind decisions made (i.e., qualitative measures), the extreme over-adjustment would only suggest that participants misunderstood the task, but not why or how they interpreted it differently.

Hence, rationality research might benefit from shifting the focus from quantitative methods to mixed methods, including qualitative methods that focus on asking participants to verbally reflect on their reasoning processes. However, such verbal descriptions might not reflect sub-conscious goals, so the inclusion of implicit measures of conflict detection, such as confidence levels, could be beneficial as well (De Neys et al., 2011).

In fact, people might very well be unaware of their goals. Although we did not collect such information, we could speculate that people's background might have influenced how they behaved in the studies in this thesis. A theory regarding differential susceptibility to the environment suggests that some people are neurobiologically more susceptible to both negative and positive environmental conditions (Ellis, Boyce, Belsky, Bakermans-Kranenburg, & Van IJzendoorn, 2011). Such susceptible individuals show larger responses to stressors in unfavourable environmental conditions with low support levels and a lack of resources, such as showing more physical illnesses (Boyce et al., 1995). Yet, they *also* flourish in favourable and supportive environments, for example by showing fewer illnesses than less susceptible children in favourable environments (Boyce et al., 1995). The susceptibility to environmental conditions is thought to have been maintained throughout evolution because it could lead to behaviours that support evolutionary fitness, in the environment in which the individual is raised as well as potential other environmental conditions which could be encountered over the course of a lifetime (Ellis et al., 2011).

Different reproductive strategies are one particular example of the effects of such susceptibility to the environment on behaviour as well as on biology. In line with life history theory, Belsky, Steinberg, and Draper (1991) suggest that children growing up in an environment with scarce resources and with an insecure attachment arising from insensitive, inconsistent, or rejecting parenting adopt a reproductive strategy that would be most beneficial in this environment. This reproductive strategy would entail earlier maturation (e.g., earlier menarche) and earlier sexual activity, as well as forming short-term, unstable bonds and showing limited parental investment. In contrast, children growing up in an environment with sufficient resources and with a secure attachment stemming from sensitive, supportive, and responsive parenting, might develop a reproductive strategy with delayed maturation and later engagement in sexual activity. Furthermore, they are more likely to form bonds that are long-term and enduring, with greater parental investment. This reproductive strategy would be most beneficial in terms of evolutionary fitness in this favourable environment (Belsky et al., 1991).

This theory could be applied to the sexual over-perception bias. Perhaps this bias is mainly shown by people who have been reared in unfavourable environments, as it might contribute to an aim to pair with as many potential mates as possible and limit parental investment. It is possible that the participants in our sexual over-perception study were more likely to have been reared in a favourable environment, being WEIRD participants (see page 233). This, then, could explain why the sexual over-perception bias was not found.

Potentially, this life-history-theory perspective could be extended to the other biases in this thesis. If people who grew up in an unfavourable environment would be more likely to form short-term bonds, perhaps they also consider others malevolent, which could contribute to the paranoid characteristics of delusions and to hasty decisions based on little evidence. Furthermore, they might be more likely to distrust others and suspect others to betray them, an expectation which, in line with our theory, would lead to ambivalence or a slight preference for a trust game when deciding whether to let a computerised lottery or another person decide on an outcome, as there is little room for unexpected betrayal and thus no reason to avoid finding out this information (as per betrayal aversion). Finally, perhaps developing an optimism bias would be harmful in an unfavourable environment, as one would be constantly disappointed (e.g., if one is more susceptible to the unfavourable environment, they might experience more illnesses than average; Ellis et al., 2011). This, in turn, might harm self-esteem, which might already be low after being reared in an unfavourable environment without having developed a secure attachment.

Considering these possibilities, the opposite outcomes would be expected for people who have grown up in favourable environments, and perhaps for people with lower susceptibility levels to environmental conditions as well. Our WEIRD participants might be more likely to have fallen into these categories of low susceptibility or high susceptibility to a favourable environment, rather than high susceptibility to an unfavourable environment. This, then, would lead the majority of our participants to have an absence of the JTC bias, an absence of the sexual over-perception bias, but show an optimism bias as well as betrayal aversion, as betrayal might not be expected as much as by those reared in unfavourable conditions. The general trends of our findings are in line with these suggested effects of having a favourable life history. However, in our dynamic task, most of our participants decided in advance of the point at which they would have maximised their *monetary* payoff (i.e., our operationalisation of showing the JTC bias). It could be that extraneous costs influenced decision making. For example, participants may have been fatigued as the task progressed, in which case it may have been rational to decide before the point that maximised expected value, to shorten the duration of the task. If this is the case, it should not have been found in the static task; indeed, the JTC bias was absent for the low-delusion-prone participants.

The fact that our participants were risk-averse across studies would also be expected if they had such backgrounds. If the environment is safe and resources are plentiful, it would be wise to exploit these favourable conditions, rather than take a risk and explore other, potentially more rewarding other environments.

This decision to explore could be a costly error (as discussed on page 43), if energy is spent on new environments while one's current environment has plentiful resources and one might be exposed to dangers in new environments.

In total, life history theory could provide a potential comprehensive framework for the findings reported in this thesis. Having been raised in favourable environments and having developed secure attachments, the majority of our WEIRD participants might be risk-averse, consider enough evidence before making decisions and thus not show the JTC bias, be more interested in forming long-term bonds and thus not show the sexual over-perception bias, have had their optimistic bias reinforced over time and thus maintained it, and show betrayal aversion as people might be trusted more than warranted.

## 7.4 Conclusion

I have investigated irrational beliefs in a variety of domains. As irrationality in several reasoning problems has been ascribed to methodological artefacts (Gigerenzer, 1996; Stanovich & West, 2000; Sturm, 2012), I aimed to (re)investigate a series of psychological phenomena through the use of rigorous methods from experimental economics, including detailed instructions and comprehension questions. I also introduced incentives so that rational decision-making would involve maximisation of payoffs (Hertwig & Ortmann, 2001); this was hypothesised to minimise the chance of low motivation confounding the results or the chance of biased reporting without biased believing (Schotter & Trevino, 2014). Yet, despite these attempts to encourage the use of the normative standard (e.g., Bayesian posterior probability), I still found considerable evidence of departures from this standard.

I found that if someone is delusion-prone, their probability reasoning might be unaffected or even more conservative than that of non-delusion-prone people. Despite relatively intact probability reasoning under most circumstances, delusion-prone people gather less data than would be optimal, and, as such, do not maximise their payoff. I even found these suboptimal decisions when gathering additional data did not require additional time or effort.

If a man tends to over-perceive sexual interest from women, this might be due to different prior beliefs about women's sexual interest, which, in turn, could be due to different socialisation. I did not find evidence that such sexual over-perception constitutes a cognitive, belief-updating bias.

If someone provides an optimistic estimate of their future outcomes, asking them to second-guess themselves may improve accuracy if the average of the two estimates is taken. However, this second guess increases bias, as participants seem to sample selectively to support even more optimistic estimates the second time around.

Finally, if someone is optimistic about others' trustworthiness, he might prefer to escape testing this belief by accepting a higher risk of a bad outcome as long as this avoids emotional costs of potential betrayal.

The fact that some results were only found for a subset of people (e.g., only high-delusion-prone participants were suboptimal in an incentivised, one-shot data-gathering task) highlights the extent of individual differences in conformity to relevant rational norms. Many departures from relevant norms stem from distortions in information processing, as per the accounts reviewed above. The fact that we documented such deviations even after major procedural improvements might seem to give a bleak image of human rationality. However, individual differences in *choosing* which norm to select, arising from differences in goals and desires (Dudley & Over, 2003), may also account for some of these differences. To the extent that people achieve their own idiosyncratic goals, they may be “personally rational” (Evans & Over, 1996). For example, if people desire to remain optimistic about their own future more than they desire to obtain financial rewards, they may better achieve their goals through self-deceptive optimism than through adherence to normative standards. As such, apparent deviations from impersonal rational norms might be personally optimal.

# References

Abbey, A. (1982). Sex differences in attributions for friendly behavior: Do males misperceive females' friendliness. *Journal of Personality and Social Psychology, 42*(5), 830-838.

Abbey, A., & Harnish, R. J. (1995). Perception of sexual intent: The role of gender, alcohol consumption, and rape supportive attitudes. *Sex Roles, 32*(5/6), 297-313.

Aimone, J. A., & Houser, D. (2011). Beneficial betrayal aversion. *PLOS One, 6*(3), e17725. doi: 10.1371/journal.pone.0017725

Aimone, J. A., & Houser, D. (2012). What you don't know won't hurt you: A laboratory analysis of betrayal aversion. *Experimental Economics, 15*(4), 571-588. doi: 10.1007/s10683-012-9314-z

Aimone, J. A., & Houser, D. (2013). Harnessing the benefits of betrayal aversion. *Journal of Economic Behavior & Organization, 89*, 1-8. doi: 10.1016/j.jebo.2013.02.001

Aimone, J. A., Houser, D., & Weber, B. (2014). Neural signatures of betrayal aversion: An fMRI study of trust. *Proceedings of the Royal Society B, 281*(1782), 20132127. doi: 10.1098/rspb.2013.2127

Alicke, M. D. (1985). Global self-evaluation as determined by the desirability and controllability of trait adjectives. *Journal of Personality and Social Psychology, 49*(6), 1621-1630.

Alicke, M. D., & Sedikides, C. (2009). Self-enhancement and self-protection: What they are and what they do. *European Review of Social Psychology, 20*(1), 1-48. doi: 10.1080/10463280802613866

Alloy, L. B., & Abramson, L. Y. (1982). Learned helplessness, depression, and the illusion of control. *Journal of Personality and Social Psychology, 42*(6), 1114-1126.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (DSM-V)* (5 ed.). Arlington, VA: American Psychiatric Association.

Andreou, C., Moritz, S., Veith, K., Veckenstedt, R., & Naber, D. (2013). Dopaminergic modulation of probabilistic reasoning and overconfidence in errors: A double-blind study. *Schizophrenia Bulletin*.

Ariely, D., & Norton, M. I. (2007). Psychology and experimental economics: A gap in abstraction. *Current Directions in Psychological Science, 16*(6), 336-339. doi: 10.1111/j.1467-8721.2007.00531.x

Arthur, W. J., & Day, D. V. (1994). Development of a short form for the Raven Advanced Progressive Matrices test. *Educational and Psychological Measurement, 54*(2), 394-403.

Baker, L. A., & Emery, R. E. (1993). When every relationship is above average: Perceptions and expectations of divorce at the time of marriage. *Law and Human Behavior, 17*(4), 439-450.

Balzan, R., Delfabbro, P., & Galletly, C. (2012). Delusion-proneness or miscomprehension? A re-examination of the jumping-to-conclusions bias. *Australian Journal of Psychology, 64*(2), 100-107. doi: 10.1111/j.1742-9536.2011.00032.x

Balzan, R., Delfabbro, P. H., Galletly, C. A., & Woodward, T. S. (2012). Over-adjustment or miscomprehension? A re-examination of the jumping to conclusions bias. *Australian and New Zealand Journal of Psychiatry, 46*(6), 532-540. doi: 10.1177/0004867411435291

Balzan, R., Delfabbro, P. H., Galletly, C. A., & Woodward, T. S. (2013). Confirmation biases across the psychosis continuum: The contribution of hypersalient evidence-hypothesis matches. *British Journal of Clinical Psychology, 52*, 53-69. doi: 10.1111/bjc.12000

Bandura, A. (2011). Self-deception: A paradox revisited. *Behavioral and Brain Sciences, 34*(01), 16-17. doi: 10.1017/S0140525X10002499

Bardsley, N. (2000). Control without deception: Individual behaviour in free-riding experiments revisited. *Experimental Economics, 3*(3), 215-240. doi: 10.1007/BF01669773

Bardsley, N., Cubitt, R., Loomes, G., Moffatt, P., Starmer, C., & Sugden, R. (2010). *Experimental economics: Rethinking the rules*. Princeton, N.J: Princeton University Press.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition, 21*, 37-46.

Baron, J. (2001). Purposes and methods. *Behavioral and Brain Sciences, 24*(3), 403.

Barrera, D., & Simpson, B. (2012). Much ado about deception: Consequences of deceiving research participants in the social sciences. *Sociological Methods & Research, 41*(3), 383-413. doi: 10.1177/0049124112452526

Belsky, J., Steinberg, L., Draper, P. (1991). Childhood experience, interpersonal development, and reproductive strategy: An evolutionary theory of socialization. *Child Development, 62*, 647-670.

Bentall, R. P., Corcoran, R., Howard, R., Blackwood, N., & Kinderman, P. (2001). Persecutory delusions: A review and theoretical integration. *Clinical Psychology Review, 21*(8), 1143-1192.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior, 10*(1), 122-142. doi: 10.1006/game.1995.1027

Bijleveld, E., Custers, R., & Aarts, H. (2011). Once the money is in sight: Distinctive effects of conscious and unconscious rewards on task performance. *Journal of Experimental Social Psychology, 47*, 865-869. doi: 10.1016/j.jesp.2011.03.002

Bohnet, I., Greig, F., Herrmann, B., & Zeckhauser, R. (2008). Betrayal aversion: Evidence from Brazil, China, Oman, Switzerland, Turkey, and the United States. *American Economic Review, 98*(1), 294-310.

Bohnet, I., & Zeckhauser, R. (2004). Trust, risk and betrayal. *Journal of Economic Behavior & Organization, 55*, 467-484. doi: 10.1016/j.jebo.2003.11.004

Bonetti, S. (1998). Experimental economics and deception. *Journal of Economic Psychology, 19*, 377-395.

Bortolotti, L. (2009). *Delusions and other irrational beliefs*. Oxford, UK: Oxford University Press.

Boyce, W. T., Chesney, M., Alkon, A., Tschann, J. M., Adams, S., Chesterman, B., . . . & Wara, D. (1995). Psychobiologic reactivity to stress and childhood respiratory illnesses: Results of two prospective studies. *Psychosomatic Medicine, 57*, 411-422.

Brissette, I., Scheier, M. F., & Carver, C. S. (2002). The role of optimism in social network development, coping, and psychological adjustment during a life transition. *Journal of Personality and Social Psychology, 82*(1), 102-111. doi: 10.1037//0022-3514.82.1.102

Bröder, A. (1998). Deception can be acceptable. *American Psychologist, 53*(7), 805-806. doi: 10.1037/h0092168

Broome, M. R., Johns, L. C., Valli, I., Woolley, J. B., Tabraham, P., Brett, C., . . . McGuire, P. K. (2007). Delusion formation and reasoning biases in those at clinical high risk for psychosis. *The British Journal of Psychiatry, 191*(51), s38-s42. doi: 10.1192/bjp.191.51.s38

Brüne, M. (2005). "Theory of mind" in schizophrenia: A review of the literature. *Schizophrenia Bulletin, 31*(1), 21-42. doi: 10.1093/schbul/sbi002

Buehler, R., Griffin, D., & MacDonald, H. (1997). The role of motivated reasoning in optimistic time predictions. *Personality and Social Psychology Bulletin, 23*, 238-247. doi: 10.1177/0146167297233003

Buehler, R., Griffin, D., & Ross, M. (1995). It's about time: Optimistic predictions in work and love. *European Review of Social Psychology, 6*(1), 1-32. doi: 10.1080/14792779343000112

Buss, D. M. (2013). The science of human mating strategies: An historical perspective. *Psychological Inquiry, 24*(3), 171-177. doi: 10.1080/1047840X.2013.819552

Cadwalladr, C. (2012). The optimism bias: Reasons to be cheerful. Retrieved June 26, 2014, from http://www.theguardian.com/books/2012/jan/01/tali-sharot-optimism-bias-interview

Cafferkey, K., Murphy, J., & Shevlin, M. (2014). Jumping to conclusions: The association between delusional ideation and reasoning biases in a healthy student population. *Psychosis: Psychological, Social and Integrative Approaches, 6*(3), 206-214. doi: 10.1080/17522439.2013.850734

Cambridge Dictionary. (Ed.) (2014) Cambridge Dictionaries Online. Cambridge University Press.

Camerer, C. F. (2003). *Behavioral game theory: Experiments in strategic interaction*. Princeton, NJ: Princeton University Press.

Camerer, C. F., Loewenstein, G., & Prelec, D. (2004). Neuroeconomics: Why economics needs brains. *Scandinavian Journal of Economics, 106*(3), 555-579. doi: 10.1111/j.1467-9442.2004.00378.x

Camerer, C. F., & Weigelt, K. (1988). Experimental tests of a sequential equilibrium reputation model. *Econometrica, 56*(1), 1-36.

Caplin, A., Dean, M., Glimcher, P. W., & Rutledge, R. B. (2010). Measuring beliefs and rewards: A neuroeconomic approach. *The Quarterly Journal of Economics, 125*(3), 923-960. doi: 10.1162/qjec.2010.125.3.923

Chang, E. C., Asakawa, K., & Sanna, L. J. (2001). Cultural variations in optimistic and pessimistic bias: Do Easterners really expect the worst and Westerners really expect the best when predicting future life events? *Journal of Personality and Social Psychology, 81*(3), 476-491. doi: 10.1037//0022-3514.81.3.476

Choi, E., & Hur, T. (2013). Is reading sexual intention truly functional? The impact of perceiving a partner's sexual intention on courtship initiation behaviors. *Archives of Sexual Behavior, 42*(8), 1525-1533. doi: 10.1007/s10508-013-0153-6

Colbert, S. M., Peters, E., & Garety, P. A. (2010). Jumping to conclusions and perceptions in early psychosis: Relationship with delusional beliefs. *Cognitive Neuropsychiatry, 15*(4), 422-440. doi: 10.1080/13546800903495684

Colbert, S. M., & Peters, E. R. (2002). Need for closure and jumping-to-conclusions in delusion-prone individuals. *The Journal of Nervous and Mental Disease, 190*(1), 27-31.

