THE DESCRIPTION – EXPERIENCE GAP IN INDIVIDUAL AND SOCIAL DECISIONS UNDER RISK AND UNCERTAINTY

ORESTIS KOPSACHEILIS

Thesis submitted to The University of Nottingham for the Degree of Doctor of Philosophy

SCHOOL OF ECONOMICS
UNIVERSITY OF NOTTINGHAM

September 2018

ABSTRACT

This thesis contributes to the understanding of the 'Description - Experience (DE) gap', which posits that risky decisions depend on the way information about uncertainty is communicated: from description or from experience. The canonical interpretation of findings in this literature suggests that when people make decisions from description, they behave as if overweighting rare events relative to their probability, whereas, when they make decisions from experiential formats, they behave as if underweighting them.

Chapter 1 provides an overview of the main topics and research methodologies presented in Chapters 2, 3 and 4.

Chapter 2 reports the results of an experimental study that provides a cohesive account of the forces behind the DE gap. Our experimental protocol allows us to quantify the effect of each factor in isolation. Moreover, to address methodological concerns in this literature, we employ an elicitation method which allows us to measure these effects both with and without the mediation of a behavioural model that accounts for probability weighting. We find an overall significant DE gap which is equal in size with the literature's average. Despite being mostly driven by an informational asymmetry (sampling bias), other factors pertaining to preferential (ambiguity) and cognitive (likelihood representation and memory) aspects of decision making proved important. Examining the shape and relative position of probability weighting curves we discuss intriguing behavioural implications of our findings.

Chapter 3 focuses on how people search for information in decisions from experience and examines how different search patterns influence ensuing risky choices. In a lab-experiment we find that people search more from options with rarer events. We also find that sampling amount decreases over time periods. Both of these findings become less salient however after the introduction of a history table which records and displays previously sampled outcomes during the lottery evaluation. The cue that dominates the treatment where such a table was present, is the table's maximum capacity. With respect to choices, we elicit and compare probability weighting functions from treatments where decisions were made from experience with a treatment where decisions were made from description. Our treatment comparison reveals evidence for a variant of the canonical interpretation of the gap. We refer to it as the 'relative underweighting hypothesis', which states that rare events in individual risky decisions are overweighted in Experience too, but less so than in Description.

Chapter 4 explores the DE gap in a social context, where we investigate whether the format in which social information is obtained -descriptive or experiential- influences cooperation in social dilemmas and if so, how. We develop and implement in an online experiment, a variation of the prisoner's dilemma game that allows us to observe cooperative responses over a range of likelihoods of cooperation. The likelihoods are communicated either in descriptive or experiential formats. We find that conditional cooperation - the willingness to cooperate if others do the same - is prevalent in our study as cooperation rates increase with the probability of cooperation across treatments. Nonetheless, there are significant differences in the cooperation patterns between Description and Experience. Interestingly, we find evidence that this gap in social decision making, is in the opposite direction from what the canonical interpretation in the individual context would have predicted. Rare events (of cooperation or defection) appear to be more overweighted in Experience rather than in Description. Another asymmetry with the individual domain is that sampling bias, the predominant driver of the DE gap in risky choices, does not affect the gap in the social domain. We conclude that this reversal is due to people being less sensitive towards social information, when they receive this information experientially compared to receiving it descriptively. Lastly, we discuss why such reversals are less likely to occur in individual decision settings.

Chapter 5 provides a summary of the previous chapters' results, identifies their limitations and discusses potentially interesting directions for future research.

PUBLICATIONS

A version of Chapter 3 has been published in *Theory and Decision*:

Kopsacheilis, O. (2018). The role of information search and its influence on risk preferences. *Theory and Decision*, 84(3):311-339.

ACKNOWLEDGMENTS

I am deeply grateful to my thesis supervisors and co-authors of Chapter 2, Chris Starmer and Robin Cubitt. The value of their wise guidance during the development of this thesis cannot be overstated. Our countless stimulating conversations have led to the development of several of the ideas you will read here. A special reference is due to my lifelong academic mentor, Christos Constantatos, for had he not insisted all those years ago to 'give this Economics course a chance', I would probably not be here today.

I want to also thank my friends and co-authors of Chapter 4, Ozan Isler and Dennie van Dolder. Although our joint project is about the 'Description - Experience gap in cooperation', I have happily discovered no such gaps in our collaboration. Also, special thanks go to Lionel Roger, Roelof van Dijk and Jörg Weber for their useful feedback and advice on this work but most importantly, for their friendship and camaraderie.

During my PhD years, my life became considerably easier thanks to receiving the Vice Chancellor's Scholarship for Academic Excellence. Many words of gratitude have to go to the fantastic support staff at the School of Economics for their valuable assistance and especially to Suzanne Robey for her tireless commitment to help. I also enjoyed the privilege of being affiliated with CeDEx[1] and NIBS[2]. These institutions offered me the opportunity to present my work in conferences around the world, attend fascinating talks by leading academics in my field and meet peers with whom I engaged in several memorable discussions. Of these moments, the annual CCC PhD meetings stand out. Participating as a young member of CeDEx, I had the chance to present my work in front of peers from CREED[3] and CBESS[4], receive valuable feedback and meet new, awesome friends like David Smerdon and Sabina Albrecht.

In closing, I want to thank my parents, Alexandra Papagrigoriou and Christos Kopsacheilis, for their unconditional love and endless support. Last but certainly not least, I thank Valeria Burdea, for her insights, careful remarks and selfless assistance that made all this possible. Most of this work was conducted in my attempt to impress her. There is still a long way to go.

1 Centre for Decision Research and Experimental Economics at the University of Nottingham
2 Network for Integrated Behavioural Sciences
3 Centre for Research in Experimental Economics and Political Decision Making at the University of Amsterdam
4 Centre for Behavioural and Experimental Social Science at the University of East Anglia

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1

GENERAL INTRODUCTION

"Hofstadter's Law: It always takes longer than you expect, even when you take into account Hofstadter's Law."

— Douglas Hofstadter, *Gödel, Escher, Bach: An Eternal Golden Braid* (1979)

Uncertainty pervades our world and taming it has been an integral goal of societies throughout the millennia. The task has not been an easy one. For the most part of human history, people would address the question of what tomorrow might bring to oracles and priests rather than to philosophers and scientists. The fact that a formal treatise of the theory of probability appeared only as late as in the 17th century (see Ore, 1960, for a private correspondence between Blaise Pascal and Pierre de Fermat) is perhaps indicative on how counter-intuitive the notion of stochasticity was at the time.

Today, social scientists have at their disposal a wide variety of theoretical tools to study and predict people's risky behaviour. Models such as Expected Utility (EU) theory (Bernoulli, 1954; Savage, 1972; Von Neumann and Morgenstern, 2007) and Cumulative Prospect Theory (CPT; Tversky and Kahneman, 1992) are readily available to help practitioners design health care plans, financial products and insurance policies that are customized to idiosyncratic risk preferences. Notwithstanding, there are still important challenges to overcome. Empirical findings, from the lab and from the field, arise every so often and challenge different aspects of the established theory, forcing the discipline to advance its account of human behaviour. This thesis focuses on one such case, namely, the 'Description - Experience (DE) gap' (Barron and Erev, 2003; Hertwig et al., 2004; Weber et al., 2004), which posits that people's risky choices are - at least partly - dependent on the way information about uncertainty is communicated: from description or from experience.

Over the past years, this discrepancy has been established in experimental studies where participants are assigned to one of two conditions. In the Description condition, they make choices between gambles whose properties (outcomes and outcome-probabilities) are explicitly and completely described in numerical form. Conversely, in the Experience condition, this information is inferred through a sequential sampling process. The key finding relates to the role of rare (low probability) events. According to the canonical interpretation of this finding, people in Description tend to make decisions *as if* overweighting such events, relative to their probability, whereas in Experience they tend to make decisions consistent with underweighting them (Hertwig et al., 2004).

The following chapters present a series of lab and online experiments that address open questions in the DE gap literature. Investigating this phenomenon offers potential benefits that transcend experimental methodological concerns. Throughout the following chapters, we mention instances where the gap could be relevant for phenomena such as insurance policies and tax-compliance.

From a methodological perspective, a key feature of the analysis throughout the next chapters is the ability to estimate a function that monitors subjective reactions to underlying probabilities. This approach is motivated by the recognition that the gap is at its core a phenomenon emerging from differential responses towards probabilistic events. Most of the key results in this thesis therefore, stem from the comparison of these functions across treatments. These treatments vary with respect to how this probability was communicated: through descriptive or experiential formats.

In Chapters 2 and 3, the construction of these functions has benefited considerably from theoretical and empirical work on the 'source method' (Abdellaoui et al., 2011a,b; Tversky and Fox, 1995). This method can be used to map different sources of uncertainty onto distinct probability weighting functions. Defining these sources at the treatment level (i.e. on variations of Description and Experience conditions) and comparing them across treatments, we can analyse our results through the lens of (potentially) differential attitudes towards rare events. In Chapter 4, where the gap is discussed in a context of social uncertainty, this function maps probabilities of cooperative outcomes to average cooperation response functions. The elicitation of these functions is based on the methods presented in Fischbacher et al. (2001).

The following paragraphs summarize the highlights of this analysis. Although the DE gap is a conceptual common denominator, each chapter can be read independently.

In Chapter 2, our goal is to understand the underpinnings of the DE gap. To this end, our aim is to provide a cohesive account of the forces that drive these two domains of decision making (Description and Experience) apart. We propose a taxonomy of the contributing factors of the gap and implement a lab-experiment that quantifies the contribution of each factor in isolation.

Moreover, to address methodological concerns in the literature, we use two distinct methods for measuring these forces. First, we study the gap by comparing choice-proportions across treatments. This approach allows us to draw inferences independent of model assumptions. Moreover, it helps us connect and compare our findings with previous studies in this literature which used a similar approach. Second, assuming a rank dependent utility model

(Quiggin, 1982), we compare decision-weighting functions across treatments. This approach allows us to address an integral component of the gap (subjective distortions of probability) while controlling for other aspects of risky behaviour (such as utility curvature).

The choice-proportions analysis, reveals an overall significant DE gap which is compatible in direction and equal in size with the literature's average. This gap is mostly driven by an informational asymmetry (sampling bias). Nonetheless, other factors pertaining to preferential (ambiguity) and cognitive (likelihood representation and memory) aspects of decisions are also important. These results are corroborated by our model-mediated approach. Through examining the shape and relative position of probability weighting curves we discuss two intriguing behavioural implications of our findings.

Chapter 3, focuses on an integral component of the DE gap which relates to information search in Experience. As mentioned earlier, participants in the Experience conditions collect information relevant to their decisions by sequentially drawing independent observations from a source of uncertainty. In this chapter we explore the criteria that influence this search as well as how different search patterns affect ensuing risky choices. We show that a lottery's variance is negatively correlated with the sampling amount. In our context, this means that people sample more from options with rarer events. We also find that sampling amount decreases over time periods. Both of these findings become less salient however after the introduction of a history table which records and displays previously sampled outcomes during the lottery evaluation. The cue that dominates the treatment where such a table was present, is the table's maximum capacity.

Our analysis of risky choices captures a significant DE gap which is, however, mitigated by the presence of the history table. To the extent that the display of previously sampled events has a similar 'descriptive' effect to the numerical summaries of uncertainty in Description, this result should not come as a surprise. We interpret this 'bridging' of the gap as evidence that the DE gap should not be seen as a dichotomy between Description and Experience. Instead, we recommend that it is viewed as a continuum over different levels of uncertainty. Moreover, by examining the shape of probability weighting functions across Description and Experience, we propose a variant of the canonical interpretation. We refer to it as the 'relative underweighting hypothesis', which states that rare events in individual risky decisions are overweighted in Experience too, but less so than in Description.

In Chapter 4 we examine the DE gap in the social domain. Conditional cooperation -the willingness to cooperate if others do the same- is a prominent explanation for the existence of cooperation in social dilemmas (Fischbacher and Gächter, 2010). In cases where there is uncertainty regarding others' intentions, conditional cooperation necessitates the formation of expectations about the likelihood of cooperation. Information about the likelihood that others will cooperate can be obtained in different ways. In this chapter we investigate whether the format in which social information is obtained -descriptive or experiential- influences cooperation in social dilemmas. To this end, we develop an experimental protocol that allows systematic manipulation of the likelihood of cooperation in a Prisoner's Dilemma game.

In accord with past evidence in favour of conditional cooperation, we verify that a majority of subjects cooperate more if their partner is expected to cooperate as well. Interestingly, we find evidence for a gap in social decision making, but in the opposite direction from what the canonical finding in the individual context would have predicted. Rare events (of cooperation or defection) appear to be more overweighted in Experience rather than in Description. Another asymmetry with the individual domain is that sampling bias, the predominant driver of the DE gap in risky choices, does not affect the gap in the social domain.

Moreover, using a separate task to elicit cooperative types and provide some robustness tests related to these findings findings. We verify that cooperative behaviour maps intuitively into these types. Moreover, we confirm that the gap in cooperation is driven by conditional cooperators. These are by definition the people who would care about social information. Similarly, we show how conditional cooperators tend to collect significantly more social information than free riders or unconditional cooperators.

To interpret this finding we develop indexes that analyse behaviour in two domains. *Cooperativeness* captures an overall propensity to cooperate while *conditionality* measures the tendency to conditionally cooperate. Through this analysis we conclude that the reversal of the canonical DE gap is due to a decrease in sensitivity towards social information in Experience. We propose that a key reason why similar reversals do not happen more often in the individual context is an asymmetry in the strength of priors (stronger in the social context) and discuss how this hypothesis can be addressed in future research.

Throughout the next three chapters, the reader will likely notice the repetition of the terms: 'probability' and 'risk'. These terms have undoubtedly an important connection with the themes of this thesis. A search for their etymological origin however, revealed yet another connection of these terms with this work. The former term comes from the Latin word 'probare', which means 'to test', while the latter, from the early Italian word 'risicare' which means 'to dare'. The most important lesson from developing this thesis, comes from learning how to balance the two. 'Daring' to pursue new ideas and intuitions and finding reliable ways for 'testing' them.

BIBLIOGRAPHY

Abdellaoui, M., Baillon, A., Placido, L., and Wakker, P. P. (2011a). The Rich Domain of Uncertainty: Source Functions and Their Experimental Implementation. *American Economic Review*, 101(2):695–723.

Abdellaoui, M., L'Haridon, O., and Paraschiv, C. (2011b). Experienced vs. described uncertainty: Do we need two prospect theory specifications? *Management Science*, 57(10):1879–1895.

Barron, G. and Erev, I. (2003). Small feedback-based decisions and their limited correspondence to description-based decisions. *Journal of Behavioral Decision Making*, 16(3):215–233.

Bernoulli, D. (1954). Exposition of a New Theory on the Measurement of Risk. *Econometrica*, 22(1):23–36.

Fischbacher, U. and Gächter, S. (2010). Social preferences, beliefs, and the dynamics of free riding in public goods experiments. *American Economic Review*, 100(1):541–56.

Fischbacher, U., Gächter, S., and Fehr, E. (2001). Are people conditionally cooperative? evidence from a public goods experiment. *Economics Letters*, 71(3):397–404.

Hertwig, R., Barron, G., Weber, E. U., and Erev, I. (2004). Decisions from Experience and the Effect of Rare Events in Risky Choice. *Psychological Science*, 15(8):534–539.

Ore, O. (1960). Pascal and the invention of probability theory. *The American Mathematical Monthly*, 67(5):409–419.

Quiggin, J. (1982). A theory of anticipated utility. *Journal of Economic Behavior & Organization*, 3(4):323–343.

Savage, L. J. (1972). *The foundations of statistics*. Courier Corporation.

Tversky, A. and Fox, C. R. (1995). Weighing risk and uncertainty. *Psychological Review*, 102(2):269–283.

Tversky, A. and Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–

323.

Von Neumann, J. and Morgenstern, O. (2007). *Theory of games and economic behavior (commemorative edition)*. Princeton University Press.

Weber, E. U., Shafir, S., and Blais, A.-R. (2004). Predicting risk sensitivity in humans and lower animals: risk as variance or coefficient of variation. *Psychological Review*, 111(2):430–445.