Coltheart, M., Langdon, R., & McKay, R. (2011). Delusional belief. *Annual Review of Psychology, 62*, 271-298. doi: 10.1146/annurev.psych.121208.131622

Coltheart, M., Menzies, P., & Sutton, J. (2010). Abductive reasoning and delusional belief. *Cognitive Neuropsychiatry, 15*(1-3), 261-287. doi: 10.1080/13546800903439120

Cook, K. S., & Yamagishi, T. (2008). A defense of deception on scientific grounds. *Social Psychology Quarterly, 71*(3), 215-221.

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition, 31*(3), 187-276.

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition, 58*, 1-73.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*(1), 7-29. doi: 10.1177/0956797613504966

Dana, J., Cain, D. M., & Dawes, R. M. (2006). What you don't know won't hurt me: Costly (but quiet) exit in dictator games. *Organizational Behavior and Human Decision Processes, 100*, 193-201. doi: 10.1016/j.obhdp.2005.10.001

De Neys, W., & Bonnefon, J.-F. (2013). The 'whys' and 'whens' of individual differences in thinking biases. *Trends in Cognitive Sciences, 17*(4), 172-178. doi: 10.1016/j.tics.2013.02.001

De Neys, W., Cromheeke, K., & Osman, M. (2011). Biased but in doubt: Conflict and decision confidence. *PLOS One, 6*(1), e15954. doi: 10.1371/journal.pone.0015954

DeKay, W. T., Haselton, M. G., & Kirkpatrick, L. A. (2000). Reversing figure and ground in the rationality debate: An evolutionary perspective. *Behavioral and Brain Sciences, 23*(5), 670-671.

Deweese-Boyd, I. (2012). Self-deception. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Spring 2012 ed.). Retrieved from http://plato.stanford.edu/archives/spr2012/entries/self-deception/.

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inferene*. New York: Palgrave MacMillan.

Dillard, A. J., Midboe, A. M., & Klein, W. M. P. (2009). The dark side of optimism: Unrealistic optimism about problems with alcohol predicts subsequent negative event experiences. *Personality and Social Psychology Bulletin, 35*(11), 1540-1550. doi: 10.1177/0146167209343124

Dohmen, T., Falk, A., Huffman, D., & Sunde, U. (2009). Homo Reciprocans: Survey evidence on behavioural outcomes. *The Economic Journal, 119*(536), 592-612. doi: 10.1111/j.1468-0297.2008.02242.x

Dudley, R., Dodgson, G., Sarll, G., Halheah, R., Bolas, H., & McCarthy-Jones, S. (2014). The effect of arousal on auditory threat detection and the relationship to auditory hallucinations. *Journal of Behavior Therapy and Experimental Psychiatry, 45*, 311-318. doi: 10.1016/j.jbtep.2014.02.002

Dudley, R., John, C. H., Young, A. W., & Over, D. E. (1997a). The effect of self-referent material on the reasoning of people with delusions. *British Journal of Clinical Psychology, 36*, 575-584.

Dudley, R., John, C. H., Young, A. W., & Over, D. E. (1997b). Normal and abnormal reasoning in people with delusions. *British Journal of Clinical Psychology, 36*, 243-258.

Dudley, R., & Over, D. E. (2003). People with delusions jump to conclusions: A theoretical account of research findings on the reasoning of people with delusions. *Clinical Psychology and Psychiatry, 10*, 263-274.

Eisenberger, R., & Cameron, J. (1996). Detrimental effects of reward: Reality or myth? *American Psychologist, 51*(11), 1153-1166.

Ellis, B., Boyce, W. T., Belsky, J., Bakermans-Kranenburg, M. J., & Van IJzendoorn, M. H. (2011). Differential susceptibility to the environment: An evolutionary-neurodevelopmental theory. *Development and Psychopathology, 23*, 7-28.

Epley, N., & Whitchurch, E. (2008). Mirror, mirror on the wall: Enhancement in self-recognition. *Personality and Social Psychology Bulletin, 34*(9), 1159-1170.

Evans, J. S., & Over, D. E. (1996). Rationality in the selection task: Epistemic utility versus uncertainty reduction. *Psychological Review, 103*(2), 356-363.

Falk, A., & Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior, 54*, 293-315. doi: 10.1016/j.geb.2005.03.001

Farris, C., Treat, T. A., Viken, R. J., & McFall, R. M. (2008a). Perceptual mechanisms that characterize gender differences in decoding women's sexual intent. *Psychological Science, 19*(4), 348-354. doi: 10.1111/j.1467-9280.2008.02092.x

Farris, C., Treat, T. A., Viken, R. J., & McFall, R. M. (2008b). Sexual coercion and the misperception of sexual intent. *Clinical Psychology Review, 28*, 48-66. doi: 10.1016/j.cpr.2007.03.002

Fehr, E. (2009). On the economics and biology of trust. *Journal of the European Economic Association, 7*(2-3), 235-266. doi: 10.1162/JEEA.2009.7.2-3.235

Fehr, E., & Gächter, S. (2002). Altruistic punishmen in humans. *Nature, 415*, 137-140. doi: 10.1038/415137a

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics, 114*(3).

Fetchenhauer, D., & Dunning, D. (2009). Do people trust too much or too little? *Journal of Economic Psychology, 30*, 263-276.

Fetchenhauer, D., & Dunning, D. (2012). Betrayal aversion versus principled trustfulness - How to explain risk avoidance and risky choices in trust games. *Journal of Economic Behavior & Organization, 81*, 534-541. doi: 10.1016/j.jebo.2011.07.017

Field, A. (2005). *Discovering statistics using SPSS* (2nd ed.). London, UK: Sage Publications Ltd.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London, UK: Sage Publications Ltd.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed.). London, UK: Sage Publications Ltd.

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. London, UK: Sage Publications Ltd.

Fine, C., Gardner, M., Craigie, J., & Gold, I. (2007). Hopping, skipping or jumping to conclusions? Clarifying the role of the JTC bias in delusions. *Cognitive Neuropsychiatry, 12*(1), 46-77. doi: 10.1080/13546800600750597

Fischbacher, U. (2007). z-Tree: Zurich Toolbox for Ready-made Economic Experiments. *Experimental Economics, 10*(2), 171-178.

Fletcher, G. J. O., Kerr, P. S. G., Li, N. P., & Valentine, K. A. (2014). Predicting romantic interest and decisions in the very early stages of mate selection: Standards, accuracy, and sex differences. *Personality and Social Psychology Bulletin*. doi: 10.1177/0146167213519481

Fraser, J., Morrison, A., & Wells, A. (2006). Cognitive processes, reasoning biases and persecutory delusions: A comparative study. *Behavioural and Cognitive Psychotherapy, 34*, 421-435. doi: 10.1017/S1352465806002852

Freeman, D. (2007). Suspicious minds: The psychology of persecutory delusions. *Clinical Psychology Review, 27*(4), 425-457. doi: 10.1016/j.cpr.2006.10.004

Freeman, D., Pugh, K., & Garety, P. A. (2008). Jumping to conclusions and paranoid ideation in the general population. *Schizophrenia Research, 102*(1-3), 254-260. doi: 10.1016/j.schres.2008.03.020

Freeman, D., Startup, H., Dunn, G., Černis, E., Wingham, G., Pugh, K., . . . Kingdon, D. (2014). Understanding jumping to conclusions in patients with persecutory delusions: Working memory and intolerance of uncertainty. *Psychological Medicine*. doi: 10.1017/S0033291714000592

Fridland, E. (2011). Reviewing the logic of self-deception. *Behavioral and Brain Sciences, 34*(01), 22-23. doi: 10.1017/S0140525X10002566

Gal, D., & Rucker, D. D. (2010). When in doubt, shout! Paradoxical influences of doubt on proselytizing. *Psychological Science, 21*(11), 1701-1707. doi: 10.1177/0956797610385953

Garety, P. A., & Freeman, D. (1999). Cognitive approaches to delusions: A critical review of theories and evidence. *The British Journal of Clinical Psychology, 38*(2), 113.

Garety, P. A., & Freeman, D. (2013). The past and future of delusions research: From the inexplicable to the treatable. *The British Journal of Psychiatry, 203*, 327-333. doi: 10.1192/bjp.bp.113.126953

Garety, P. A., Hemsley, D. R., & Wessely, S. (1991). Reasoning in deluded schizophrenic and paranoid patients: Biases in performance on a probabilistic inference task. *The Journal of Nervous and Mental Disease, 179*(4), 194-201.

Garety, P. A., Joyce, E., Jolley, S., Emsley, R., Waller, H., Kuipers, E., . . . Freeman, D. (2013). Neuropsychological functioning and jumping to conclusions in delusions. *Schizophrenia Research, 150*(2-3), 570-574. doi: 10.1016/j.schres.2013.08.035

Garrett, N., & Sharot, T. (2014). How robust is the optimistic update bias for estimating self-risk and population base rates? *PLOS One, 9*(6), e98848. doi: 10.1371/journal.pone.0098848

Gerrans, P. (2001). Delusions as performance failures. *Cognitive Neuropsychiatry, 6*(3), 161-173. doi: 10.1080/1354680004200016

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky (1996). *Psychological Review, 103*(3), 592-596.

Gigerenzer, G. (2004). The irrationality paradox. *Behavioral and Brain Sciences, 27*(3), 336-338.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *1996, 103*(4), 650-669.

Gigerenzer, G., & Sturm, T. (2012). How (far) can rationality be naturalized? *Synthese, 187*, 243-268. doi: 10.1007/s11229-011-0030-6

Glaeser, E. L., Laibson, D. I., Scheinkman, J. A., & Soutter, C. L. (2000). Measuring trust. *The Quarterly Journal of Economics, 115*(3), 811-846. doi: 10.1162/003355300554926

Glimcher, P. W., Fehr, E., Camerer, C. F., & Poldrack, R. (2008). *Neuroeconomics: Decision making and the brain*.

Gosling, S. D., Vazire, S., Srivastasa, S., & John, O. P. (2004). Should we trust Web-based studies? A comparative analysis of six preconceptions about Internet questionnaires. *American Psychologist, 59*(2), 93-104. doi: 10.1037/0003-066X.59.2.93

Green, C. E. L., Freeman, D., & Kuipers, E. (2011). Paranoid explanations of experience: A novel experimental study. *Behavioural and Cognitive Psychotherapy, 39*, 21-34. doi: 10.1017/S1352465810000457

Greene, A. J., & Levy, W. B. (2000). Individual differences: Variation by design. *Behavioral and Brain Sciences, 23*(5), 676-677.

Greiner, B. (2004). The online recruitment system ORSEE 2.0: A guide for the organization of experiments in economics. *Working Paper Series in Economics, University of Cologne, Germany, 10*.

Gur, R. C., & Sackeim, H. A. (1979). Self-deception: A concept in search of a phenomenon. *Journal of Personality and Social Psychology, 37*(2), 147-169.

Harris, A. J. L., & Hahn, U. (2011). Unrealistic optimism about future life events: A cautionary note. *Psychological Review, 118*(1), 135-154. doi: 10.1037/a0020997

Harris, A. J. L., & Osman, M. (2012). The illusion of control: A Bayesian perspective. *Synthese, 189*, 29-38. doi: 10.1007/s11229-012-0090-2

Harris, P. R., & Napper, L. (2005). Self-affirmation and the biased processing of threatening health-risk information. *Personality and Social Psychology Bulletin, 31*, 1250-1263. doi: 10.1177/0146167205274694

Haselton, M. G. (2003). The sexual overperception bias: Evidence of a systematic bias in men from a survey of naturally occurring events. *Journal of Research in Personality, 37*, 34-47.

Haselton, M. G., & Buss, D. M. (2000). Error management theory: A new perspective on biases in cross-sex mind reading. *Journal of Personality and Social Psychology, 78*(1), 81-91. doi: 10.1037//0022-3514.78.1.81

Haselton, M. G., & Nettle, D. (2006). The paranoid optimist: An integrative evolutionary model of cognitive biases. *Personality and Social Psychology Review, 10*(1), 47-66. doi: 10.1207/s15327957pspr1001_3

Haselton, M. G., Nettle, D., & Murray, D. R. (in press). The evolution of cognitive bias. In D. M. Buss (Ed.), *The evolutionary psychology handbook* (2nd ed.): Wiley.

Helweg-Larsen, M., & Shepperd, J. A. (2001). Do moderators of the optimistic bias affect personal or target risk estimates? A review of the literature. *Personality and Social Psychology Review, 5*(1), 74-95. doi: 10.1207/S15327957PSPR0501_5

Hemsley, D. R., & Garety, P. A. (1986). The formation of maintenance of delusions: A Bayesian analysis. *British Journal of Psychiatry, 149*, 51-56.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*, 61-135. doi: 10.1017/S0140525X0999152X

Heriot-Maitland, C., Knight, M., & Peters, E. (2012). A qualitative comparisons of psychotic-like phenomena in clinical and non-clinical populations. *British Journal of Clinical Psychology, 51*(1), 37-53. doi: 10.1111/j.2044-8260.2011.02011.x

Hertwig, R., & Gigerenzer, G. (1999). The 'conjunction fallacy' revisited: How intelligent inferences look like reasoning errors. *Journal of Behavioral Decision Making, 12*, 275-305.

Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences, 24*, 383-451.

Hertwig, R., & Ortmann, A. (2008a). Deception in experiments: Revisiting the arguments in its defense. *Ethics & Behavior, 18*(1), 59-92. doi: 10.1080/10508420701712990

Hertwig, R., & Ortmann, A. (2008b). Deception in social psychologial experiments: Two misconceptions and a research agenda. *Social Psychology Quarterly, 71*(3), 222-227.

Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science, 20*(2), 231-237.

Herzog, S. M., & Hertwig, R. (2014). Think twice and then: Combining or choosing in dialectical bootstrapping? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(1), 218-232. doi: 10.1037/a0034054

Hilbert, M. (2012). Toward a synthesis of cognitive biases: How noisy information processing can bias human decision making. *Psychological Bulletin, 138*(2), 211-237.

Hollingshead, A. B. (1975). *Four factor index of social status*. Department of Sociology. Yale University. Retrieved from http://psy6023.alliant.wikispaces.net/file/view/hollingshead+ses.pdf

Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *The American Economic Review, 92*(5), 1644-1655.

Holt, C. A., & Smith, A. M. (2009). An update on Bayesian updating. *Journal of Economic Behavior & Organization, 69*, 125-134. doi: 10.1016/j.jebo.2007.08.013

Hong, K., & Bohnet, I. (2007). Status and distrust: The relevance of inequality and betrayal aversion. *Journal of Economic Psychology, 28*(2), 197-213. doi: 10.1016/j.joep.2006.06.003

Hoorens, V., Smits, T., & Shepperd, J. A. (2008). Comparative optimism in the spontaneous generation of future life-events. *British Journal of Social Psychology, 47*(3), 441-451. doi: 10.1348/014466607X236023

Howell, D. C. (2010). *Statistical methods for psychology* (International; 7th ed.). Belmont, CA: Wadsworth, Cengage Learning.

Huq, S. F., Garety, P. A., & Hemsley, D. R. (1988). Probabilistic judgements in deluded and non-deluded subjects. *The Quarterly Journal of Experimental Psychology, 40A*, 801-812.

Jamison, J., Karlan, D., & Schechter, L. (2008). To deceive or not to deceive: The effect of deception on behavior in future laboratory experiments. *Journal of Economic Behavior & Organization, 68*(477-488). doi: 10.1016/j.jebo.2008.09.002

Johnson, D. D. P., Blumstein, D. T., Fowler, J. H., & Haselton, M. G. (2013). The evolution of error: Error management, cognitive constraints, and adaptive decision-making biases. *Trends in Ecology & Evolution, 28*(8), 474-481. doi: 10.1016/j.tree.2013.05.014

Johnson, D. D. P., & Fowler, J. H. (2011). The evolution of overconfidence. *Nature, 477*(7364), 317-320. doi: 10.1038/nature10384

Johnson, N. D., & Mislin, A. A. (2011). Trust games: A meta-analysis. *Journal of Economic Psychology, 32*, 865-889. doi: 10.1016/j.joep.2011.05.007

Jonason, P. K., & Li, N. P. (2013). Playing hard-to-get: Manipulating one's perceived availability as a mate. *European Journal of Personality, 27*(5), 458-469. doi: 10.1002/per.1881

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review, 93*(5), 1449-1475.

Kapur, S. (2003). Psychosis as a state of aberrant salience: A framework linking biology, phenomenology, and pharmacology in schizophrenia. *American Journal of Psychiatry, 160*, 13-23.

Kelleher, I., Keeley, H., Corcoran, P., Lynch, F., Fitzpatrick, C., Devlin, N., . . . Cannon, M. (2012). Clinicopathological significance of psychotic experiences in non-psychotic young people: Evidence from four population-based studies. *The British Journal of Psychiatry, 201*, 26-32. doi: 10.1192/bjp.bp.111.101543

Kimmel, A. J. (1998). In defense of deception. *American Psychologist, 53*(7), 803-805. doi: 10.1037/0003-066X.53.7.803

Kinsey, A. C., Pomeroy, W. B., & Martin, C. E. (1948/1998). *Sexual behavior in the human male*. Philadelphia: W. B. Saunders.

Knack, S., & Keefer, P. (1997). Does social capital have an economic payoff? A cross-country investigation. *The Quarterly Journal of Economics, 112*(4), 1251-1288. doi: 10.1162/003355300555475

Knapp, M. (2005). Costs of schizophrenia. *Psychiatry, 4*(10), 33-35.

Korn, C. W., Sharot, T., Walter, H., Heekeren, H. R., & Dolan, R. J. (2014). Depression is related to an absence of optimistically biased belief updating about future life events. *Psychological Medicine, 44*, 579-592. doi: 10.1017/S0033291713001074

Kühberger, A. (2000). What about motivation? *Behavioral and Brain Sciences, 23*(5), 685.