2

UNDERSTANDING THE DESCRIPTION - EXPERIENCE GAP

*

"Although dichotomies promise to bring order to chaos, they do so at the cost of being simplistic"

— Katharina Barbe. , *The dilemma with dichotomies.* (2001)

2.1 INTRODUCTION

In early 2000, a catastrophic cyclone, "Cyclone Eline" hit the Mozambique coast, causing devastating floods. The consequences of this event include the death of eight hundred people, thousands without homes, two million affected and the destruction of over 90% of the irrigation systems in Mozambique. To reduce the vulnerability of farmers against future catastrophes, the Ministry of Environmental Affairs (MICOA) and the Ministry of Public Work and Housing (MOPH) spent an estimate of $13 million to build entire villages in the hills overlooking the floodplain, for those living in areas that were most prone to future flooding. The resettlement program however failed as merely a few months after living in their new houses, farmers started to return to the floodplain to farm and rebuilt their residence in their old villages (see Patt and Schröter, 2008, for a more detailed account of these events).

At face value, this disparity could be attributed to differences in risk appetites: farmers are more risk seeking than policy makers. However, according to recent developments in the field of cognitive and experimental psychology, the conflict might be arising from the way information regarding the rare event of a flood is communicated. Farmers experience the likelihood of a flood on a day-to-day basis. Throughout their entire lives, most farmers in the area are likely to have never experienced a second catastrophic event, other than the Cyclone Eline incident (this was in fact the worst flood in 50 years). On the other hand, policy makers have access to and can process descriptive information regarding the likelihood of a flood, such as probabilistic forecasts provided by experts. In this case, factoring in the possibility that this flood

* This chapter is based on joint work with Robin Cubitt and Chris Starmer.

was correlated with climate change, increased the probability that another flood would occur soon.

According to the 'Description - Experience (DE) gap' (Barron and Erev, 2003; Hertwig et al., 2004; Weber et al., 2004), people's risky choices are - at least partly - dependent on the way information about uncertainty is communicated: from description or from experience. Over the past years, this empirical discrepancy has been established in experimental laboratories. In these studies, participants are typically divided in two conditions. In the *Description* condition, they make choices between gambles whose properties (outcomes and outcome-probabilities) are explicitly and completely described in numerical form. Conversely, in the *Experience* condition, this information is inferred through a sequential sampling process. The most common interpretation of this disparity relates to the role of rare (low probability) events. People in *Description* tend to make decisions *as if* overweighting such events, relative to their probability, whereas in *Experience* they tend to make decisions consistent with underweighting them (Hertwig et al., 2004).

Both tendencies are departures from expected utility (EU) theory, but in different directions. Overweighting rare events in *Description* is in line with the tenets of Cumulative Prospect Theory (CPT; Tversky and Kahneman, 1992) - arguably the most influential alternative of EU. The (apparent) underweighting of these events in *Experience* however, is a surprising result which has generated a vivid debate among decision theorists. Among the several interesting open questions that this debate has instigated, this study is focusing on the factors underpinning the DE gap.

Why do risky choices differ between *Description* and *Experience*?

Unlike in *Description*, information in *Experience* is often *biased* and *ambiguous*. Even mathematically equivalent information, however, can be perceived or processed differently due to differences in its presentation. This is numerical and simultaneous in *Description* but sequential and analogical in *Experience*. Despite there being studies that have investigated the contribution of some of these factors (see Hertwig (2012) for a review), the evidence has been relatively scant. This study contributes to the literature, therefore, by providing a cohesive account of how and by how much each factor contributes to the gap. Conceptually, we distinguish between factors that pertain to informational (sampling bias), preferential (ambiguity) and cognitive (likelihood representation and memory) differences. With this taxonomy in mind, we design a treatment protocol and conduct a lab-experiment that allows us to isolate and quantify each of these factors. The benefits of this endeavour are threefold.

First, it can shed light onto why conflicts of interests such as that between the farmers and experts arise. Understanding the source of these differences and the way they map onto ensuing choices is very likely to produce valuable insights regarding the implementation of future relevant policies.

Second, it can inform the theoretical modelling of risky behaviour. Economists usually elicit risk preferences in the lab by providing participants with explicit descriptions of the uncertainty at hand. If preferences are significantly modulated by the presentation of this information, that would put into question the external validity of some of the hitherto stylized facts. For example, according to CPT, the leading descriptive preference model for choices in the lab (Barberis, 2013), people tend to overweight rare events. Is overweighting of rare events a genuine behavioural phenomenon or is it an artefact of the descriptive presentation of uncertainty? To the extent that experienced uncertainty warrants different weighting patterns, our study can provide a roadmap towards adjusting the model according to the influence of each factor. For example, if the gap owns only to informational differences due to limited sampling in *Experience*, then replacing objective probabilities with experienced relative frequencies in CPT's functional form should suffice to account for the gap (see Fox and Hadar, 2006). The verdict is not as clear however with respect to factors that pertain to preferential or cognitive aspects. We explore the impact of such factors through the lens of a rank dependent expected utility model (RDEU) that accounts for probability distortions. Using a large and diverse decision set and applying Abdellaoui et al.'s (2008) semi-parametric method of eliciting RDEU's components at the individual level, we qualify and quantify each factor's impact.

Finally, it can address an increasingly relevant tension in the literature. Over the past few years the size of the DE gap and even its overall direction (Glöckner et al., 2016) have been a point of debate. Differences in the experimental design, particularly related to the implementation of *Experience* in the lab, as well as different ways of measuring the gap, are very likely part of the reason for this volatility. First, different implementations could be fostering a different mix of driving forces. In this study, instead of eliciting a singular DE gap, we focus on the impact of specific factors that are inherently different between *Experience* and *Description*. Second, measuring the gap through direct choice comparisons or through certainty equivalents can produce markedly different conclusions. For this reason, by manipulating a feature of our elicitation method, we measure the gap through both methods.

2.2 BACKGROUND

The "sampling paradigm" (Hertwig et al., 2004) is the most common lab-setting in which the DE gap has been studied. Participants are typically divided into two treatments: *Description* and *Experience* where they make one-off choices between a risky and a safe gamble. Importantly, the risky gamble always contains an outcome that occurs with low probability.

In *Description* the properties of these gambles are explicitly and numerically described so that there is no uncertainty regarding their possible outcomes nor the probability distribution of these outcomes. In *Experience* on the other hand, participants are asked to gather information about these properties by themselves. The two gambles usually appear on screen in the form of two buttons. Every time a button is pressed, one of the outcomes of the gamble at hand appears on screen. These outcomes are tallied to appear in a relative frequency that matches their objective probability in *Description*.

In this framework, the DE gap is commonly detected through direct comparisons between choice proportions across the two conditions. The "*canonical finding*" consists of the observation that subjects in the *Description* condition tend to prefer the risky option when the rare event gives a desirable outcome, and to prefer the safe option when the rare event gives an undesirable outcome; whereas the opposite is observed in the *Experience* condition. Taken together, this pattern has been commonly interpreted as reflecting a tendency to overweight rare events in *Description* but underweight rare events in *Experience* (Hertwig et al., 2004).

The focus of our study is the identification and comparison of candidate forces that drive *Description* and *Experience* apart. In this section, we start by summarizing evidence from previous literature regarding these forces. We then describe a recently emerged discord in this literature owning to a multiplicity of DE gap manifestations and argue how our analysis contributes to its resolution.

2.2.1 *DE gap contributors*

Following our conceptual taxonomy, "*sampling bias*" refers to the informational dimension, "*ambiguity*" to the preferential one while "*likelihood representation*" and "*memory*" both pertain to the cognitive aspect of decision making. Next, we discuss these factors in greater detail.