Kurzban, R. (2011). Two problems with "self-deception": No "self" and no "deception". *Behavioral and Brain Sciences, 34*(01), 32-33. doi: 10.1017/S0140525X10002116

Lamba, S., & Nityananda, V. (2014). Self-deceived individuals are better at deceiving others. *PLOS One, 9*(8), e104562. doi: 10.1371/journal.pone.0104562

Langdon, R., & Bayne, T. (2010). Delusion and confabulation: Mistakes of perceiving, remembering and believing. *Cognitive Neuropsychiatry, 15*(1-3), 319-345. doi: 10.1080/13546800903000229

Langdon, R., Still, M., Connors, M. H., Ward, P. B., & Catts, S. V. (2014). Jumping to delusions in early psychosis. *Cognitive Neuropsychiatry, 19*(3), 241-256. doi: 10.1080/13546805.2013.854198

Langdon, R., Ward, P. B., & Coltheart, M. (2010). Reasoning anomalies associated with delusions in schizophrenia. *Schizophrenia Bulletin, 36*(2), 321-330. doi: 10.1093/schbul/sbn069

LaRocco, V. A., & Warman, D. M. (2009). Probability estimations and delusion-proneness. *Personality and Individual Differences, 47*(3), 197-202. doi: 10.1016/j.paid.2009.02.021

Lee, G., Barrowclough, C., & Lobban, F. (2011). The influence of positive affect on jumping to conclusions in delusional thinking. *Personality and Individual Differences, 50*(5), 717-722. doi: 10.1016/j.paid.2010.12.024

Lench, H. C., Smallman, R., Darbor, K. E., & Bench, S. W. (2014). Motivated perception of probabilistic information. *Cognition, 133*, 429-442. doi: 10.1016/j.cognition.2014.08.001

Lincoln, T. M., Lange, J., Buau, J., Exner, C., & Moritz, S. (2010). The effect of state anxiety on paranoid ideation and jumping to conclusions: An experimental investigation. *Schizophrenia Bulletin, 36*(6), 1140-1148. doi: 10.1093/schbul/sbp029

Lincoln, T. M., Ziegler, M., Lüllmann, E., Müller, M. J., & Rief, W. (2010). Can delusions be self-assessed? Concordance between self- and observer-rated delusions in schizophrenia. *Psychiatry Research, 178*, 249-254. doi: 10.1016/j.psychres.2009.04.019

Lincoln, T. M., Ziegler, M., Mehl, S., & Rief, W. (2010). The jumping to conclusions bias in delusions: Specificity and changeability. *Journal of Abnormal Psychology, 119*(1), 40-49. doi: 10.1037/a0018118

Lindgren, K. P., Parkhill, M. R., George, W. H., & Hendershot, C. S. (2008). Gender differences in perceptions of sexual intent: A qualitative review and integration. *Psychology of Women Quarterly, 32*(4), 423-439. doi: 10.1111/j.1471-6402.2008.00456.x

Low, B. S. (1989). Cross-cultural patterns in the training of children: An evolutionary perspective. *Journal of Comparative Psychology, 103*(4), 311-319.

Madden, G. J., Raiff, B. R., Lagorio, C. H., Begotka, A. M., Mueller, A. M., Hehli, D. J., & Wegener, A. A. (2004). Delay discounting of potentially real and hypothetical rewards: II. Between- and within-subject comparisons. *Experimental and Clinical Psychopharmacology, 12*(4), 251-261. doi: 10.1037/1064-1297.12.4.251

Manapat, M. L., Nowak, M. A., & Rand, D. G. (2013). Information, irrationality, and the evolution of trust. *Journal of Economic Behavior & Organization, 90S*, S57-S75. doi: 10.1016/j.jebo.2012.10.018

Marshall, J. A. R., Trimmer, P. C., Houston, A. I., & McNamara, J. M. (2013). On evolutionary explanations of cognitive biases. *Trends in Ecology & Evolution, 28*(8), 469-473. doi: 10.1016/j.tree.2013.05.013

Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods, 43*(3), 679-690. doi: 10.3758/s13428-010-0049-5

Matthews, R. (2005). Why do people believe weird things? *Significance, 2*(4), 182-184. doi: 10.1111/j.1740-9713.2005.00134.x

McCrone, P., Craig, T. K. J., Power, P., & Garety, P. A. (2010). Cost-effectiveness of an early intervention service for people with psychosis. *The British Journal of Psychiatry, 196*, 377-382. doi: 10.1192/bjp.bp.109.065896

McGuire, L., Junginger, J., Adams Jr., S. G., Burright, R., & Donovick, P. (2001). Delusions and delusional reasoning. *Journal of Abnormal Psychology, 110*(2), 259-266. doi: 10.1037/0021-843X.110.2.259

McKay, R. (2012). Delusional inference. *Mind & Language, 27*(3), 330-355.

McKay, R., & Dennett, D. C. (2009). The evolution of misbelief. *Behavioral and Brain Sciences, 32*(06), 493. doi: 10.1017/s0140525x09990975

McKay, R., & Efferson, C. (2010). The subtleties of error management. *Evolution and Human Behavior, 31*, 309-319. doi: 10.1016/j.evolhumbehav.2010.04.005

McKay, R., Langdon, R., & Coltheart, M. (2006). Need for closure, jumping to conclusions, and decisiveness in delusion-prone individuals. *The Journal of Nervous and Mental Disease, 164*(6), 422-426.

McKay, R., Langdon, R., & Coltheart, M. (2007). Jumping to delusions? Paranoia, probabilistic reasoning, and need for closure. *Cognitive Neuropsychiatry, 12*(4), 362-376.

McKay, R., Mijović-Prelec, D., & Prelec, D. (2011). Protesting too much: Self-deception and self-signaling. *Behavioral and Brain Sciences, 34*(01), 34-35. doi: 10.1017/S0140525X10002608

McNamara, J. M., Stephens, P. A., Dall, S. R. X., & Houston, A. I. (2009). Evolution of trust and trustworthiness: Social awareness favours personality differences. *Proceedings of the Royal Society B, 276*(1657), 605-613. doi: 10.1098/rspb.2008.1182

Mele, A. R. (1997). Real self-deception. *Behavioral and Brain Sciences, 20*, 91-136.

Menon, M., Pomarol-Clotet, E., McKenna, P. J., & McCarthy, R. A. (2006). Probabilistic reasoning in schizophrenia: A comparison of the performance of deluded and nondeluded schizophrenic patients and exploration of possible cognitive underpinnings. *Cognitive Neuropsychiatry, 11*(6), 521-536. doi: 10.1080/13546800544000046

Mijovic-Prelec, D., & Prelec, D. (2010). Self-deception as self-signalling: a model and experimental evidence. *Philosophical Transactions of the Royal Society B: Biological Sciences, 365*(1538), 227-240. doi: 10.1098/rstb.2009.0218

Moritz, S., & Woodward, T. S. (2005). Jumping to conclusions in delusional and non-delusional schizophrenic patients. *British Journal of Clinical Psychology, 44*(2), 193-207. doi: 10.1348/014466505x35678

Moritz, S., Woodward, T. S., & Hausmann, D. (2006). Incautious reasoning as a pathogenetic factor for the development of psychotic symptoms in schizophrenia. *Schizophrenia Bulletin, 32*(2), 327-331. doi: 10.1093/schbul/sbj034

Moritz, S., Woodward, T. S., & Lambert, M. (2007). Under what circumstances do patients with schizophrenia jump to conclusions? A liberal acceptance account. *British Journal of Clinical Psychology, 46*(2), 127-137. doi: 10.1348/014466506x129862

Moutoussis, M., Bentall, R. P., El-Deredy, W., & Dayan, P. (2011). Bayesian modelling of jumping-to-conclusions bias in delusional patients. *Cognitive Neuropsychiatry, 16*(5), 422-447.

Naef, M., & Schunk, D. (2009). *Once bitten, twice shy: On the causal effect of prior experiences on trusting behaviour*. Working paper.

Neuhoff, J. G. (2001). An adaptive bias in the perception of looming auditory motion. *Ecological Psychology, 13*(2), 87-110.

Nicholls, M. E. R., Loveless, K. M., Thomas, N. A., Loetscher, T., & Churches, O. (2014). Some participants may be better than others: Sustained attention and motivation are higher early in semester. *The Quarterly Journal of Experimental Psychology*. doi: 10.1080/17470218.2014.925481

Nowak, M. A., Page, K. M., & Sigmund, K. (2000). Fairness versus reason in the ultimatum game. *Science, 289*(5485), 1773-1775. doi: 10.1126/science.289.5485.1773

Nowlis, S. M., Kahn, B. E., & Dhar, R. (2002). Coping with ambivalence: The effect of removing a neutral option on consumer attitude and preference judgments. *Journal of Consumer Research, 29*(3), 319-334. doi: 10.1086/344431

Nyhan, B., Reifler, J., Richey, S., & Freed, G. (2014). Effective messages in vaccine promotion: A randomized trial. *Pediatrics, 133*(4), e835 -e842. doi: 10.1542/peds.2013-2365

Ochoa, S., Haro, J. M., Huerta-Ramos, E., Cuevas-Esteban, J., Stephan-Otto, C., Usall, J., . . . Brebion, G. (in press). Relation between jumping to conclusions and cognitive functioning in people with schizophrenia in contrast with healthy participants. *Schizophrenia Research*. doi: 10.1016/j.schres.2014.07.026

Ortmann, A., & Hertwig, R. (1998). The question remains: Is deception acceptable? *American Psychologist, 53*(7), 806-807. doi: 10.1037/0003-066X.53.7.806

Ortmann, A., & Hertwig, R. (2002). The costs of deception: Evidence from psychology. *Experimental Economics, 5*, 111-131.

Parco, J. E., Rapoport, A., & Stein, W. E. (2002). Effects of financial incentives on the breakdown of mutual trust. *Psychological Science, 13*(3), 292-297. doi: 10.1111/1467-9280.00454

Perilloux, C., Easton, J. A., & Buss, D. M. (2012). The misperception of sexual interest. *Psychological Science, 23*(2), 146-151. doi: 10.1177/0956797611424162

Perugini, M., Gallucci, M., Presaghi, F., & Ercolani, A. P. (2003). The personal norm of reciprocity. *European Journal of Personality, 17*, 251-283. doi: 10.1002/per.474

Peters, E. (2010). Are delusions on a continuum? The case of religious and delusional beliefs. In I. Clarke (Ed.), *Psychosis and spirituality: Consolidating the new paradigm* (2nd ed.). Chichester, England: John Wiley & Sons.

Peters, E., Joseph, S., Day, S., & Garety, P. A. (2004). Measuring delusional ideation: The 21-item Peters et al. Delusions Inventory (PDI). *Schizophrenia Bulletin, 30*(4), 1005-1022.

Peters, E., Joseph, S., & Garety, P. A. (1999). Measurement of delusional ideation in the normal population: Introducing the PDI (Peters et al. Delusions Inventory). *Schizophrenia Bulletin, 25*(3).

Phillips, L. D., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology, 72*(3), 346-354.

Puri, M., & Robinson, D. T. (2007). Optimism and economic choice. *Journal of Financial Economics, 86*(1), 71-99. doi: 10.1016/j.jfineco.2006.09.003

Raven, J. C., Court, J. H., & Raven, J. (1992). *Manual for Raven's Progressive Matrices and Vocabulary Scales.* Oxford: Oxford Psychologists Press LTD.

Rawlings, D., & Freeman, J. L. (1996). A questionnaire for the measurement of paranoia/suspiciousness. *British Journal of Clinical Psychology, 35*, 451-461.

Rosenboim, M., & Shavit, T. (2012). Whose money is it anyway? Using prepaid incentives in experimental economics to create a natural environment. *Experimental Economics, 15*, 145-157. doi: 10.1007/s10683-011-9294-4

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. F. (1998). Not so different after all: A cross-discipline view of trust. *Academy of Management Review, 23*(3), 393-404. doi: 10.5465/AMR.1998.926617

Rudski, J. (2001). Competition, superstition and the illusion of control. *Current Psychology, 20*(1), 68-84.

Salvatore, G., Lysaker, P. H., Popolo, R., Procacci, M., Carcione, A., & Dimaggio, G. (2012). Vulnerable self, poor understanding of others' minds, threat anticipation and cognitive biases as triggers for delusional experience in schizophrenia: A theoretical model. *Clinical Psychology and Psychotherapy, 19*, 247-259. doi: 10.1002/cpp.746

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decision-making in the ultimatum game. *Science, 300*(5626), 1755-1758. doi: 10.1126/science.1082976

Schotter, A., & Trevino, I. (2014). Belief elicitation in the laboratory. *Annual Review of Economics, 6*, 103-128. doi: 10.1146/annurev-economics-080213-040927

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology, 80*, 1-27.

Schwitzgebel, E. (2014). Belief. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy. Retrieved from http://plato.stanford.edu/archives/spr2014/entries/belief/.

Segerstrom, S. C. (2007). Optimism and resources: Effects on each other and on health over 10 years. *Journal of Research in Personality, 41*, 772-786. doi: 10.1016/j.jrp.2006.09.004

Sevincer, A. T., Wagner, G., Kalvelage, J., & Oettingen, G. (2014). Positive thinking about the future in newspaper reports and presidential addresses predicts economic downturn. *Psychological Science*. doi: 10.1177/0956797613518350

Shah, P. (2012). Toward a neurobiology of unrealistic optimism. *Frontiers in Psychology, 3*(344), 1-2. doi: 10.3389/fpsyg.2012.00344

Shariff, A. F., & Norenzayan, A. (2007). God is watching you: Priming god concepts increases prosocial behavior in an anonymous economic game. *Psychological Science, 18*(9), 803-809. doi: 10.1111/j.1467-9280.2007.01983.x

Sharot, T. (2011a). The optimism bias. *Current Biology, 21*(23), R941-R945. doi: 10.1016/j.cub.2011.10.030

Sharot, T. (2011b). *The optimism bias*. New York, NY: Pantheon Books.

Sharot, T. (2012). Tali Sharot: The optimism bias [Video file]. Retrieved June 26, 2014, from https://www.ted.com/talks/tali_sharot_the_optimism_bias

Sharot, T., Guitart-Masip, M., Korn, C. W., Chowdhury, R., & Dolan, R. J. (2012). How dopamine enhances an optimism bias in humans. *Current Biology, 22*, 1477-1481. doi: 10.1016/j.cub.2012.05.053

Sharot, T., Korn, C. W., & Dolan, R. J. (2011). How unrealistic optimism is maintained in the face of reality. *Nature Neuroscience, 14*(11), 1475-1479. doi: 10.1038/nn.2949

Shepperd, J. A., Klein, W. M. P., Waters, E. A., & Weinstein, N. D. (2013). Taking stock of unrealistic optimism. *Perspectives on Psychological Science, 8*(4), 395-411. doi: 10.1177/1745691613485247

Sheremeta, R. M., & Shields, T. W. (2013). Do liars believe? Beliefs and other-regarding preferences in send-receiver games. *Journal of Economic Behavior & Organization, 94*, 268-277. doi: 10.1016/j.jebo.2012.09.023

Shotland, R. L., & Craig, J. M. (1988). Can men and women differentiate between friendly and sexually interested behavior? *Social Psychology Quarterly, 51*(1), 66-73.

Simmons, J. P., & Massey, C. (2012). Is optimism real? *Journal of Experimental Psychology: General, 141*(4), 630-634. doi: 10.1037/a0027405

Simon, H. A. (1990). Invariants of human behavior. *Annual Review of Psychology, 41*, 1-19.

Smith, D. L. (2011). Aiming at self-deception: Deflationism, intentionalism, and biological purpose. *Behavioral and Brain Sciences, 34*(01), 37-38. doi: doi:10.1017/S0140525X10002657

Smith, V. L. (1991). Rational choice: The contrast between economics and psychology. *Journal of Political Economy, 99*(4), 877-897.

Solberg Nes, L., & Segerstrom, S. C. (2006). Dispositional optimism and coping: A meta-analytic review. *Personality and Social Psychology Review, 10*(3), 235-251. doi: 10.1207/s15327957pspr1003_3

Soll, J. B., & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(3), 780-805. doi: 10.1037/a0015145

Speechley, W. J., Ngan, E. T.-C., Moritz, S., & Woodward, T. S. (2012). Impaired evidence integration and delusions in schizophrenia. *Journal of Experimental Psychopathology, 3*(4), 688-701. doi: 10.5127/jep.018411

Speechley, W. J., Whitman, J. C., & Woodward, T. S. (2010). The contribution of hypersalience to the "jumping to conclusions" bias associated with delusions in schizophrenia. *Journal of Psychiatry and Neuroscience, 35*(1), 7-17.

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. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences, 23*(5), 645-726.

Stanovich, K. E., & West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. *94, 4*(672-695). doi: 10.1037/0022-3514.94.4.672

Sturm, T. (2012). The "rationality wars" in psychology: Where they are and where they could go. *Inquiry, 55*(1), 66-81. doi: 10.1080/0020174X.2012.643628

Taylor, P., Hutton, P., & Dudley, R. (2014). Rationale and protocol for a systematic review and meta-analysis on reduced data gathering in people with delusions. *Systematic Reviews, 3*(44). doi: 10.1186/2046-4053-3-44

Taylor, S. E., Kemeny, M. E., Reed, G. M., Bower, J. E., & Gruenewald, T. L. (2000). Psychological resources, positive illusions, and health. *American Psychologist, 55*(1), 99-109.

Tone, E. B., & Davis, J. S. (2012). Paranoid thinking, suspicion, and risk for aggression: A neurodevelopmental perspective. *Development and Psychopathology, 24*, 1031-1046. doi: 10.1017/S0954579412000521

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: Conjunction fallacy in probability judgment. *Psychological Review, 90*, 293-315.

Tversky, A., & Kahneman, D. (1986). Rational choice and the framing of decisions. *The Journal of Business, 59*(4), S251-S278.