Sampling bias refers to an information asymmetry. In *Description*, participants are informed accurately about the objective stochastic properties (outcomes and their probability distribution) of each option. In *Experience* however, objective and experienced information does not always match. Usually, people collect small samples (e.g. the median subject of Hills and Hertwig (2010) samples each option 9 times) and according to a property of the binomial distribution, such samples tend to under-represent rare events. If rare events are systematically under-represented in the samples observed by individual subjects, it would come as no surprise if their impact on choices is also attenuated. Despite the crucial contribution that sampling bias has been found to exert in the DE gap (Fox and Hadar, 2006; Rakow et al., 2008), there have been several studies arguing that the phenomenon survives its removal (Hau et al., 2010; Ungemach et al., 2009). To explore factors beyond sampling bias, researchers typically control the sampling process so that experienced relative frequencies and objective probabilities coincide. In this study, we follow the same approach in some of our treatments. By comparing risky choices between two conditions that differ only with respect to whether information was biased we are able to isolate the effect of this information asymmetry.

Ambiguity refers to the fact that unlike in *Description*, subjects in *Experience* cannot be certain about the objective parameters of the lotteries they face. Abdellaoui et al. (2011b) estimate CPT components at the individual level and find that decision weights elicited in *Experience* are systematically smaller than those elicited in *Description*. Drawing a parallel with Ellsberg's famous urn experiments, they attribute their findings to differences in attitudes towards ambiguity. Just as people prefer to gamble in the known urn over the unknown urn, so their willingness to bet is higher in *Description* (known distributions) than in *Experience* (partially unknown distributions). In this study, we investigate this claim further by isolating the effect of ambiguity. We do so by introducing a treatment manipulation across which participants observe mathematically equivalent information and which is presented in the same format but ambiguous in one case and not in the other.

Likelihood representation refers to the format with which the stochastic properties of uncertainty are communicated. In *Description*, probabilities are typically communicated through percentages (e.g. "win with 10% chance") but in *Experience* this information is obtained sequentially, resulting to a representation closer to natural frequencies (e.g. "there are 10 out of 100 winning cards"). These differences are likely to evoke different cognitive mechanisms and result in different actions. It has been demonstrated for example that communicating probabilities through natural frequencies (10 out of 100) instead of percentages (10%), improves Bayesian inference (Gigerenzer and Hoffrage, 1995). In this study, we isolate the potential effect of this representation difference by comparing two treatments offering mathematically equivalent information in two different formats, one based on explicit and numerical format and the other on a form of sequential sampling.

Memory and its boundaries is the second cognitive factor we investigate. When likelihoods are presented sequentially, claims about what information subjects have in mind are contingent on assumptions about their recall. For example, even a subject who has seen all the balls drawn from an urn without replacement may not have complete information about the proportion of red balls, if she can only remembers some of the draws. To control for this, we use in most of our treatments a history. table[1]

Typical experimental designs based on the sampling paradigm do not provide subjects with this kind of aid in the *Experience* condition, whereas in the *Description* condition subjects have full information about the options at the moment of choice. As a result, imperfect memory is a possible driver of any gap in behaviour between the two conditions. We examine the impact of this potential determinant of the DE gap by comparing risky behaviour between two conditions that are identical except that one of them does not offer this memory aid. Though the possible role of imperfect memory in the DE gap has been noticed in previous literature, it has proved hard to pinpoint (see for example the discussion by Wulff et al., 2018).

2.2.2 *A robust but diverse phenomenon*

In a recent meta-analysis, Wulff et al., 2018 find evidence for a significant DE gap, showcasing thus the robustness of this phenomenon. At the same time however, this analysis demonstrates a sizeable dispersion with respect to the size of the gap, ranging from very small to very large. In fact, recent evidence by Glöckner et al. (2016) has even suggested that under certain conditions, the gap can be reversed, with subjects in *Experience* making choices consistent with more overweighting than in *Description*.

This diversity is common also in studies that go beyond direct choice comparisons by fitting a model accounting for probability weighting. Although there is a consensus regarding the inverse S-shaped probability weighting curve - consistent with overweighting - in *Description*, the same is not true for the shape of the curves elicited in *Experience*. For example, Ungemach et al. (2009) estimated parameters that are compatible with a probability weighting curve that is S-shaped, which is indicative of underweighting of rare events. Hau et al. (2008) on the other hand, found this curve to be linear. Furthermore, Abdellaoui et al. (2011b), find the curve from *Experience* to be inverse S-shaped but lying beneath that from *Description* throughout the probability interval.

We distinguish between two factors that may underlie the diversity of this evidence. The first pertains to differences in the experimental design - particularly with respect to how *Experience* is implemented. The second attributes this diversity to differences in the way the gap is measured.

1 The idea of using a history table as a memory aid in *Experience* originates with Hau et al. (2010). This table records the sampled events for the subject to observe after the sampling process is completed, enabling thus perfect recall.

First, adaptations of its canonical designs are quite common. For example, Glöckner et al. (2016) fix sampling so that objective probabilities (in *Description*) and experienced ones (in *Experience*) coincide. They also focus on choices between non-degenerate binary lotteries, excluding choices between gambles and certainties that were part of the decision set of previous studies. Abdellaoui et al. (2011b) allow subjects to sample freely but only from one option at a time. They then evaluate these options through the method of certainty equivalents. The methodological diversity in the previous literature motivates our experimental design as this enables the isolation of the different factors by sequentially comparing treatments which differ in only one aspect at a time.

Second, most studies in this literature use direct choice comparisons while others have used the mediation of a behavioural model that accounts for probability weighting, such as CPT. To increase the precision of this estimation and to be able to separate between probability weighting and other aspects of risky behaviour such as subjective transformation of outcomes, Abdellaoui et al. (2011b) introduce a method that relies on the elicitation of certainty equivalents. To reconcile these differences, we measure the gap with multiple approaches. To achieve this, we manipulate Abdellaoui et al.'s (2011b) elicitation method in order to obtain both direct choice comparison scores and non-parametric decision weights at the individual level through certainty equivalents.

The next section provides details on the implementation of these methods.

2.3 DESIGN AND METHODS

2.3.1 *Treatments and decision problems*

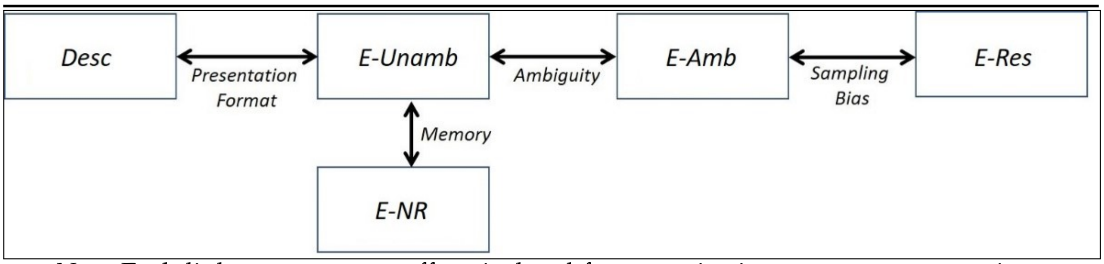
Our experimental design includes one *Description* (Desc) treatment and four variations of *Experience*: Unambiguous (E-Unamb), Ambiguous (E-Amb), No Records (E-NR) and Restricted (E-Res). In each treatment, subjects evaluate a series of binary gambles. These are represented by virtual decks, each containing two types of card demarcated by different colours. The probability of each outcome in this gamble is equal to the relative frequency of its corresponding colour. The valuation of each gamble takes place right after subjects receive information about these probabilities. During this evaluation, participants in all five treatments are able to see which outcome corresponds to which colour.

Treatments differ with respect to the type of information available to participants regarding the probabilities of different outcomes in a given gamble and the way this information is communicated. In Desc, it is provided in explicit and numerical form (percentages) at the time of evaluation. In our *Experience* treatments on the other hand, subjects discover this information by sampling sequentially through the content of each deck on a screen.

Figure 2.1 summarizes the key properties of each treatment (top panel) and the comparisons that isolate the potential DE gap factors (bottom panel).

Figure 2.1: Summary of treatments and treatment - comparisons

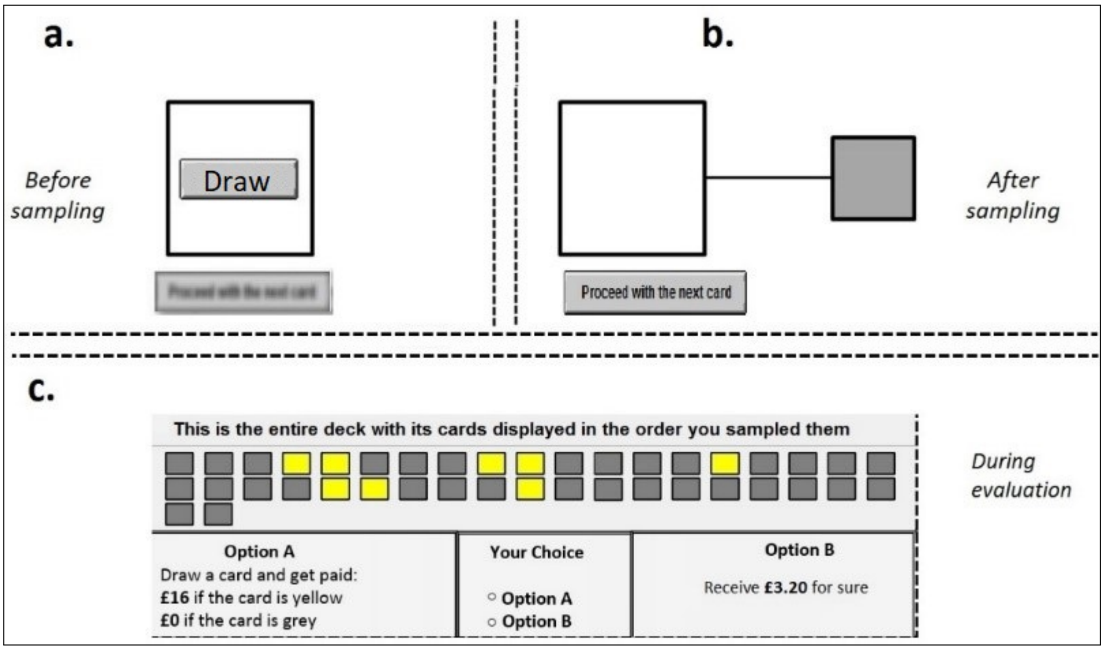
Desc	Decisions from description with numerical (percentages) likelihood-representations
E-Unamb	Subjects *knowingly* sample all 40 cards in each deck with a history table
E-NR	Same as **E-Unamb** but without the history table
E-Amb	Same as **E-Unamb** but subjects are unaware that they sample all cards
E-Res	Same as **E-Amb** but sampling restricted to 18 instead of 40 cards



Note. Each link represents an effect, isolated from a pairwise, treatment-comparison.

More specifically, E-Unamb is designed to remove ambiguity regarding the properties of lotteries (outcomes and their likelihoods) but also sampling bias and memory limitations. Therefore, comparing risky choices between Desc and E-Unamb, isolates the effect of likelihood representation. To remove sampling bias, we fix sampling amount so that the final sampling distribution matches the objective probabilities in Desc. We do this by fixing the size of the deck (to 40 cards) and implement a protocol of sampling without replacement. To remove memory limitations, we introduce a history table during the valuation screen, that records the colours of cards that were sampled in the order they were sampled. Finally, to remove ambiguity we stress to participants during instructions that the cards they sample are not replaced in the deck so that each card in the deck will be encountered exactly once. To highlight this, a pop-up message appears at the end of each sampling process reassuring people that this was the last card in the deck. They are reminded of this fact during the valuation stage where the message on top of the history table reads: "This is the entire deck with its cards displayed in the order you sampled them". Figure 2.2 depicts three instances of the experimental procedure for E-Unamb. Panels a. and b. capture the before and after of the sampling process while panel c. demonstrates an example of the valuation process.

Figure 2.2: Instances of E-Unamb's interface



Similarly, the other three versions of *Experience* are designed so that the appropriate binary treatment comparisons isolate the remaining effects. E-NR is identical to E-Unamb, except that there is no history table. To the extent that the history table removes memory constraints, the comparison of E-Unamb with E-NR captures the effect of memory boundaries. E-Amb is identical to E-Unamb except that participants are no longer informed that they sample the entire deck of cards. There is no pop-up message at the end of their sampling process and the message on top of the history table is less informative, reading: "These are the colours you sampled in the order you sampled them". Therefore, participants in E-Amb are not aware that the relative frequencies they experience are accurate (although, importantly, in fact they are) and thus the comparison between E-Unamb and E-Amb isolates the effect of ambiguity. Lastly, E-Res has been designed to be identical to E-Amb, except that the sampling amount is restricted to a smaller amount[2] and therefore the experienced relative frequency does not match the objective one. Consequently, comparing E-Amb with E-Res isolates the effect of sampling bias. Notice that sampling bias can take two directions: people can over- or under- represent a certain event in their sample. We find this distinction important and therefore split observations in E-Res into two subsets: E-Over and E-Under. Since rare events are the loci of our interest, we taxonomize observations in E-Over and E-Under according to whether the event with the smallest probability to occur was over- or under-represented.[3]

Besides isolating the factors of the DE gap, this treatment protocol enables us to study a variety of DE gaps. Comparing Desc with each E-variation, we obtain a distinct DE gap. Notice that the number of factors underlying each gap are increasing as we move from left to right in Figure 2.1. Comparing Desc with E-Unamb reveals a potential gap due to differences in likelihood representation. Moving further to the right, comparing Desc with E-Amb, one obtains a potential gap that combines likelihood representation and ambiguity, whereas the comparison between Desc with E-Res would add sampling bias to this mix. The comparison between Desc and E-NR is somewhat less straight-forward (thus not represented on the same line). In principle, this gap would be combining the factors of likelihood representation and memory. Nonetheless, to the extent that memory limitations induce uncertainty regarding the gamble's properties, ambiguity should also be included as a potential factor underlying this comparison. Lastly, in order to get an estimate of the average DE gap we elicit, we compare Desc with E-All, a compilation of observations across all 4 variations of *Experience*.

The valuation of each gamble took place at the bottom of the screen - just as it is depicted in Figure 2.2 - and the protocol was similar for all five treatments. The risky option (Option A) is associated with the deck of cards while the safe option (Option B) comprises of a certain amount. As in Abdellaoui et al. (2011b), we implement the bisection method whereby the certain amount is updated according to the previous choice of each subject. During the first iter-

2 This restriction is implemented by taking subsets of 18 cards from the unbiased 40-card distribution that was used in E-Unamb, E-NR and E-Amb.

3 There is one gamble with a 50-50 probability distribution that cannot fit this criterion. In this case, we taxonomize the observation according to observed relative frequency of the event corresponding to the highest of the two outcomes.

ation of each valuation, the certain amount is the expected value of Option A. In the second iteration, this amount is updated upwards (downwards) to the mid-point of the gamble's highest (lowest) outcome and the certain amount just rejected (accepted). The certainty equivalent is the certain amount of Option B after 5 such iterations.

Participants evaluated the lotteries summarized in Table 2.1. The order of these lotteries was randomized within two clusters for each subject. The order of the cards within each deck was also randomized for each subject. Lotteries in the first cluster (1.1 - 1.7) had a winning probability fixed at $p = 0.25$ and varying outcomes. Lotteries in the second cluster had a pair of fixed outcomes and varying probability. It is in this second cluster that we focus explicitly on the role of rare events.[4] Following the convention in this literature we consider an event rare[5] if its corresponding probability is less than $p < 0.20$ (see Hertwig et al., 2004). Among the lotteries in Table 2.1, the subset of lotteries containing a rare event includes problems: $\{2.4, 2.5, 2.6\}$, where $p < 0.20$ and problems: 2.7, 2.8, 2.9 where $(1 - p) < 0.20$. Since p represents the probability of the highest of the two outcomes in each gamble the first cluster of problems contains a desirable rare event while the second cluster contains undesirable rare events.

In total, 198 participants were recruited through ORSEE (Greiner, 2015) and randomly assigned to one of the five treatments summarized in Table 2.1. All sessions were conducted in CeDEx's laboratory at the University of Nottingham and lasted for approximately one hour. On average, subjects were paid £11.50, with each subject's payment dependent on their choices and on the resolution of gambles. The experiment was computerized, having been programmed in Z-tree (Fischbacher, 2007).