Ubel, P. (2009). Human nature and the financial crisis. *Forbes.* Retrieved February 5th, 2014, from http://www.forbes.com/2009/02/20/behavioral-economics-mortgage-opinions-contributors_financial_crisis.html

Van der Leer, L., Hartig, B., Goldmanis, M., & McKay, R. (in press). Delusion-proneness and jumping to conclusions: Relative and absolute effects. *Psychological Medicine*.

Van der Leer, L., & McKay, R. (2014). "Jumping to conclusions" in delusion-prone participants: An experimental economics approach. *Cognitive Neuropsychiatry, 19*(3), 257-267. doi: 10.1080/13546805.2013.861350

Van Os, J., Hanssen, M., Bijl, R. V., & Ravelli, A. (2000). Strauss (1969) revisited: A psychosis continuum in the general population? *Schizophrenia Research, 45*(1-2), 11-20. doi: 10.1016/S0920-9964(99)00224-8

van Os, J., Linscott, R. J., Myin-Germeys, I., Delespaul, P., & Krabbendam, L. (2009). A systematic review and meta-analysis of the psychosis continuum: Evidence for a psychosis proneness-persistence-impairment model of psychotic disorder. *Psychological Medicine, 39*, 179-195. doi: 10.1017/S0033291708003814

Varki, A., & Brower, D. (2013). *Denial: Self-deception, false beliefs, and the origins of the human mind*. New York, NY: Twelve.

Vlaev, I. (2012). How different are real and hypothetical decisions? Overestimation, contrast and assimilation in social interaction. *Journal of Economic Psychology, 33*, 963-972. doi: 10.1016/j.joep.2012.05.005

von Hippel, W., & Trivers, R. (2011). The evolution and psychology of self-deception. *Behavioral and Brain Sciences, 34*(01), 1-16. doi: 10.1017/s0140525x10001354

Vul, E. (n.d.). Retrieved December 19, 2013, from http://edvul.com/crowdwithin.php

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probablistic representations within individuals. *Psychological Science, 19*(7), 645-647.

Wang, X. T. (2000). Beyond "pardonable errors by subjects and unpardonable ones by psychologists". *Behavioral and Brain Sciences, 23*(5), 699-670.

Warman, D. M. (2008). Reasoning and delusion proneness: Confidence in decisions. *The Journal of Nervous and Mental Disease, 196*(1), 9-15. doi: 10.1097/NMD.0b013e3181601141

Warman, D. M., Lysaker, P. H., Martin, J. M., Davis, L., & Haudenschield, S. L. (2007). Jumping to conclusions and the continuum of delusional beliefs. *Behaviour Research and Therapy, 45*(6), 1255-1269. doi: 10.1016/j.brat.2006.09.002

Warman, D. M., & Martin, J. M. (2006). Jumping to conclusions and delusion proneness: The impact of emotionally salient stimuli. *The Journal of Nervous and Mental Disease, 194*(10), 760-765. doi: 10.1097/01.nmd.0000239907.83668.aa

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology, 20*(3), 273-281. doi: 10.1080/14640746808400161

Weber, J. M., Malhotra, D., & Murnighan, J. K. (2004). Normal acts of irrational trust: Motivated attributions and the trust development process. *Research in Organizational Behavior, 26*, 75-101. doi: 10.1016/S0191-3085(04)26003-8

Weinstein, N. D. (1980). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology, 39*(5), 806-820.

Weinstein, N. D. (1989). Optimistic biases about personal risks. *Science, 246*(4935), 1232-1233.

Weinstein, N. D., & Klein, W. M. P. (1995). Resistance of personal risk perceptions to debiasing interventions. *Health Psychology, 14*(2), 132-140.

Weiss, D. J. (2001). Deception by researchers is necessary and not necessarily evil. *Behavioral and Brain Sciences, 24*(3), 431-432.

White, C. M., & Antonakis, J. (2013). Quantifying accuracy improvement in sets of pooled judgments: Does dialectical bootstrapping work? *Psychological Science, 24*(1), 115-116. doi: 10.1177/0956797612449174

White, L. O., & Mansell, W. (2009). Failing to ponder? Delusion-prone individuals rush to conclusions. *Clinical Psychology & Psychotherapy, 16*(2), 111-124. doi: 10.1002/cpp.607

Whitman, J. C., Menon, M., Kuo, S. S., & Woodward, T. S. (2013). Bias in favour of self-selected hypotheses is associated with delusion severity in schizophrenia. *Cognitive Neuropsychiatry, 18*(5), 376-89. doi: 10.1080/13546805.2012.715084

Whitman, J. C., & Woodward, T. S. (2011). Evidence affects hypothesis judgments more if accumulated gradually than if presented instantaneously. *Psychonomic Bulletin & Review, 18*(6), 1156-1165. doi: 10.3758/s13423-011-0141-6

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing* (2nd ed.). Burlington, MA: Elsevier.

Williams, E. F., & Gilovich, T. (2008). Do people really believe they are above average? *Journal of Experimental Social Psychology, 44*, 1121-1128. doi: 10.1016/j.jesp.2008.01.002

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13*(1), 103-128. doi: 10.1016/0010-0277(83)90004-5

Wischniewski, J., & Brüne, M. (2011). Moral reasoning in schizophrenia: An explorative study into economic decision making. *Cognitive Neuropsychiatry, 16*(4), 348-363. doi: 10.1080/13546805.2010.539919

Woodward, T. S., Buchy, L., Moritz, S., & Liotti, M. (2007). A bias against disconfirmatory evidence is associated with delusion proneness in a nonclinical sample. *Schizophrenia Bulletin, 33*(4), 1023-1028. doi: 10.1093/schbul/sbm013

Woodward, T. S., Munz, M., LeClerc, C., & Lecomte, T. (2009). Change in delusions is associated with change in "jumping to conclusions". *Psychiatry Research, 170*, 124-127.

Young, H. F., & Bentall, R. P. (1995). Hypothesis testing in patients with persecutory delusions: Comparison with depressed and normal subjects. *British Journal of Clinical Psychology, 34*, 353-369.

Zawadzki, J. A., Woodward, T. S., Sokolowski, H. M., Boon, H. S., Wong, A. H. C., & Menon, M. (2012). Cognitive factors associated with subclinical delusional ideation in the general population. *Psychiatry Research, 197*, 345-349. doi: 10.1016/j.psychres.2012.01.004

Ziegler, M., Rief, W., Werner, S.-M., Mehl, S., & Lincoln, T. M. (2008). Hasty decision-making in a variety of tasks: Does it contribute to the development of delusions? *Psychology and Psychotherapy: Theory, Research and Practice, 81*, 237-245. doi: 10.1348/147608308X297104

Zolotova, J., & Brüne, M. (2006). Persecutory delusions: Reminiscence of ancestral hostile threats? *Evolution and Human Behavior, 27*, 185-192. doi: 10.1016/j.evolhumbehav.2005.08.001

# Appendices

Images were presented in colour in the experiments, but have been rendered to greyscale for the appendices.

## Appendix A: Calculation of Optimal Decision Points (Chapter 2) [19]

First, let us introduce some notation:

| Symbol | Description |
|---|---|
| $w$ | The event that the next fish caught is white |
| $b$ | The event that the next fish caught is black |
| $W$ | The event that the true lake is White |
| $B$ | The event that the true lake is Black |
| $n_w$ | The number of white fish caught so far |
| $n_b$ | The number of black fish caught so far |
| $\Delta$ | $= n_w - n_b$ |
| $\Delta'$ | = The value of $\Delta$ after catching one more fish |
| $l$ | The event that the next fish is of the currently leading fish colour<br>( $l = w$ if $n_w > n_b$ and $l = b$ if $n_w < n_b$ ) |
| $L$ | The event that the true lake is the currently leading lake<br>( $L = W$ if $n_w > n_b$ and $L = B$ if $n_w < n_b$ ) |
| $p$ | $= \Pr(w \mid W) = \Pr(b \mid B) > 0.5$ |
| $\rho$ | $= p/(1-p) > 1$ |
| $\pi$ | The probability of making a correct guess if guessing now |
| $c$ | The cost of seeing one more fish |
| $R$ | The reward for a correct guess |

[19] This material has been created by Dr. Maris Goldmanis (Department of Economics, RHUL).

### *Dynamic Task*

Suppose that $n$ fish have been caught so far, of which $n_w$ are white and $n_b$ are black. We are interested in the probability that the true lake is White, conditional on having caught $n_w$ white fish: $\Pr(W\,|\,(n_w,n_b))$ (note that $\Pr(B\,|\,(n_w,n_b)) = 1 - \Pr(W\,|\,(n_w,n_b))$). We will find this by Bayes' Rule:

$$\Pr(W\,|\,(n_w,n_b)) = \frac{\Pr((n_w,n_b)\,|\,W)\Pr(W)}{\Pr((n_w,n_b)\,|\,W)\Pr(W) + \Pr((n_w,n_b)\,|\,B)\Pr(B)}.$$

Because we have assumed a diffuse prior (i.e., both lakes are *a priori* equally likely, $\Pr(W) = \Pr(B) = 0.5$), the formula simplifies to:

$$\Pr(W\,|\,(n_w,n_b)) = \frac{\Pr((n_w,n_b)\,|\,W)}{\Pr((n_w,n_b)\,|\,W) + (\Pr(n_w,n_b)\,|\,B)}.$$

The conditional probabilities that $n_w$ of the $n$ fish are white given the type of lake are:

$$\Pr((n_w,n_b)\,|\,W) = p^{n_w}(1-p)^{n_b}\binom{n_w}{n_w+n_b}; \text{ and } \Pr((n_w,n_b)\,|\,B) = (1-p)^{n_w}(p)^{n_b}\binom{n_w}{n_w+n_b}.$$

Inserting these expressions into the formula for $\Pr W(n_w,n_b)$ and dividing through by $p^{n_w}(1-p)^{n_b}$, we finally obtain:

$$\boxed{\Pr(W\,|\,(n_w,n_b)) = \frac{1}{1+\rho^{n_b-n_w}}, \quad \text{where } \rho = \frac{p}{1-p} > 1.} \qquad (1)$$

Note that $\Pr(B\,|\,(n_w,n_b)) = 1 - \Pr(W\,|\,(n_w,n_b)) = 1/\left(1+\rho^{n_w-n_b}\right)$. Because $p > 0.5$, $\rho > 1$, so that $\Pr(W\,|\,(n_w,n_b)) > \Pr(B\,|\,(n_w,n_b))$ if and only if $n_w > n_b$. Therefore, if the decision maker decides to make a guess, she should always guess the lake corresponding to the most fish caught so far, and the probability of a correct guess is:

$$\pi = \begin{cases} \Pr(W \mid (n_w, n_b)) = \dfrac{1}{1+\rho^{n_b - n_w}} & \text{if } \mathrm{n_w} > \mathrm{n_b}; \\ \Pr(B \mid (n_w, n_b)) = \dfrac{1}{1+\rho^{n_w - n_b}} & \text{if } \mathrm{n_b} > \mathrm{n_w}; \\ \Pr(W \mid (n_w, n_b)) = \Pr(B \mid (n_w, n_b)) = \dfrac{1}{2} & \text{if } \mathrm{n_w} = \mathrm{n_b}. \end{cases}$$

Note that this simplifies to the following extremely simple rule, where the probability of a correct guess depends only on the absolute value of the *difference* of the numbers of white and black fish caught so far:

$$\boxed{\pi(\Delta) = \frac{1}{1+\rho^{-\Delta}}, \quad \text{where } \Delta = n_w - n_b.} \tag{2}$$

It follows that the only relevant state variable for our problem is $\Delta$, and we can write the value function as $V(\Delta)$. Here $V(\Delta)$ is the expected value to the decision maker of having observed $\Delta$ more fish of one color than of the other. As in any stopping-time problem, this value is the maximum of (1) the expected value of guessing immediately and (2) the expected option value of seeing one more fish.

The only missing element remaining in the formulation of this problem is the state transition matrix. This, however, is easy to find. Clearly, if one more fish is caught, $\Delta$ will change to either $\Delta+1$ or $\Delta-1$. Furthermore, if $\Delta = 0$, the change will be to $\Delta' = 1 = \Delta+1$ with probability one. If $\Delta > 1$, the probability of it changing to $\Delta+1$ is simply the probability of getting one more fish of the currently leading color:

$$\Pr(\Delta' = \Delta+1) = \Pr(l \mid L) + \Pr(l \mid \neg L) = p\pi(\Delta) + (1-p)(1-\pi(\Delta)).$$

To summarise:

$$\boxed{\Pr(\Delta' = \Delta + 1) = \begin{cases} 1 & \text{if } \Delta = 0; \\ p\pi(\Delta) + (1-p)(1-\pi(\Delta)) & \text{otherwise.} \end{cases}} \tag{3}$$

Now, we are ready to formulate the Bellman equation for the optimal stopping time problem. The expected value of guessing now is $\pi(\Delta) \cdot R$. The expected value of drawing one more fish is:

$$\Pr(\Delta' = \Delta + 1)V(\Delta + 1) + (1 - \Pr(\Delta' = \Delta + 1))V(\Delta - 1) - c\,.$$

The value function is therefore defined recursively by:

$$\boxed{V(\Delta) = \max\{\pi(\Delta)R; \Pr(\Delta' = \Delta + 1)V(\Delta + 1) + (1 - \Pr(\Delta' = \Delta + 1))V(\Delta - 1) - c\}.} \tag{4}$$

This equation is easy to solve by value function iteration. It can also be proven analytically that the stopping rule will always take the form "Stop if and only if $\Delta > \Delta_0$" for some $\Delta_0$.

### *Static Task*

Let the observer have a uniform prior over lakes:

$$\Pr(\mathrm{W}) = \Pr(\mathrm{B}) = \frac{1}{2}.$$

Now, suppose we have sampled $n$ fish and found that $n_w \leq n$ of them are white (so that $n_b = n - n_w$ are black). What is our best guess of a lake? Given that the prior probabilities of both lakes are the same, it is clear that we should guess "Lake White" if we have sampled more white than black fish ($n_w > n_b = n - n_w$) and "Lake Black" if we have sampled more black than white fish ($n_w < n_b = n - n_w$). If we have observed $n_w = n_b = n/2$, the sample gives us no information, so we can make either guess, and it will be correct with probability 1/2.

Given this decision rule, what is our probability of making a correct guess based on a sample of $n$ fish? Clearly, for an odd $n$ this is simply the probability that we get more fish of the "correct" than of the "incorrect" colour (where the "correct" colour is white if the true lake is Lake White and black if the true lake is Lake Black). For an even $n$, we need to add to this one half of the probability that we draw equal numbers of fish of both colours. To calculate these quantities, we simply note that we can get $n_c$ of the $n$ fish in the correct color in $\binom{n}{n_c}$ ways, and each of these occurs with probability $p^{n_c}(1-p)^{n-n_c}$, so that the total probability of getting $n_c$ of the $n$ fish in the "correct" color is

$$\Pr(n_c \text{of the } n \text{ fish are of the correct color}) = \binom{n}{n_c} p^{n_c}(1-p)^{n-n_c}.$$

Thus the probability of a correct decision for any odd $n$, i.e., for any $n = 2k+1$, where $k$ is a natural number, is

$$\Pr(\text{correct decision with sample size } n = 2k+1) = \sum_{n_c=k+1}^{n} \binom{n}{n_c} p^{n_c}(1-p)^{n-n_c}.$$

The probability of a correct decision for any even n, i.e., for any $n = 2k$, where $k$ is a natural number, is

$$\Pr(\text{correct decision with sample size } n = 2k)$$
$$= \sum_{n_c=k+1}^{n} \binom{n}{n_c} p^{n_c}(1-p)^{n-n_c} + \frac{1}{2}\binom{n}{k} p^{k}(1-p)^{k}.$$

Suppose that the cost is c per fish, while the reward for a correct guess is R. Then the total expected payoff after sampling $n$ fish is:

$$E[\Pi(n)] = R \cdot \Pr(\text{correct decision with sample size } n) - c \cdot n.$$

The optimal sample size maximises this expression:

$$n^* = \arg\max_n E[\Pi(n)]\,.$$

An even $n$ is never optimal: Intuitively, increasing the sample size by one from any odd number $n = 2k + 1$ can never be optimal, since adding the $(2k + 2)^{th}$ fish can never meaningfully change the optimal decision. Mathematically, we can show that for any $k$:

$$\begin{aligned}&\Pr(\text{correct decision with sample size } n = 2k + 1)\\ &\qquad = \Pr(\text{correct decision with sample size } n = 2k + 2),\end{aligned}$$

so that $E[\Pi(2k + 2)] - E[\Pi(2k + 1)] = -c < 0$. The problem thus reduces to

$$n^* = \arg\max_{n=2k+1} \left( R \times \sum_{n_c=k+1}^{n} \binom{n}{n_c} p^{n_c}(1-p)^{n-n_c} - c \times n \right).$$

To illustrate, Figure 1 shows the expected payoff $\Pi[n]$ as a function of sample size $n$ for parameter values $p = 0.8; R = 25; c = 0.5$. In this example, the optimal sample size is $n^* = 5$. Figure 2 shows the optimal sample size as function of the parameter $p$ when the other two parameters are fixed at $R = 10$ and $c = 0.1$.