2.3.2 *Analysis*

We distinguish between two levels of analysis. The model-free analysis operates through direct choice comparisons. The advantage of this approach is that it facilitates the comparison of our findings with those in early literature, where the DE gap was established. The model mediated analysis operates through comparisons of RDEU components (utility curvature and weighting

4 As we explain in the estimation section, it was important that subjects realized that lotteries in this first cluster are related to the same probability distribution but different outcomes. To communicate this in the *Experience* treatments, these lotteries were associated with only one deck and one sampling process, but seven valuations. See Appendix A.1 for the instructions that were handed to participants in this experiment.

5 Since what constitutes a rare event is usually context dependent, using a specific threshold for this distinction might seem somewhat arbitrary. Nonetheless, since the goal of this section of the analysis is to compare findings with previous literature, we adopt this convention.

Table 2.1: Decision problems and characterisation

Decision Problem	Risky	Safe (1^{st} iteration)
1.1	(4, 0.25; 0)	1.0
1.2	(8, 0.25; 0)	2.0
1.3	(12, 0.25; 0)	3.0
1.4	(16, 0.25; 0)	4.0
1.5	(16, 0.25; 4)	8.0
1.6	(16, 0.25; 8)	10.0
1.7	(16, 0.25; 12)	13.0
2.1	(16, 0.25; 0)	4.0
2.2	(16, 0.5; 0)	8.0
2.3	(16, 0.75; 0)	12.0
2.4	(16, 0.025; 0)	0.4
2.5	(16, 0.05; 0	0.8
2.6	(16, 0.1; 0)	1.6
2.7	(16, 0.90; 0)	14.4
2.8	(16, 0.95; 0)	15.2
2.9	(16, 0.975; 0)	15.6

Note. Decision problems in grey cells contain a rare event.

curves, estimated semi-parametrically and at the individual level). The advantage of this approach is that we can examine the effect of probability weighting while controlling for other aspects of risky behaviour.

2.3.2.1 *Model-free analysis*

In the model-free analysis, we use tests that do not rely on any behavioural model and instead perform only cross-treatment comparisons of choices. We consider choices from the first iteration of each bisection. These are choices between the gamble (risky choice) and the certain amount (safe choice) equal to the gamble's expected value. This choice structure is similar to that of the early studies in the sampling paradigm. As these early studies focused only on situations involving rare events, this part of the analysis will focus only on the subset of decision problems containing a rare event (see decision problems in grey from Table 2.1).

We monitor risky behaviour through the "OVRW" index[6]. This index takes the value 1 if the risky option was chosen over the safe one in problems containing desirable rare events or if the safe option was chosen over the risky one in problems containing undesirable rare events. Intuitively, this in-

6 This is an abbreviation of "overweighting". Glöckner et al. (2016) refer to the same index as p(overweighting)".

dex accounts for the influence of rare events. According to Wulff et al. (2018): "this scoring follows from the assumed as-if over (description) versus underweighting (experience) of rare events". Summing up OVRW for each subject and dividing with the total of relevant choices (6) we obtain a measure of propensity to overweight which does not suffer from repeated observations. We refer to this index as %OVRW. Aggregating this propensity for each treatment we obtain the %OVRW score for each treatment. Therefore, according to canonical DE gap every time the %OVRW score is higher in *Description* than in *Experience*, that is an instance of the DE gap.

2.3.2.2 *Model-mediated analysis: RDEU*

In this section we provide a brief account of the formal requirements that are necessary to study the DE gap through a RDEU model (Quiggin, 1982). Since the canonical finding in this literature relates to how people treat rare events (underweight or overweight them), we decided that a model that can incorporate probability weighting is the most appropriate choice. Therefore, although expected utility (EU) theory is considered by many economists the benchmark for studying risky behaviour, in this study we employ a RDEU model that accounts for probability distortions.

More formally, let $x_{E_p}y$ stand for a binary gamble where x, y are non-negative outcomes, contingent on mutually exclusive events and $x > y$, so that x is the high (or desirable) outcome and y is the low (or undesirable outcome). Let E_p stand for an event occurring with probability p. For example, $E_{0.25}$ can represent the event of drawing a yellow card from a deck containing 100 cards where only 25 of those are yellow. According to RDEU, given a strictly increasing utility function: u and a probability weighting function W, subjects maximize:

$$W(E_p)u(x) + (1 - W(E_p))u(y) \tag{2.1}$$

An additional advantage of RDEU is that it allows preferences consistent with EU to emerge, in the special case where probabilities are weighted linearly so that: $W(E_p) = p, \forall p$. Moreover, for the type of non-negative, binary lotteries that we consider in this study, the RDEU model has the same functional form with most transitive non-expected utility models (Ghirardato and Marinacci, 2001; Luce, 1991; Miyamoto, 1988; Tversky and Kahneman, 1992).

To study the DE gap through the lens of a RDEU model, we use the source method (Abdellaoui et al., 2011a; Tversky and Fox, 1995). This method was later developed to accommodate the DE gap by Abdellaoui et al. (2011b). According to this approach, sufficiently different environments -or equivalently, sources of uncertainty- give rise to different probability weighting curves. In our setting, these environments are defined at the treatment level. More formally, for an event E_p, such as drawing a yellow card from a deck in a specific treatment, where $(100 \times p)\%$ of its cards are yellow, the decision weight is given by:

$$W(E_p) = w_\sigma(\pi(E_p)) \tag{2.2}$$

In Equation 2.2, $W(E_p)$ is the decision weight associated with event E_p, w_σ is the source function which transforms probabilities into decision weights according to the source of uncertainty σ. In our experiment, each treatment corresponds to a different such source. Moreover, $\pi()$ is the belief of the likelihood of E_p. It is commonly assumed that in *Description*, $\pi(E_p) = p$. In *Experience* on the other hand, this belief depends on the observed relative frequency (f_p) of each event E_p. We therefore assume that $\pi(E_p) = f_p$ for our *Experience* treatments[7]. Since in most of our *Experience* variations, $f_p = p$, it follows that $\pi(E_p) = p$. Therefore, Equation 2.2 can be re-written simply as:

$$W(E_p) = w_\sigma(p)) \tag{2.3}$$

Notice that conditions that include observations drawn from biased samples are an exception to this. In these cases, although $f_p \neq p$, we are still operating under Equation 2.3. In these cases, p corresponds to the probability that would have generated the accurate f_p had it not been pre-maturely interrupted. Therefore, we take the difference between $w_\sigma(p)$ and $w_\sigma(f_p)$ to be capturing the effect of sampling bias.

To study these source functions, we need to estimate the utility curvature and decision weights at the individual level. We use the seven certainty equivalents elicited from decision problems: 1.1 - 1.7, to fit the utility curvature parameter of the power utility function: $U(x) = x^\alpha$. We do so by minimizing the non-linear least squares: $(||z - \hat{z}||)^2$ where z_i refers to the observed certainty equivalent of risky gamble i, and $\hat{z}_i$ is the estimated certainty equivalent. An important feature of this estimation protocol is that the event E_{p^*} corresponding to outcome x_i is common for $i = 1, \ldots, 7$ (and therefore associated with the same probability: p^*). Therefore, we can treat the corresponding decision weight: $W(E_{p^*})$ as a free parameter to be estimated together with the utility curvature parameter: α. We can therefore re-write Equation 2.2 as:

7 This is a common assumption in the literature and it is founded in findings reporting high correlations between judged probabilities and observed relative frequencies (e.g. Fox and Hadar, 2006).

$$\hat{z}_i = [W(E_{p^*})(x_i^{\alpha} - y_i^{\alpha}) + y_i^{\alpha}]^{\frac{1}{\alpha}} \tag{2.4}$$

Having obtained an estimate of each subject's utility curvature, we proceed to calculate non-parametrically decision weights for lotteries in 2.1 - 2.9. Notice that these lotteries have fixed outcomes: $x^* = 16$ and $y^* = 0$ and varying probability: p_j. Deriving Equation 2.4, we can calculate therefore decision weights for each p_j level according to:

$$W(E_{p_j}) = \left(\frac{z_j^{'}}{x^*}\right)^{\alpha} \tag{2.5}$$

Where, $z_j^{'}$ stands for the elicited CE from risky option j, with $j = 1, 2, \ldots, 9$. Aggregating weights across individuals, we obtain an aggregated source function under each treatment. By studying the shape of the elicited weighting curves and comparing them across treatments, we examine the DE gap and its driving forces from the perspective of a model that accounts for weighting[8].