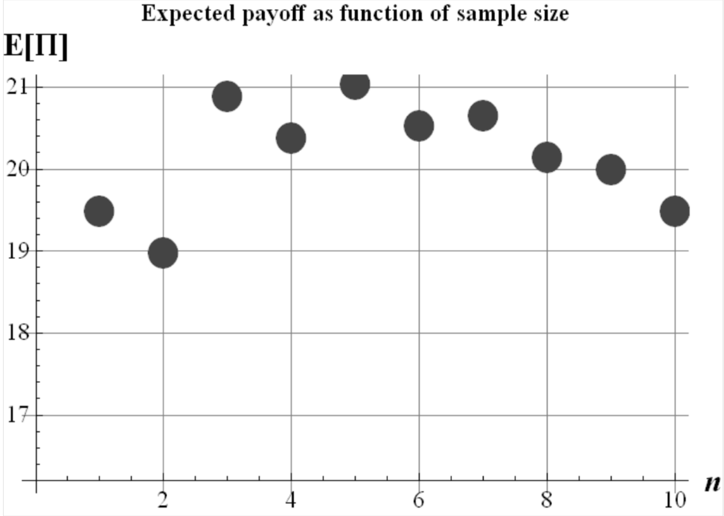


**Figure 1** Expected payoff as a function of sample size when p=0.8, R=25, and c=0.5

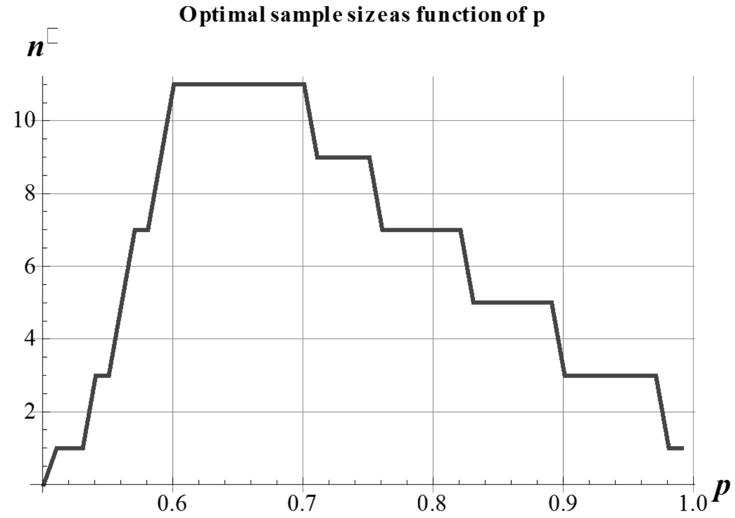


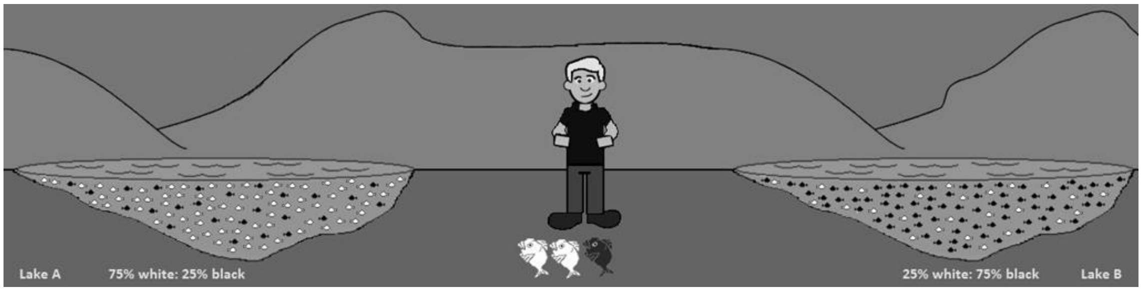
**Figure 2** Optimal sample size as a function of p when R=10 and c=0.1

# Appendix B: Instructions and Comprehension Questions (Chapter 2)

## *Instructions*

### First Task

In each round you will see a fisherman and two lakes: Lake A on the left and Lake B on the right (which lake is which is indicated on the screen; see an example of a screen shot below). Both lakes contain black and white fish. The proportion of black to white fish in each lake is always either 75:25 or 25:75 and is always indicated below the lake. The relative proportion of black and white fish in one lake is always exactly opposite to that in the other lake. The lakes have so many fish in them, that the fishing of several fish does not affect the overall proportions.



Each time he goes fishing, the fisherman flips a coin. If the coin shows heads, he goes to Lake A, if it shows tails, he goes to Lake B. In other words, he is equally likely to go to each lake. However, when he visits a lake, he always stays there until he has caught at least twelve fish, although he might stay and catch more. He never visits both lakes on the same trip.

In this task, you will get the opportunity to see the fisherman's collection of caught fish for seven different fishing trips (rounds). He will wear a shirt with a different colour on each trip, just to make it clear it is a different trip. The ratios in the two lakes may alternate on different trips. Your task is to guess which lake the fisherman went to on each trip. To do that, you will be shown one fish at a time. After seeing a fish you will be given the choice to decide on either Lake A or B or to see more fish before deciding on a lake. The fisherman always catches at least twelve fish, but often catches more. You can see as many fish as you want before making your decision about which lake the fisherman has been fishing from, up to the last fish, after which you will have to make a decision between the lakes.

If you choose the correct lake, you receive 100 points, but 2 points are deducted for each additional fish after the first one that you ask to see. If you are wrong when you decide which lake the fisherman has been fishing from, you will receive zero points, regardless of the number of fish you have seen (i.e. you will not lose points).

IMPORTANT: At the end of the experiment, we will pay you for *one* round in which the lake and the fish are randomly drawn. All points you receive in this round will be converted into pounds at a rate of 1 point = £0.05. These earnings will be paid to you in addition to your show-up fee of £6 at the end of the experiment. You will not know in advance for which round you will be paid, so it is in your interest to treat each round as if it would determine your earnings.

Each time you see a fish, we will also ask you to rate your confidence that the fish are coming from either lake. You will be asked to do this regardless of whether you chose to see another fish or decided on a lake. After you have provided your confidence ratings, you will move on to the next fish if you chose to see another fish or to the next round if you decided on a lake. Even though it does not influence your earnings, we ask you to state your confidence deliberately and truthfully (you could consider your show-up fee a payment for this).

You have now finished reading the instructions for the first task; please complete the questions on the computer to ensure that you have fully understood the instructions.

**Second Task**

In this second task, you will again see the fisherman in new fishing trips. The proportion of black and white fish in each lake is still 75:25 or vice versa, and the proportions are still opposite in the two lakes. As before, the fisherman only goes to one lake each time.

As before, you will be able to earn points for choosing the correct lake and points will be deducted the more fish you see. The number of points you can receive for a correct decision differs each time, but is indicated on the screen.

In this task, you will need to indicate *in advance* how many fish you would like to see in total. You will then be shown the number of fish you requested all at once. You will not be able to request to see additional fish afterwards. After seeing your requested number of fish, you again decide on a lake. Finally, you will again indicate your confidence level in each lake.

As in the first task, you will be paid for one of the five rounds in which the lake and fish are chosen randomly. This amount will be paid in addition to your show-up fee and to what you earned in the first task, again at a rate of 1 point = £0.05. You will not know in advance for which round you will be paid, so it still is in your interest to treat each round as if it would determine your earnings.

### *Comprehension Questions*

[Correct answers are given in square brackets for clarity; these were not shown on the actual paper sheets.]

**First task**

1) If Lake A contains 25% white fish, what percentage of the fish in Lake A will be black? 1) [75%]

2) If Lake B contains 75% white fish, what percentage of the fish in Lake A will be white? 2) [25%]

3) Which of the following is not possible: 3) [B]

   a) Lake A (75% black: 25% white); Lake B (25% black: 75% white)

   b) Lake A (25% black: 75% white); Lake B (25% black: 75% white)

   c) Lake A (25% black: 75% white); Lake B (75% black: 25% white)

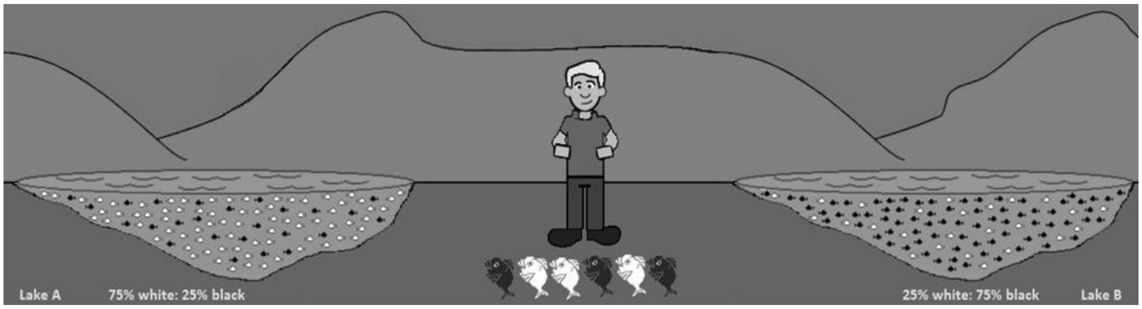
4) Which of the following is not possible: 4) [A]

   a) Lake A (20% black: 80% white); Lake B (80% black: 20% white)

   b) Lake A (25% black: 75% white); Lake B (75% black: 25% white)

   c) Lake A (75% black: 25% white); Lake B (25% black: 75% white)

5) Which of the following is not possible in the scenario below: 5) [C]



a) All the fish are from Lake A

b) All the fish are from Lake B

c) The black fish are from Lake A, the white fish are from Lake B

6) If the second fish caught on a certain trip is from Lake A, the last fish caught on that trip must also be from Lake A. 6) [A: True]

a) True

b) False

7) The fisherman visits Lake A more often than Lake B. 7) [B: False]

a) True

b) False

8) On any given trip the fisherman can catch a maximum of 20 fish, 10 from Lake A and 10 from Lake B. 8) [B: False]

a) True

b) False

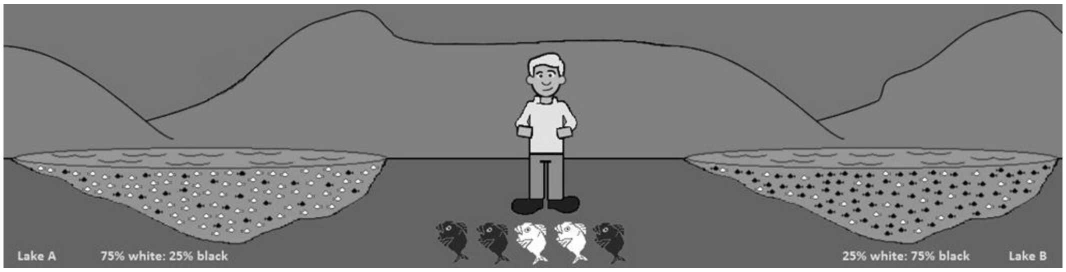
9) The fewer fish you choose to see, the fewer points you will earn if you choose the correct lake. 9) [B: False]

a) True

b) False

**Second task**

10) Which of the following is possible in this scenario: 10) [B]



a) Some of the fish are from Lake A, some from Lake B

b) All the fish are from Lake A

c) The black fish are from Lake B, the white fish are from Lake A

11) If you choose to see seven fish, these fish will be shown to you once at a time and after each fish you will have the option of deciding on a lake or waiting for the next fish. 11) [B: False]

a) True

b) False

# Appendix C: Instructions and Comprehension Questions (Chapter 3)

Below are the instructions and comprehension questions for the lakes-and-fish task. Differences in instructions and in questions for the betting and control groups are noted in square brackets. Correct answers are also given in italics in square brackets for clarity; these were not shown on the actual paper sheets.

## *Instructions*

For this study, you will be required to first read these instructions and complete some comprehension questions, and then complete some practice trials, followed by 26 experimental trials. At the end we will ask you to fill out some questionnaires. Please do not access any other programmes, such as internet browsers, or you will not be paid for the session. In each of the trials you will see two lakes: Lake A (always on the left) and Lake B (always on the right).

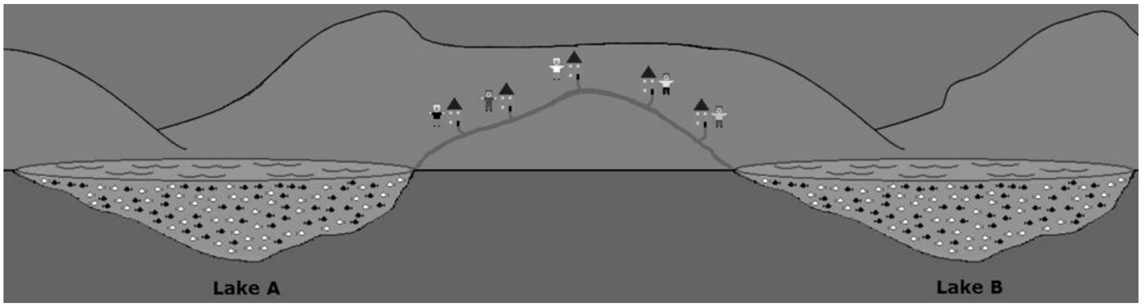


Figure 1

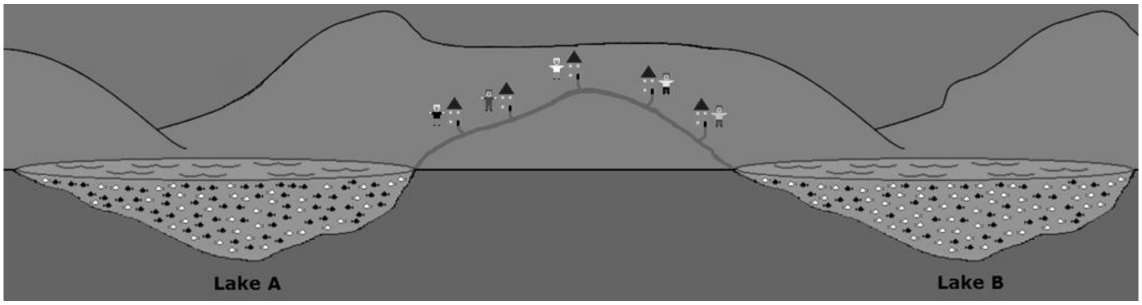


Figure 2

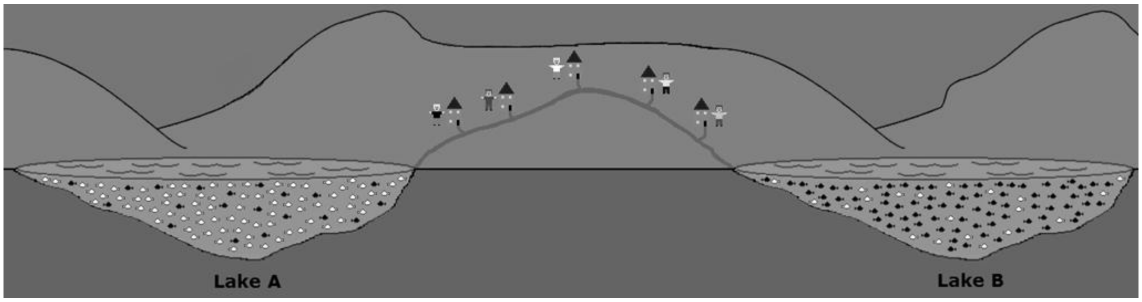
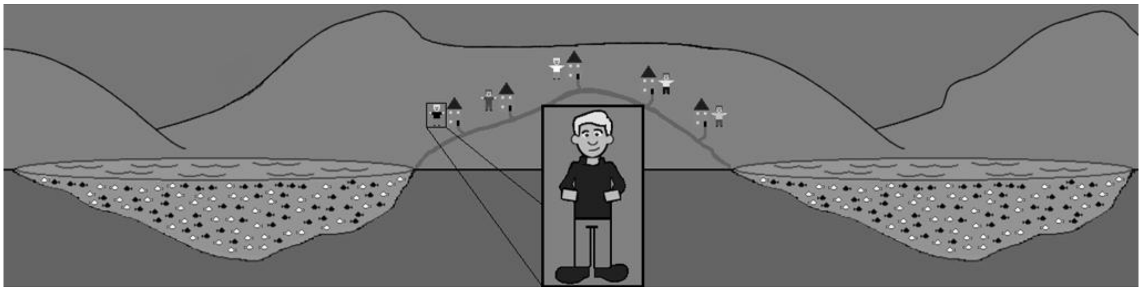


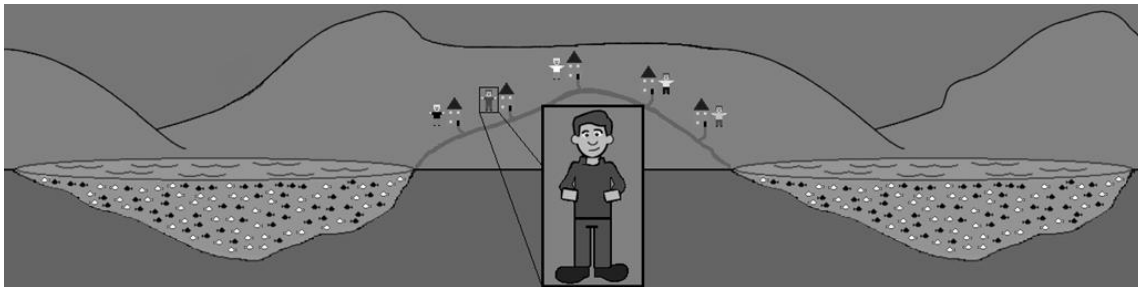
Figure 3

Both lakes contain black and white fish, in opposite proportions. The proportions in each lake differ with the seasons. Sometimes, as in Figure 1 above, equal numbers of black and white fish are found in each of the lakes (50 black: 50 white). At other times, as in Figure 2 above, one of the lakes has slightly more black than white fish (60 black: 40 white), and the other lake has slightly more white than black fish (40 black: 60 white). At yet other times, as in Figure 3 above, one of the lakes has many more black than white fish (85 black: 15 white), and the other lake has many more white than black fish (15 black: 85 white). The relative proportions of black and white fish in Lake A are always exactly opposite to those in Lake B.

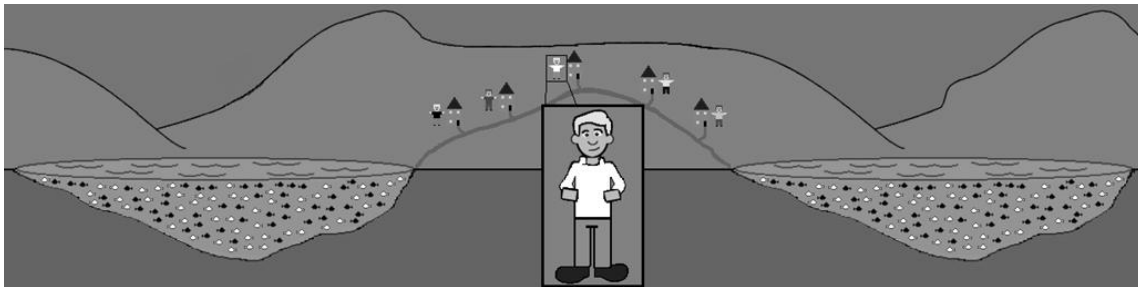
In the background you see the houses of five fishermen. All these fishermen like to fish for a hobby and they always return any fish they catch back to the lake. They all fish six days a week, but stay at home on Sunday. As they all value convenience but also like a change of scenery every now and then, the number of times each fisherman visits a given lake is directly proportional to how close he lives to it. Hence, the closer a fisherman lives to a lake, the more often he visits that lake.



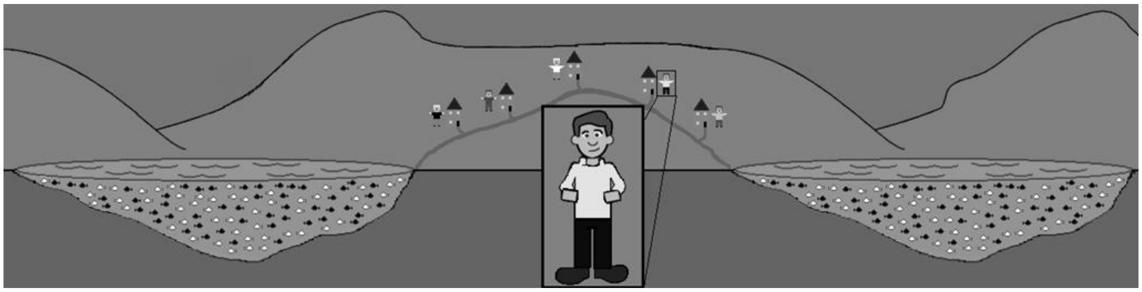
Here you see John, who lives closest to Lake A. He visits Lake A on 5 out of 6 fishing days and he visits Lake B on 1 out of 6 fishing days.



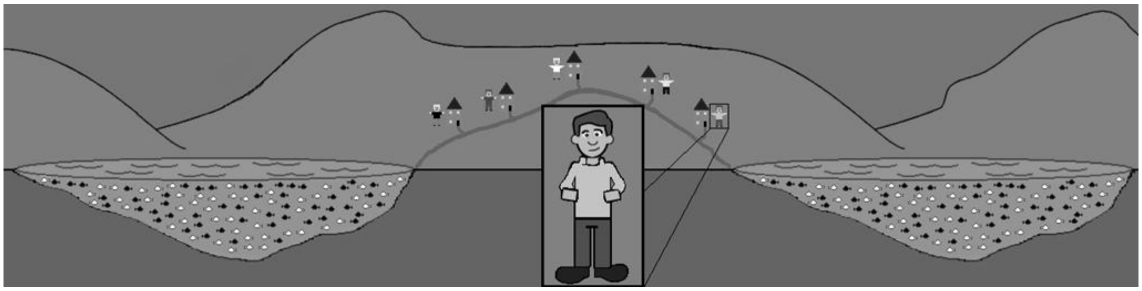
Here you see Paul, who lives closer to Lake A than to Lake B. He visits Lake A on 4 out of 6 fishing days and he visits Lake B on 2 out of 6 fishing days.



Here you see Bob, who lives halfway between Lake A and Lake B. He visits Lake A on 3 out of 6 fishing days and he visits Lake B on 3 out of 6 fishing days.



Here you see Luke, who lives closer to Lake B than to Lake A. He visits Lake A on 2 out of 6 fishing days and he visits Lake B on 4 out of 6 fishing days.



Here you see Mark, who lives closest to Lake B. He visits Lake A on 1 out of 6 fishing days and he visits Lake B on 5 out of 6 fishing days.

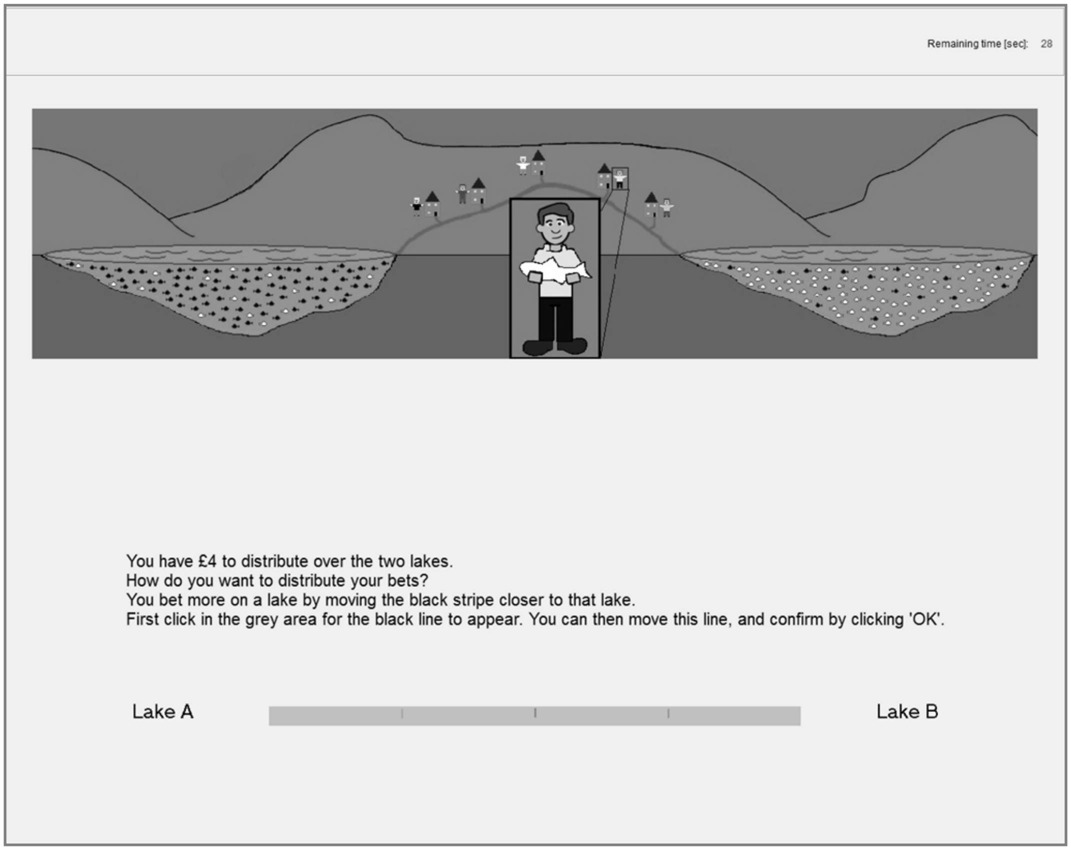
In each trial you will be shown a picture from a randomly picked day of the year (not Sundays). In each picture a fisherman will show you what he caught that day: a black fish or a white fish. Your task is to [*betting group*: bet on/ *control group*: indicate] which lake he was actually fishing from that day. [*betting group*: For each trial, you get £4 to distribute over the two lakes (see the example on the next page). One of the trials will be picked at random and you will be paid the amount you bet on the correct lake for that trial as a bonus (i.e. you will receive this money on top of your show-up fee)./ *This information was omitted for the control group*].

It might be difficult to judge which lake the fisherman has been fishing from, but just make the best decision you can based on the information you have, within the time allocated. Each picture is shown for 20 seconds, and once time is up the next trial will begin automatically. If you have not made a decision by the end of the time limit, [*betting group*: you will not win anything if that trial gets chosen/ *control group*: no response will be recorded for that trial].

To enter your response, you first click anywhere along the grey area, so that a black line appears (see example below). This black line can then be placed anywhere along the grey area, to indicate how likely you think either lake is. Press the "OK" button that appears after placing your response to confirm your decision.

**An example**

A trial starts with the presentation of a fisherman, the fish he happens to have caught that day (whether black or white), and the two lakes. The picture makes clear how close the fisherman lives to each of the lakes, and also shows the proportions of black and white fish in each lake on that day. Below this image is the response bar (grey rectangle area), which represents how [*betting group*: your £4.00 will be distributed across the two lakes/ *control group*: likely you think either lake is]. Again, your task is to [*betting group*: bet on which lake/ *control group*: indicate which lake you think] the fisherman was actually fishing from that day. You can make your response by clicking anywhere along this grey rectangle. Darker grey stripes indicate one quarter, half, and three quarters of the grey rectangle.



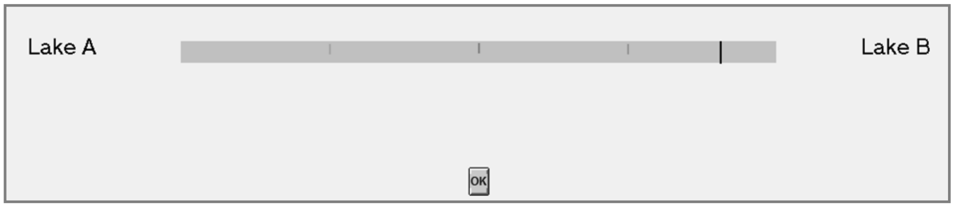
Remaining time [sec]: 28

You have £4 to distribute over the two lakes.
How do you want to distribute your bets?
You bet more on a lake by moving the black stripe closer to that lake.
First click in the grey area for the black line to appear. You can then move this line, and confirm by clicking 'OK'.

[*on-screen text for the control group*: How likely do you think either lake is? You indicate that a lake is more likely by moving the black stripe closer to that lake. First click in the grey area for the black line to appear. You can then move this line, and confirm by clicking “OK”.]

After you’ve taken all information into consideration, you can click somewhere in the grey area to have the black stripe appear there. In case you want to adjust the position of the black stripe, you can do this by dragging it to the correct place. After placing the black stripe, an OK button appears which needs to be clicked to confirm your response.

In the example below, you [*betting group*: have bet £0.40 on Lake A and £3.60 on Lake B (these amounts are not shown, but the fraction of your £4.00 that you bet on each lake is directly proportional to the position of the black stripe inside the grey rectangle). Assuming this trial is chosen for payment, you would win £0.40 if Lake A was the correct answer and £3.60 if Lake B was the correct answer./ *control group*: think Lake B is more likely than Lake A (here you are approximately 90% sure that Lake B is the lake being fished from). These percentages are not shown, but the closer the black stripe is to either end of the grey rectangle, the more likely you think that the relevant lake is.]



### *Comprehension Questions*

1) If Lake A contains 40 black fish, how many white fish will Lake A contain? 1) [*60*]

2) If Lake A contains 85 black fish, how many white fish will Lake B contain? 2) [*85*]

3) Which of the following is not possible: 3) [*C*]

a) Lake A (50 black: 50 white); Lake B (50 black: 50 white)

b) Lake A (15 black: 85 white); Lake B (85 black: 15 white)

c) Lake A (60 black: 40 white); Lake B (60 black: 40 white)

4) Which of the following is not possible: 4) [*B*]

a) Lake A (85 black: 15 white); Lake B (15 black: 85 white)

b) Lake A (15 black: 85 white); Lake B (60 black: 40 white)

c) Lake A (40 black: 60 white); Lake B (60 black: 40 white)

5) Which of the following is not possible: 5) [*B*]

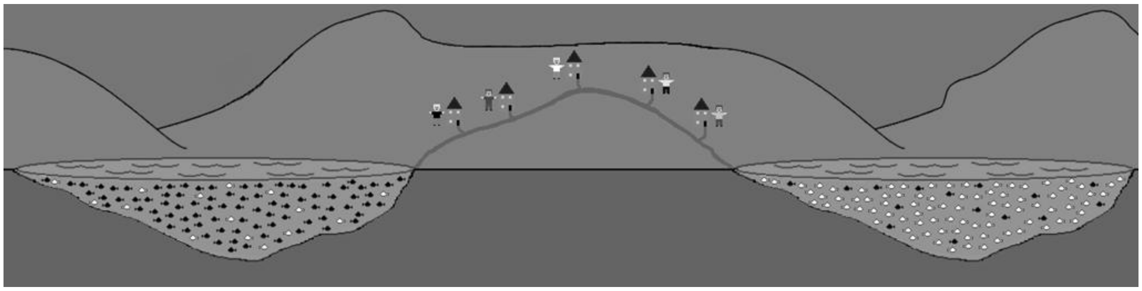
a) Lake A (60 black: 40 white); Lake B (40 black: 60 white)

b) Lake A (90 black: 10 white); Lake B (10 black: 90 white)

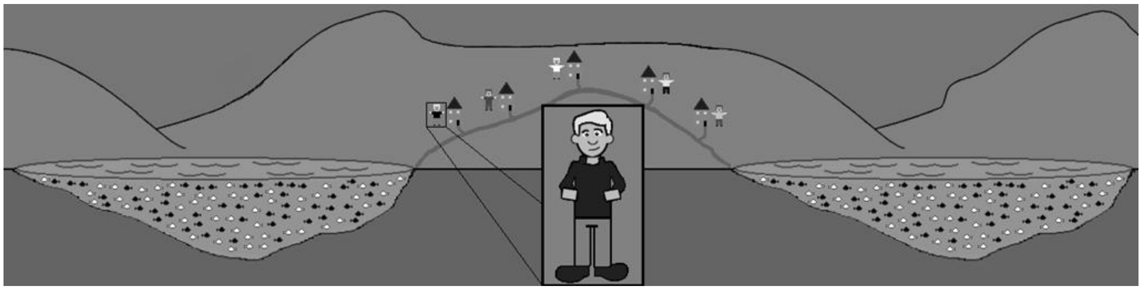
c) Lake A (50 black: 50 white); Lake B (50 black: 50 white)

(hint: the answer is in the second paragraph of the instructions)

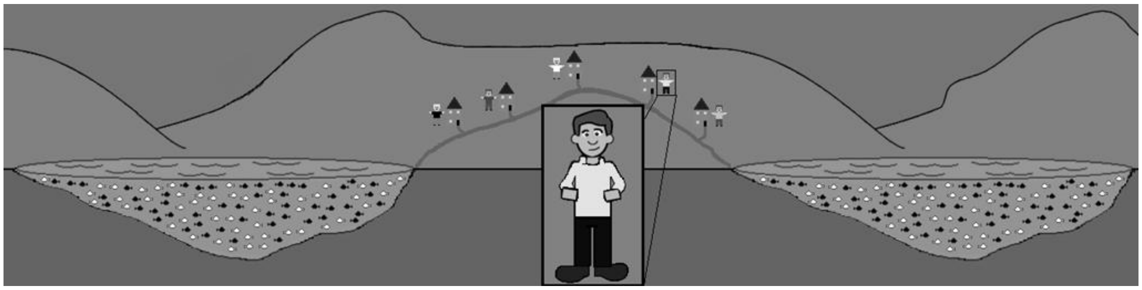
6) How many white fish are in Lake B below? 6) [*85*]



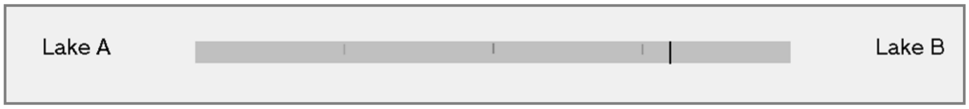
7) Below you see John. How likely is he to visit Lake B? 7) [*1/6*]



8) Below you see Luke. How likely is he to visit Lake B? 8) [*4/6*]



9) True or false: The person below [*betting group*: has bet most of their money on Lake A/ *control group*: thinks Lake A is more likely than Lake B]. 9) [*False*]



# Appendix D: Instructions and Comprehension Questions (Chapter 4)

Instructions and comprehension questions were computerised for the two experiments in Chapter 4. Here, instructions and comprehension questions are presented for the order in which the neutral condition preceded the bias condition. The same type of questions, but for different sources and pieces of information were used in the opposite order. Screenshots of each condition accompany the unformatted instructions and comprehension questions for completeness. Correct answers to questions are underlined in this appendix.

## *Experiments 3a and 3b – Neutral (Jars) Condition*

### *Instructions*

In this task, the computer draws beads from one of two jars. Jar A contains 70% white beads and 30% black beads, while Jar B contains 30% white beads and 70% black beads. These ratios will be displayed on the screen throughout the task. The computer randomly selects one jar; both jars are equally likely to be selected (i.e., the probability for each jar is 50%). After the jar is selected, the computer draws 10 beads from this jar. The computer puts back each drawn bead, but a record of drawn beads is shown.

The computer will show you a series of ten beads drawn from the selected jar. After seeing each bead, you must indicate how likely you think it is that the beads are drawn from Jar A or Jar B. After seeing the tenth bead, you will have to decide which jar the beads were being drawn from. If you are correct, you will win £1; if you pick the wrong jar, you do not get any extra money (i.e., you do not lose money).

You will do this task for four rounds. In three of those four rounds, the computer will show you predetermined sequences. In one of the four rounds, the computer will actually select a jar and draw beads from it as described above. Important: You will only be paid for this round. Since you do not know in advance which of the four rounds you will be paid for, you should treat each round as if you would be paid for it.

You have finished reading the instructions for this task. You will now have to complete a few questions to ensure that you have fully understood the instructions. If you select the wrong answer, a pop-up with the explanation will appear.

*Comprehension Questions*

1) Which of the following is **not possible** in a given round?

a) All the beads are drawn from Jar A.

b) All the beads are drawn from Jar B.

c) The black beads are drawn from Jar A, the white beads are drawn from Jar B.