8 For more details on the theoretical background and its implementation, see Abdellaoui et al. (2011b) and Abdellaoui et al. (2011a).

2.4 RESULTS

2.4.1 *Model-free analysis*

We begin our analysis by examining choice proportions through %OVRW scores. These scores derive from the choice made in the first iteration of each bisection process. As detailed in Section 2.3.2.1, a popular interpretation of this score is that it captures a propensity to overweight rare events. According to this interpretation, taking the average %OVRW for each treatment captures a treatment-level propensity to overweight. Figure 2.3 plots average %OVRW scores across all 5 treatments, including our 3 treatment derivatives: E-Over, E-Under and E-All.

As we can see from Figure 2.3, the propensity to overweight is higher in Desc than in any variation of *Experience*. This is in line with the canonical finding. To examine the size and statistical significance of the ensuing DE gaps as well as the isolation of the factors that drive them, we report the appropriate treatment comparisons in Table 2.2.

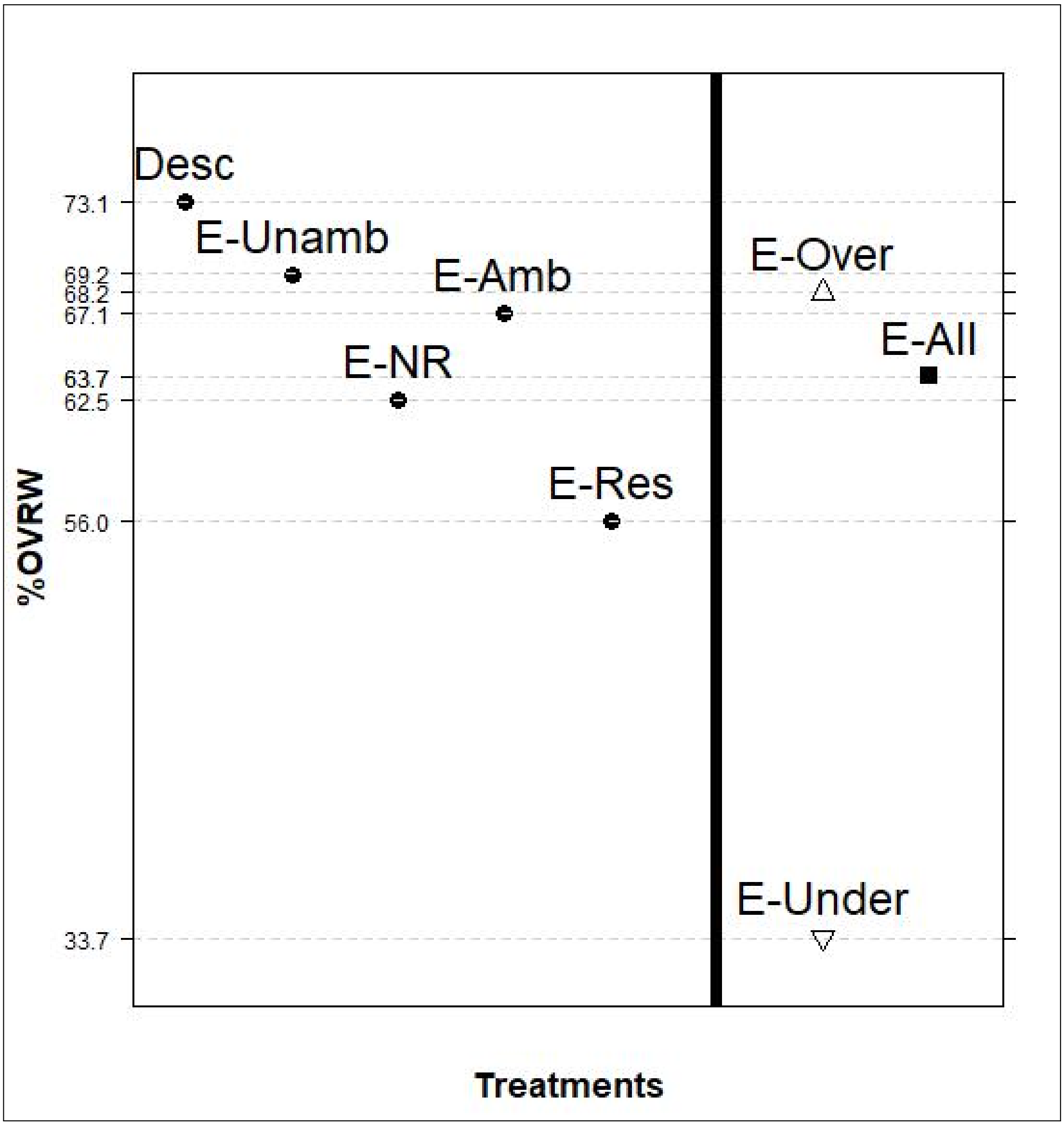
Table 2.2: Treatment comparisons and effect size measured in %OVRW

	Treatment comparisons	Difference in %OVRW	$P - value^{+}$
	Desc vs. E-All	9.35^{**}	0.041
DE gaps	Desc vs. E-Unamb	3.91	0.284
	Desc vs. E-NR	10.58	0.121
	Desc vs. E-Amb	6.00	0.209
	Desc vs. E-Res	17.09^{***}	0.008
	Desc vs. E-Over	4.87	0.555
	Desc vs. E-Under	39.34^{***}	0.000
Effects	Likelihood representation	3.91	0.284
	Memory	6.67	0.468
	Ambiguity	2.08	0.727
	Sampling Bias	11.10	0.126
	Overrepresentation	-1.12	0.579
	Underrepresentation	33.35^{***}	0.000

Note. $^{+}$ Reported P-values from 2-sided, Exact Wilcoxon Rank Sum tests on %OVRW differences. We obtain only one %OVRW per subject so there is no concern regarding repeated observations. $^{***}P < 0.01;^{**} P < 0.05;^{*} P < 0.1$.

The top rows of Table 2.2 report a series of DE gaps. The average DE gap (derived from Desc vs. E-All) was 9.35 percentage points wide and significant ($P = 0.041$; MW exact). The size of our average gap is very close to this

Figure 2.3: Average %OVRW scores across treatments



Note. Average %OVRW scores plotted across treatments (Desc, E-Unamb, E-Amb, E-NR, E-Res) and treatment-derivatives (E-Over, E-Under, E-All).

literature's average (9.7 percentage points) which was calculated in a large meta-analysis from 80 data sets (Wulff et al., 2018).

Result 1. *"A standard DE gap"*
The average DE gap we elicit is in accord with the canonical finding and of similar size with the literature's average.

Comparing Desc with each *Experience*-treatment separately, we observe that the strongest evidence for the gap stems from the "Desc vs. E-Res" comparison, where sampling bias is present. Not surprisingly, the gap is widest when rare events are under-represented rather than when they are over-represented. In comparisons that do not involve sampling bias we find no evidence for a statistically significant DE gap, although the comparison between Desc and E-NR comes close ($P = 0.121$; MW test).

Interestingly, the two widest gaps derive from the comparisons in which most factors were at play. The "Desc vs. E-NR" comparison combines the effects of: memory, likelihood representation and, potentially, ambiguity (to the extent that someone is aware of her memory limitations during the valuation of a gamble). The "Desc vs. E-Restr" comparison combines the effects of: sampling bias, likelihood representation and ambiguity. This suggests that there is a synergistic relation between the factors we isolate: DE gaps comprised of more factors are wider than DE gaps comprised of fewer factors.

Our isolation protocol sheds some light on this finding. Looking at the bottom panel of Table 2.2 ("Effects"), we see that all effects, besides overrepresentation, have a positive sign. This means that each factor contributes to the gap in the direction predicted by the canonical finding: each reducing %OVRW and therefore the propensity to overweight rare events. As a result, comparing treatments that vary in more than one factor at a time, is bound to capture wider effect sizes.