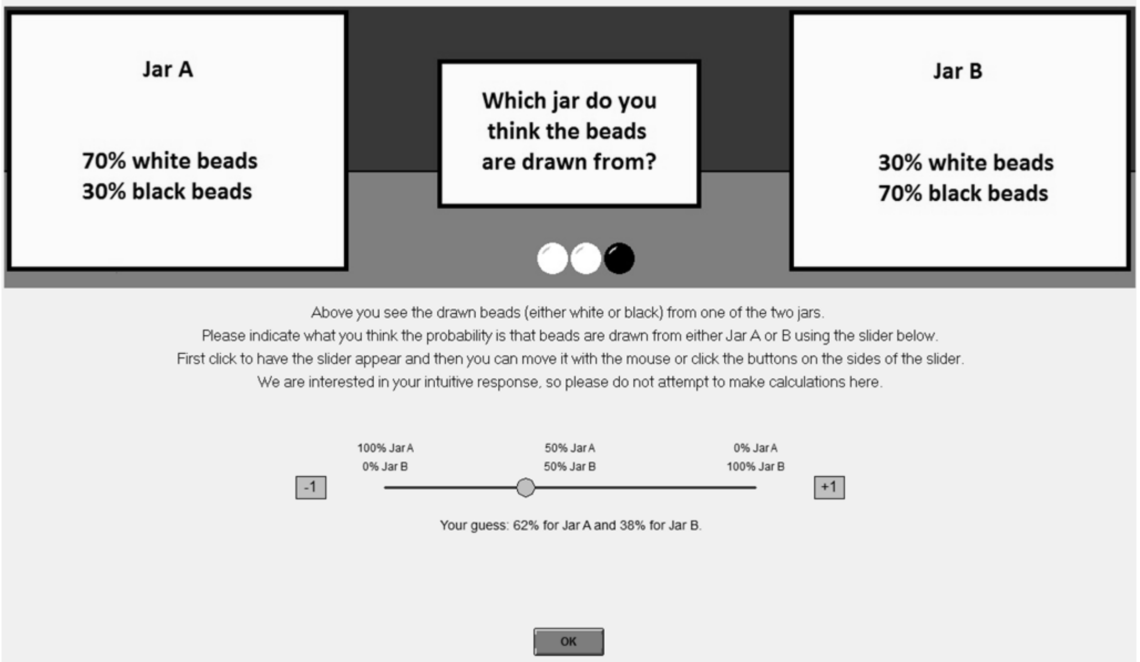
2) True or false?

a) If the second bead is from Jar A, the last bead in the same sequence is also from Jar A. True

b) It is more likely that Jar B, rather than Jar A, is initially selected by the computer. False

c) In the paid round, if you pick the correct jar, you win £1; if you pick the wrong jar, you do not get any extra money. True

*Screenshot*



Jar A

70% white beads
30% black beads

Which jar do you think the beads are drawn from?

Jar B

30% white beads
70% black beads

Above you see the drawn beads (either white or black) from one of the two jars.
Please indicate what you think the probability is that beads are drawn from either Jar A or B using the slider below.
First click to have the slider appear and then you can move it with the mouse or click the buttons on the sides of the slider.
We are interested in your intuitive response, so please do not attempt to make calculations here.

Your guess: 62% for Jar A and 38% for Jar B.

### *Experiments 3a – Bias (Bars) Condition*

*Instructions*

John, Mark, Tom, and Luke each went to a separate bar on Saturday night. John, Mark, Tom, and Luke are all heterosexual, but they each have homosexual friends so on any weekend they are just as likely to visit a gay bar as a straight bar. At the bar on this occasion, all four men flirted with ten women each. All men are equally charming and attractive, but on average they are less successful with the opposite sex in a gay bar compared to a straight bar. This might be because the women are lesbians and not interested in men or perhaps because they were hoping for a girls' night out without being chatted up by men.

Your task is to judge whether each man (i.e., John, Mark, Tom, and Luke) went to a gay bar or to a straight bar, based on women's reactions. On average, men's flirtations with women are reciprocated 70% of the time in a straight bar and are ignored 30% of the time, whereas in a gay bar men's flirtations with women are reciprocated 30% of the time and ignored 70% of the time. These ratios will be displayed on the screen throughout the task. Each man will show you the reactions of the ten women he flirted with: a green check when his flirting is reciprocated,  a red cross when it is ignored. After seeing each reaction you must indicate how likely it is that he went to a gay bar or a straight bar. After seeing the reactions of all ten women, you have to indicate which type of bar he was in. If you are correct, you will win £1; if you are wrong, you do not get any extra money (i.e., you do not lose any money).

You will do this task for four rounds. In three of those four rounds, the computer will show you predetermined sequences of responses. In one of the four rounds, the computer will actually select a bar (each with 50% probability of being selected) and show reactions from women at that bar as described above. Important: You will only be paid for this round. Since you do not know in advance which of the four rounds you will be paid for, you should treat each round as if you would be paid for it.

You have finished reading the instructions for this task. You will now have to complete a few questions to ensure that you have fully understood the instructions. If you select the wrong answer, a pop-up with the explanation will appear.

*Comprehension Questions*

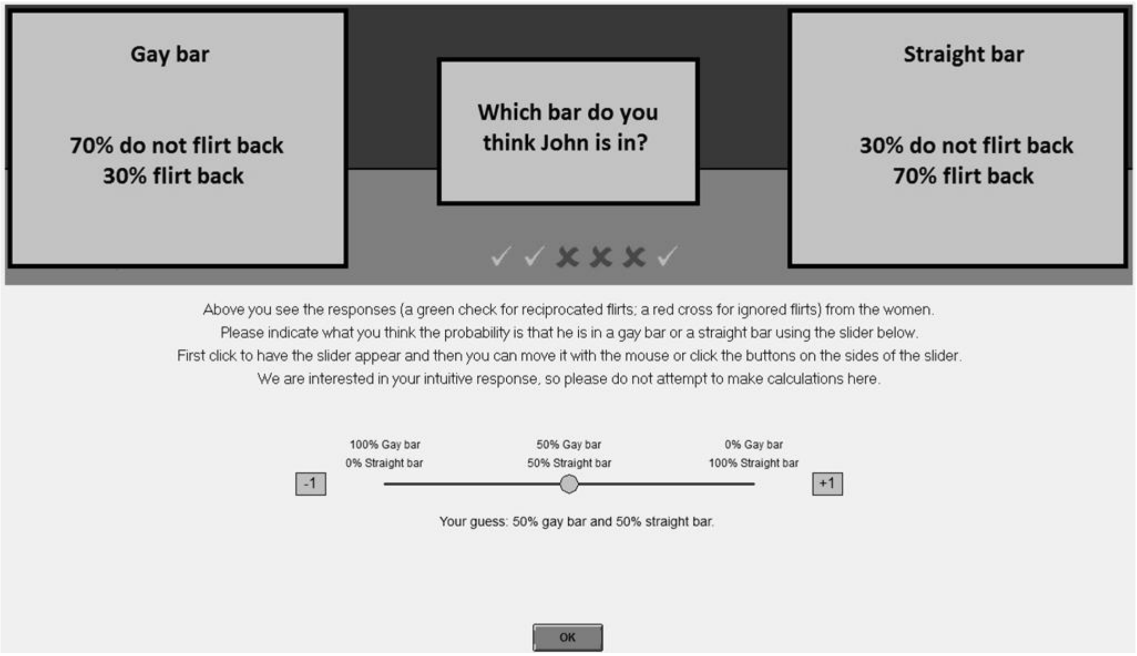
1) True or false?

a) It is more likely that the man is in a straight bar than in a gay bar, as he is heterosexual. False

b) The man went to only one of the two bars. True

c) In the paid round, if you make a correct decision about which type of bar a man visited, you will get £1; if you make an incorrect decision, you do not get any extra money. True

*Screenshot*



Gay bar

70% do not flirt back
30% flirt back

Which bar do you think John is in?

Straight bar

30% do not flirt back
70% flirt back

✓ ✓ ✗ ✗ ✗ ✓

Above you see the responses (a green check for reciprocated flirts; a red cross for ignored flirts) from the women.
Please indicate what you think the probability is that he is in a gay bar or a straight bar using the slider below.
First click to have the slider appear and then you can move it with the mouse or click the buttons on the sides of the slider.
We are interested in your intuitive response, so please do not attempt to make calculations here.

Your guess: 50% gay bar and 50% straight bar.

### *Experiments 3a and 3b – Bias (Dates) Condition*

*Instructions*

John, Mark, Tom, and Luke went speed-dating on Saturday night. Each of them spoke to a series of ten women, and each woman indicated whether or not she wanted to see each of them again for a further date. In the general population men are equally likely to be attractive or unattractive. On average, attractive males score dates 70% of the time and are rejected 30% of the time, whereas unattractive males are successful 30% of the time and rejected 70% of the time. These ratios will be displayed on the screen throughout the task.

Your task is to judge whether each man (i.e., John, Mark, Tom, and Luke) is generally attractive or unattractive to the opposite sex, based on the number of dates he scored. Each man will show you the reactions of the ten women he speed-dated: a green check if they would say “yes” to another date, a red cross if they would say “no” to another date. After seeing each reaction you must indicate how likely it is that he is generally attractive or unattractive. After seeing the reactions of all ten women, you have to indicate whether he is generally attractive or unattractive. If you are correct, you will win £1; if you are wrong, you do not get any extra money (i.e., you do not lose any money).

You will do this task for four rounds. In three of those four rounds, the computer will show you predetermined sequences. In one of the four rounds, the computer will actually select either attractive or unattractive (each with 50% probability of being selected) and show reactions from the women as described above. Important: You will only be paid for this round. Since you do not know in advance which of the four rounds you will be paid for, you should treat each round as if you would be paid for it.

You have finished reading the instructions for this task. You will now have to complete a few questions to ensure that you have fully understood the instructions. If you select the wrong answer, a pop-up with the explanation will appear.

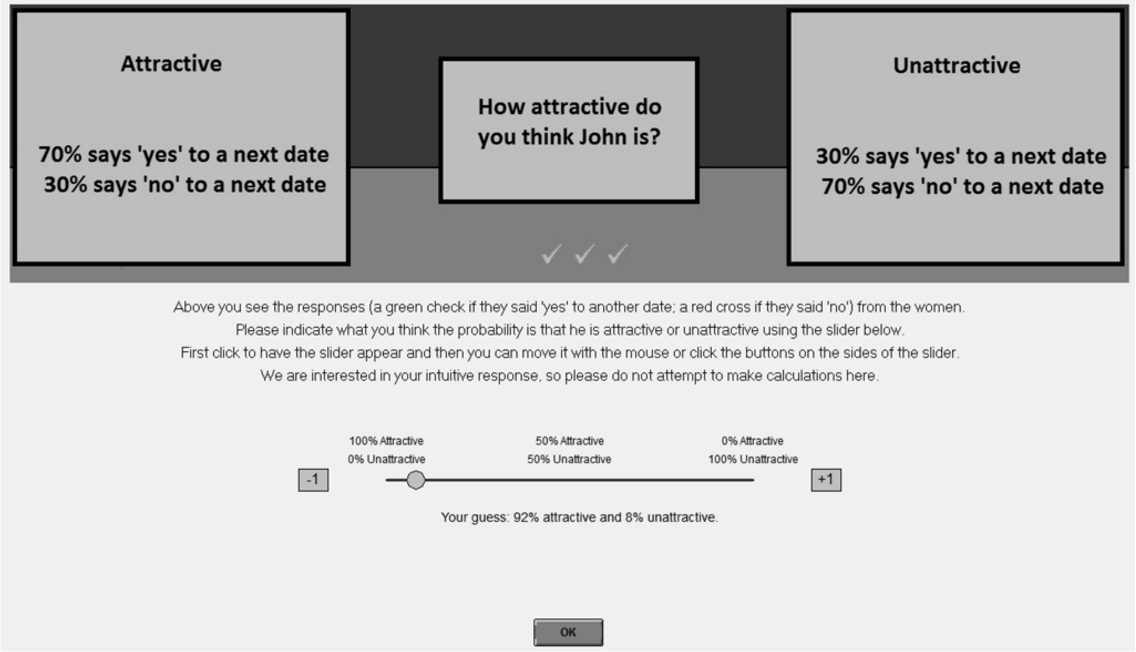
*Comprehension questions*

1) True or false?

a) It is more likely that any man is attractive than that he is unattractive. <u>False</u>

b) In the paid round, if you make a correct decision about whether a man is attractive or unattractive, you will get £1; if you are wrong, you do not get any extra money. <u>True</u>

*Screenshot*



70% says 'yes' to a next date
30% says 'no' to a next date

30% says 'yes' to a next date
70% says 'no' to a next date

Above you see the responses (a green check if they said 'yes' to another date; a red cross if they said 'no') from the women.
Please indicate what you think the probability is that he is attractive or unattractive using the slider below.
First click to have the slider appear and then you can move it with the mouse or click the buttons on the sides of the slider.
We are interested in your intuitive response, so please do not attempt to make calculations here.

Your guess: 92% attractive and 8% unattractive.

# Appendix E: Instructions (Chapter 5)

Instructions were computerised for the study in Chapter 5; here they are shown in appendix format.

### *Round 1 – First own estimates*

In this experiment, we will ask you to estimate answers to various questions. The answers to which your estimates will be compared were taken from official, peer-reviewed or governmental sources (e.g., Office for National Statistics). The answers to some of the questions concern averages for people of your age range and socio-cultural background. However, try to give estimates that would apply to you personally, in your life overall.

You will be paid for your performance on a preselected set of questions in this round (i.e., which questions are paid has been determined beforehand and this does not depend on your performance). You do not know which questions these are, and it is in your best interest to treat all questions as though you are paid for each. The payment scheme is as follows:

- If you estimate the correct answer, you will get 10 points.
- If you are within 2% of the correct answer (on either side), you will get 4 points.
- If you are within 5% of the correct answer (on either side), you will get 2 points.
- If you are within 10% of the correct answer (on either side), you will get 1 point.
- You will not get any points for answers that are further from the correct answer.

It might be difficult to provide the exact answer, but please rest assured that your payment is based on how close your estimate is to the correct answer.

Click OK to start the task.

### *Round 2 – First others' estimates*

We have compiled the answers of all participants in this session to get average estimates to the same questions.

Your task now is to give an estimate of these average estimates of the other participants to the questions shown.

The payment scheme is the same as before. The closer you are to the true average estimate, the more points you will earn. The points from this round will be added to the points from the previous round. Again, a preselected set of questions is paid for.

- If you estimate the correct answer, you will get 10 points.
- If you are within 2% of the correct answer (on either side), you will get 4 points.
- If you are within 5% of the correct answer (on either side), you will get 2 points.
- If you are within 10% of the correct answer (on either side), you will get 1 point.
- You will not get any points for answers that are further from the correct answer.

Click OK to start the task.

### *Round 3 – Second own estimates*

Now, assume your previous personal answers were incorrect. We will ask you to answer the questions again. Your previous estimate will be shown.

For each question, please think about why your previous answers may have been incorrect. Which assumptions and considerations could have been wrong? What do new considerations imply? Was your first estimate too high or too low?

The payment scheme is the same as before. The closer you are to the true estimate, the more points you will earn. The points from this round will be added to the points from the previous rounds. Again, a preselected set of questions is paid for.

- If you estimate the correct answer, you will get 10 points.
- If you are within 2% of the correct answer (on either side), you will get 4 points.
- If you are within 5% of the correct answer (on either side), you will get 2 points.
- If you are within 10% of the correct answer (on either side), you will get 1 point.
- You will not get any points for answers that are further from the correct answer.

If your answers are better in this round, you will be rewarded the points for this round; if your first guess was better, you will be rewarded the points for the first round.

Click OK to start the task.

### *Round 4 – Second others' estimates*

Now, assume your estimates of the average estimates provided in the first round were incorrect.

For each question, please think about why your previous answers may have been incorrect. Which assumptions and considerations could have been wrong? What do new considerations imply? Was your first estimate too high or too low?

The payment scheme is the same as before. The closer you are to the true average estimate, the more points you will earn. The points from this round will be added to the points from the previous rounds. Again, a preselected set of questions is paid for.

- If you estimate the correct answer, you will get 10 points.
- If you are within 2% of the correct answer (on either side), you will get 4 points.
- If you are within 5% of the correct answer (on either side), you will get 2 points.
- If you are within 10% of the correct answer (on either side), you will get 1 point.

- You will not get any points for answers that are further from the correct answer.

If your answers are better in this round, you will be rewarded the points for this round; if your first guess was better, you will be rewarded the points for the first round.

Click OK to start the task.

## Appendix F: Instructions and Comprehension Questions (Chapter 6)

Instructions were computerised for the study in Chapter 6. Instructions from several screens have been formatted for the appendix; screenshots are shown where relevant. The order depicted here is one where investors first indicated their minimal number of white beads and then their trustworthiness-beliefs.

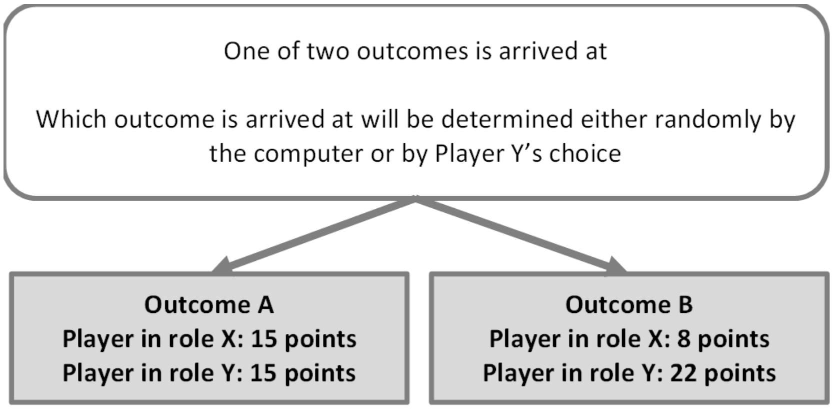
### *General instructions presented to all participants on a paper sheet*

**General overview of the game**

You will be either a player in role X or a player in role Y.

More instructions will follow on the computer screen.

Numbers given are in experimental currency units, where 1 point = £ 0.50.



One of two outcomes is arrived at

Which outcome is arrived at will be determined either randomly by the computer or by Player Y's choice

**Outcome A**
**Player in role X: 15 points**
**Player in role Y: 15 points**

**Outcome B**
**Player in role X: 8 points**
**Player in role Y: 22 points**

### *Computerised instructions to all participants*

During the study, we do not speak of pounds (£). Instead, all earnings are given in points. At the end of the study, all points are transferred into pounds with the following exchange rate: **1 point = 50 pence**.

There are general instructions on the paper by the computer. More instructions will be shown on the screen.

Today's study consists of a single round. In the study, all participants will be either in role X or role Y. Roles are not switched at any point. The computer will randomly assign you either role X or role Y later. Then, the computer will randomly match each participant in role X with a different participant in role Y into a pair. On the next pages, we will explain the decision situation of today's study in more detail.

Each pair of two players will end up with either **outcome A** or **outcome B** (see below). If a pair ends up with outcome A, the player in role X receives 15 points and the player in role Y receives 15 points. If a pair ends up with outcome B, the player in role X receives 8 points and the player in role Y receives 22.

| **Outcome A** | **Outcome B** |
|---|---|
| Player in role X receives 15 points | Player in role X receives 8 points |
| Player in role Y receives 15 points | Player in role Y receives 22 points |

There are two possible methods how the final outcome is determined. Which of these two methods is actually used to determine the final outcome depends on the decisions of the player in role X and other factors. We will explain this later when it becomes relevant.

**Method 1**: The computer randomly chooses one of the two outcomes. The randomly chosen outcome is then paid out at the end of the study.

**Method 2**: The player in role Y chooses one of the two outcomes. The chosen outcome is then paid out at the end of the study.

### *Computerised comprehension questions for all participants*

1: Every participant will play both in role X and in role Y. True/False

2: Each player in role X is matched with how many other players?

- With one player in role Y
- With one other player in role X
- With two players in role Y
- With several players in role Y

3: How can the final outcome (potentially) be determined?

- By the player in role X
- By the player in role Y
- Randomly by the computer

4a: If outcome A is implemented,

how many points does the player in role X receive? 15

4b: If outcome A is implemented,

how many points does the player in role Y receive? 15

5a: If outcome B is implemented,

how many points does the player in role X receive? 8

5b: If outcome B is implemented,

how many points does the player in role Y receive? 22

### *Computerised instructions for trustees*

The computer has randomly determined that **you are in role Y** and it has matched you with one participant in role X.

As mentioned before, there are two methods how a pair's outcome is determined:

**Method 1**: The computer randomly selects one of the two outcomes.

**Method 2**: The player in role Y (i.e. you) selects one of the two outcomes.

At this point, it is not known which of these two methods is used for your pair. That depends on the decision of your matched player in role X and random factors. However, on the next page, we ask you to assume that your choice will determine your pair's outcome and to select one of the two outcomes. If method 1 is used to determine the outcome (i.e. the computer randomly selects an outcome), your choice will not be relevant for payment and player X will not be informed about your decision. If method 2 is used to determine the outcome (i.e. you as the player in role Y determine the outcome), the choice you make on the next screen determines your payoff and the payoff of your matched player in role X. That means if you choose outcome A, the player in role X will receive 15 points and you will receive 15 points. If you choose outcome B, the player in role X will receive 8 points and you will receive 22 points. If you are ready to make your decision, please click ok.

Please select the outcome you want to implement below. Please confirm your decision when you are done.

### *Computerised instructions for investors to set their number of white beads*

The computer has randomly determined that **you are in role X** and it has matched you with one participant in role Y.

As you know, your final outcome will either be determined randomly by the computer or chosen by the player in role Y that you are matched with. Now, it is your task to decide which of the two methods will be used, i.e. you decide if the decision of your matched player in role Y determines your final outcome or if the computer makes a random draw instead. The player in role Y will be informed about which method was used to determine the outcome at the end of the study, but only after he or she has already chosen one of the outcomes. The player will not be informed about how exactly you made the decision between both methods. Before you have to make your decision whether to let the player in role Y choose or to let the computer make a random draw, we will explain what happens depending on your choice on the next pages.

<u>If you decide to let the player in role Y select the outcome</u>

All players in role Y are asked to make a decision between outcome A and outcome B in case this is how their pair's final outcome is determined. If you decide to let the player in role Y select the final outcome, the program simply looks at the decision that your matched player in role Y has made. The outcome that the player has selected will then be implemented and later paid out. That means that if the player has selected outcome A, you receive 15 points and the player in role Y receives 15 points. If the player has selected outcome B, you receive 8 points and the player in role Y receives 22 points. It does not matter for your outcome what other players in role X or in role Y have chosen. Only the decision of your matched player in role Y is relevant for your outcome.

<u>If you decide to let the computer randomly select the outcome</u>

If you decide to let the computer randomly select an outcome, the decision of your matched player in role Y is **NOT** relevant for your outcome. Instead, the computer draws a bead (a small ball) from a container with 1000 white and black beads. If a **white** bead is drawn, **outcome A** will be implemented, meaning that you receive 15 points and your matched player in role Y receives 15 points. If a **black** bead is drawn, **outcome B** will be implemented, meaning you receive 8 points and your matched player in role Y receives 22 points. Again, it does not matter for your outcome what other players in role X or in role Y have decided. Only the colour of the bead that the computer draws for you is relevant for your outcome. If you let the computer decide, you will not be informed about the outcome your matched player in role Y has chosen.

We will further explain how to make your decision between both methods on the next pages.

<u>How to choose the method to determine your outcome</u>

The computer has filled a container with 1000 beads. First, the computer has randomly selected a number between 1 and 1000 and put that many white beads into the container (all numbers from 1 to 1000 are equally likely, i.e., each number has a 0.1% chance of being chosen). Then, the computer filled up the container with black beads until there were 1000 beads in the container in total (i.e. Number of white beads + Number of black beads = 1000).

If you let the computer decide the outcome, the computer will randomly draw one bead from the 1000 beads in the container. Every bead is equally likely to be picked. Remember that a white bead represents outcome A and a black bead represents outcome B.

It is your task to indicate how many white beads you want in the container, **at minimum**, so that you let the computer draw one bead from the container to determine the outcome instead of letting the player in role Y decide. If there are as many or more white beads in the container as the number you gave, the computer will draw one bead to determine the outcome. If there are fewer white beads in the container, the computer will NOT draw a bead and the decision of the player in role Y determines the outcome instead.

For example: Let's say you have indicated that you want at least 591 of the beads to be white. So as long as there are 591 or more white beads in the container, the computer will randomly draw a bead to determine the outcome. As long as there are fewer than 591 white beads in the container, the choice of the player in role Y determines the outcome (and the computer does not pick a bead).

Another example: Let's say you requested at least 753 beads and the container actually holds 814 white and 186 black beads. Then, because 814 is greater than (or equal to) 753, the computer will randomly pick one of the beads to determine the outcome. The probability for outcome A (i.e. that a white bead is drawn) is 814/1000 (81.4%) and the probability for outcome B (i.e. that a black bead is drawn) is 186/1000 (18.6%). If the container holds 165 white and 835 black beads instead, the player in role Y will determine the outcome (and the computer does not pick a bead), because 165 is smaller than 753.

### *Computerised comprehension questions for trustees*

1: Assume you indicated that you want at least 482 of the beads to be white. The container actually holds 390 white beads and 610 black beads. How would the outcome be determined?

- The computer randomly draws a bead from the container
- The player in role Y picks an outcome
- Either the computer draws an outcome or the player in role Y chooses an outcome
- Another player in role X decides

2: The number of white beads in the container is higher than the number you chose. The computer has now drawn a black bead from the container. What is the outcome that will be paid out?

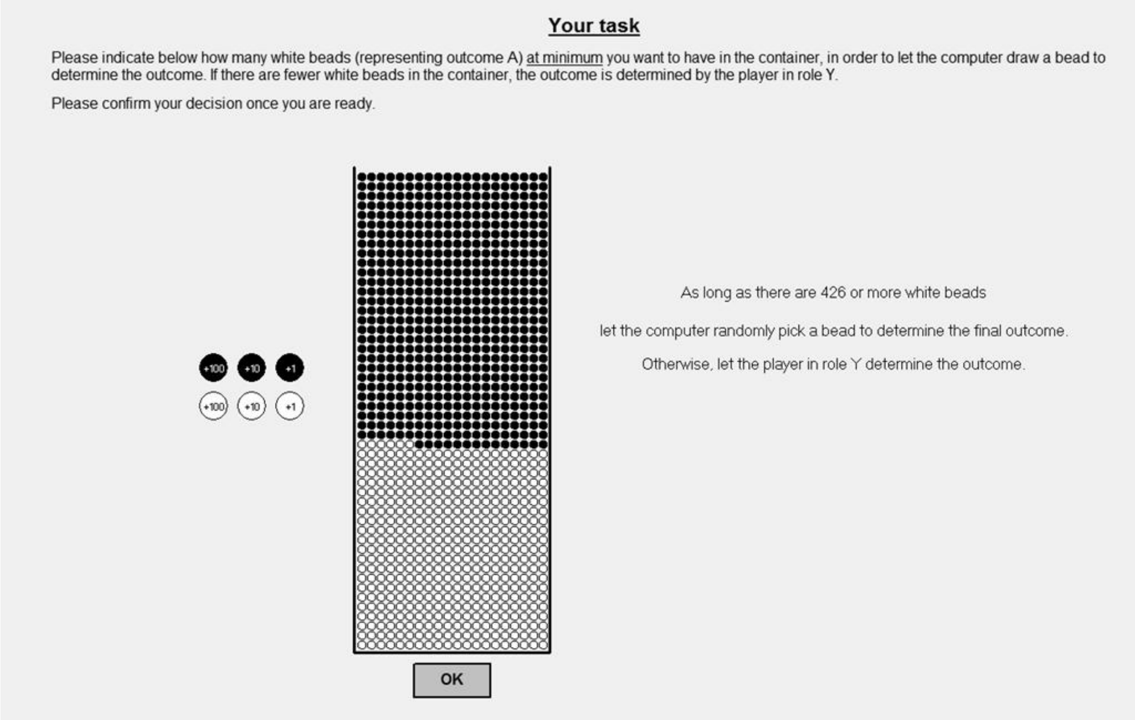
- That depends on the choice made by the player in role Y
- Outcome A
- Outcome B

3: Assume you indicated you want at least 261 beads to be white in order for the computer to determine the outcome. The player in Role Y has picked outcome B. The number of white beads is 836. The computer draws a white bead from the container. What is the outcome that will be paid out?

- Outcome A
- Outcome B
- Either of the outcomes is implemented randomly by the computer

4: Correct or false: The player in role Y is informed about how many beads you want to be white in order to let the computer draw the outcome. False

### *Computerised task to set the minimal acceptable number of white beads*



Your task

Please indicate below how many white beads (representing outcome A) at minimum you want to have in the container, in order to let the computer draw a bead to determine the outcome. If there are fewer white beads in the container, the outcome is determined by the player in role Y.

Please confirm your decision once you are ready.

As long as there are 426 or more white beads

let the computer randomly pick a bead to determine the final outcome.

Otherwise, let the player in role Y determine the outcome.

### *Computerised instructions for investors to state their trustworthiness-beliefs*

Before the study continues, we have another task for you. Please click continue for more information.

There are 8 players in role Y in the room. All of them are asked to make a decision between outcome A and outcome B in case this is how their pair's final outcome is determined. On the next screen, we will ask you to give your best guess about how many of the 8 players in role Y choose outcome A and how many choose outcome B. The players in role Y will not be informed about your guess. **IMPORTANT**: If your guess is correct, you will receive 10 points in addition to anything else you earn during the experiment. If your guess is not correct, you will not receive any additional points and you will not be informed about the correct number.

Please indicate how many players in role Y you think choose outcome A and how many choose outcome B by selecting one of the answers below. Confirm your decision when you are done. Remember that if you guess correctly, you will receive 10 points.

| | | |
|---|---|---|
| **0** players in role Y will pick **outcome A** | **&** | **8 players in role Y will pick outcome B** |
| **1** player in role Y will pick **outcome A** | & | **7** players in role Y will pick **outcome B** |
| **2** players in role Y will pick **outcome A** | & | **6** players in role Y will pick **outcome B** |
| **3** players in role Y will pick **outcome A** | & | **5** players in role Y will pick **outcome B** |
| **4** players in role Y will pick **outcome A** | & | **4** players in role Y will pick **outcome B** |
| **5** players in role Y will pick **outcome A** | & | **3** players in role Y will pick **outcome B** |
| **6** players in role Y will pick **outcome A** | & | **2** players in role Y will pick **outcome B** |
| **7** players in role Y will pick **outcome A** | & | **1** player in role Y will pick **outcome B** |
| **8** players in role Y will pick **outcome A** | & | **0** players in role Y will pick **outcome B** |

### *Computerised instructions for the risk-aversion measure*

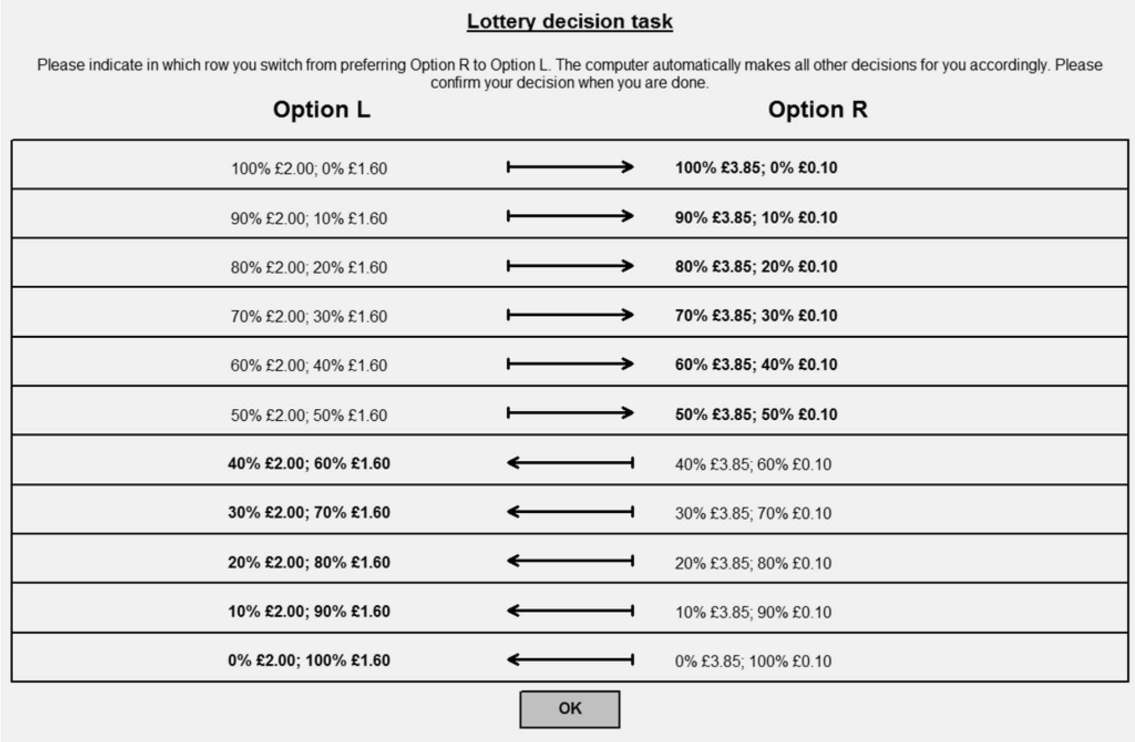
On the next screen, you will see 11 rows. In each row, you have the choice between option L and option R. Both in option L and option R, you can win some amount of money with some probability *x* or another lower amount of money with probability *1-x*. For example, in the third row, if you chose option L, you would receive £2.00 with probability 20% and £1.60 with probability 80%. If you chose option R instead, you would receive £3.85 with probability 20% and £0.10 with probability 80%.

After all participants have made their decisions, the computer plays two lotteries:

**1. Lottery**: The computer randomly selects 1 of the 11 rows (each row is equally likely to be chosen). The option you have selected in this row becomes relevant for your payment.

**2. Lottery**: The computer uses the probabilities of the relevant option in the selected row to randomly pick one of the two possible outcomes of the option.

At the end of the study, ONE participant will be randomly selected by the computer and receive the amount of money earned in this decision situation in addition to everything else earned during the study. When you have read and understood these instructions, please click "continue". If you have any questions, please raise your hand and a study organizer will come to you and answer your question in private.



**Lottery decision task**

Please indicate in which row you switch from preferring Option R to Option L. The computer automatically makes all other decisions for you accordingly. Please confirm your decision when you are done.

| Option L | | Option R |
|---|---|---|
| 100% £2.00; 0% £1.60 | → | **100% £3.85; 0% £0.10** |
| 90% £2.00; 10% £1.60 | → | **90% £3.85; 10% £0.10** |
| 80% £2.00; 20% £1.60 | → | **80% £3.85; 20% £0.10** |
| 70% £2.00; 30% £1.60 | → | **70% £3.85; 30% £0.10** |
| 60% £2.00; 40% £1.60 | → | **60% £3.85; 40% £0.10** |
| 50% £2.00; 50% £1.60 | → | **50% £3.85; 50% £0.10** |
| **40% £2.00; 60% £1.60** | ← | 40% £3.85; 60% £0.10 |
| **30% £2.00; 70% £1.60** | ← | 30% £3.85; 70% £0.10 |
| **20% £2.00; 80% £1.60** | ← | 20% £3.85; 80% £0.10 |
| **10% £2.00; 90% £1.60** | ← | 10% £3.85; 90% £0.10 |
| **0% £2.00; 100% £1.60** | ← | 0% £3.85; 100% £0.10 |