Result 2. *"Stronger together"*
Every factor we isolate contributes positively to the canonical DE gap. Therefore, combining the effect of multiple factors together produces wider gaps.

In terms of magnitude, the effect of sampling bias is the most impactful one, especially when focusing on the subset of cases where rare events have been under-represented. This is in fact the only instance where a factor in isolation is significant ($P < 0.01$, MW-exact). Not surprisingly, over-representation of rare events acts in the opposite direction (albeit, not significantly) mitigating the contribution of sampling bias to the gap.

Result 3. *"Key driver: under-representation"*
Sampling bias driven from under-representation of rare events is the largest and only significant factor in isolation.

To summarize, our direct choice comparison analysis revealed that the average DE gap we elicited, coincides both in terms of direction and in terms of size with the literature's average finding (Result 1). We interpret this as evidence that the phenomenon is robust even though we applied significant adaptations to the sampling paradigm - the original experimental set up for studying the DE gap. Our isolation protocol revealed that each of the factors we set to analyse contributes to the canonical gap positively. Consequently, the DE gaps we elicited widen as a function of the number of factors they comprise of (Result 2). However, the effect of sampling bias stemming from under-representation of rare events is the key and only significant driving force (Result 3) in isolation.

Does this imply that the sampling bias is the only significant driver of the gap, as some papers have suggested (e.g. Fox and Hadar, 2006; Rakow et al., 2008) or are other factors also relevant as another section of the literature maintains (Hau et al., 2008; Ungemach et al., 2009)?

To explore this question, we turn to a more detailed examination of the DE gap and its factors. Mediating a risky preferences decision model to our data allows us to examine probability weighting while controlling first for other potentially confounding factors such as subjective transformation of outcomes (utility curvature). Moreover, our non-parametric calculation of decision weights, allows us to zero-in on sub-intervals of the unit interval, elaborating therefore the search for the DE gap and the effect of its drivers.

2.4.2 *Analysis through the RDEU model*

In this analysis we use CEs which we derive from all choices made throughout each bisection process. We then use these CEs to estimate for each individual: utility curvature first (as per Equation 2.4) and then calculate decision weights (as per Equation 2.5).

Table 2.3 reports median values for the utility curvature parameter (α) across treatments. On aggregate, our findings suggest a near linear[9] utility over money. Median values are very similar across treatments; a Kruskal-Wallis test does not reject the null hypothesis that utility curvature does not differ across treatments ($P = 0.613$). Absence of difference with respect to utility curvature suggests that potential treatment effects are more likely to occur due to differences in probability weighting rather than due to differences in preferences over money. Nonetheless, the interquartile ranges provide evidence of significant heterogeneity within each treatment. This verifies our concerns that inferring probability weighting from direct choice comparisons may be ignoring a significant source of behavioural heterogeneity.

Table 2.3: Utility curvature across treatments (medians)

	Desc	E-Unamb	E-NR	E-Amb	E-Res
Median	1.06	1.08	1.07	1.10	0.98
IQR	0.84-1.38	0.82-1.47	0.81-1.89	0.83-1.35	0.66-1.17

Note. Parametric estimations of utility curvature: α from x^{α}. These estimates derive from a non-linear least squares algorithm, commonly specified for all 198 subjects.

9 These estimates fall within the range of contemporaneous studies, reporting the power of the utility function to be between 0.8 and 1.1 (Abdellaoui, 2000; Booij et al., 2010; Etchart-Vincent, 2004; Fehr-Duda et al., 2006; Murad et al., 2016).

Next, we calculate decision weights for each subject at each probability level, following Equation 2.5. Table 2.4 reports median values for these decision weights and statistical comparisons with the diagonal.

Table 2.4: Median decision weights and comparison with the diagonal

p	**Desc**	**E-Unamb**	**E-Amb**	**E-NR**	**E-Res^{+}**	**E-Over^{+}**	**E-Under^{+}**	**E-All^{+}**
0.025	0.096^{***}	0.045^{***}	0.047^{***}	0.017^{ns}	0.051^{***}	0.064^{***}	0.012^{ns}	0.039^{***}
0.050	0.125^{***}	0.100^{***}	0.071^{***}	0.064^{**}	0.069^{*}	0.070^{*}	0.038^{ns}	0.061^{***}
0.100	0.185^{***}	0.146^{***}	0.090^{ns}	0.063^{ns}	0.101^{ns}	0.146^{ns}	0.083^{ns}	0.098^{ns}
0.250	0.184^{ns}	0.238^{ns}	0.196^{**}	0.187^{ns}	0.177^{ns}	0.200^{ns}	0.172^{ns}	0.204^{ns}
0.500	0.358^{***}	0.336^{**}	0.349^{***}	0.306^{***}	0.353^{***}	0.413^{ns}	0.385^{*}	0.336^{***}
0.750	0.474^{***}	0.496^{***}	0.446^{***}	0.370^{***}	0.490^{***}	0.512^{***}	0.493^{***}	0.468^{***}
0.900	0.764^{***}	0.727^{***}	0.603^{***}	0.559^{***}	0.754^{***}	0.637^{***}	0.996^{ns}	0.658^{***}
0.950	0.628^{***}	0.693^{***}	0.729^{***}	0.839^{***}	0.791^{***}	0.742^{***}	0.997^{*}	0.775^{***}
0.975	0.842^{***}	0.898^{***}	0.822^{***}	0.894^{***}	0.943^{***}	0.862^{***}	0.999^{ns}	0.890^{***}

Note. Differences with the diagonal derive from MW tests for: $H_0 : W(E_{p_j}) = p_j$ vs. $H1 : W(E_{p_j}) \neq p_j$

"$***$": significantly different from the diagonal at $P < 0.01$

"$**$": significantly different from the diagonal at $P < 0.05$

"$*$": significantly different from the diagonal at $P < 0.10$

"ns": not significantly different from the diagonal at $P < 0.10$

$^{+}$Containing cases where objective and observed relative frequencies do not (always) coincide.

Recall that rare events can be found in two regions: for $p \in \{0.025, 0.05, 0.10\}$ and for $p \in \{0.90, 0.95, 0.975\}$. As p is associated with x, the highest of the two possible outcomes, rare events in the first interval are associated with desirable outcomes and their decision weight corresponds to $W(E_p)$. Conversely, rare events associated with probabilities in the second interval are associated with the probability of not obtaining x and are therefore characterised as undesirable. According to RDEU, their associated decision weight is calculated by $1 - W(E_p)$. Since Table 2.4 reports only $W(E_p)$, overweighting of rare events is compatible with $W(E_p) > p$ for $p < 0.25$ and $W(E_p) < p$ for $p > 0.75$. Conversely, rare events are underweighted when $W(E_p) < p$ for $p < 0.25$ and $W(E_p) > p$ for $p > 0.75$.

As we see in Table 2.4, decision weights are almost everywhere above the diagonal for low values of p (positive values for t-statistics) and under it for high values of p (negative values for t-statistics), with a crossover near $p = 0.25$. This is evidence that probability weighting takes an inverse S-shaped form, suggesting overweighting. There are two exceptions to this pattern. First, in E-NR, the weighting curve is under the diagonal for small values of p, suggesting underweighting of rare events (except for $p = 0.05$). Second, and even more strikingly, in E-Under, the weighting curve appears to be S-shaped: decision weights are below the diagonal for $p < 0.25$ and above it for $p > 0.75$. There is a caveat to these two exceptions. Even though nominally there seems to be a case for underweighting in E-Under and E-NR, statistical analysis suggests otherwise. In these cases, decision weights are in fact not significantly different from the diagonal (see Table 2.4 for how $H_0 : W(E_{p_j}) = p_j$ cannot be rejected in these cases).

Result 4. *"Mostly overweighting"*
Aggregate decision weighting functions reveal a pattern consistent with overweighting of rare events both in Description and in Experience. Rare events in E-Under and desirable rare events in E-NR are two notable exceptions to this pattern.

Result 4 suggests that CPT's claim that people overweight rare events, can be found beyond the narrow frame of described risk. In fact, overweighting seems compatible with cases where uncertainty is experienced. Nonetheless, the two exceptions: E-NR and E-Under, warrant a note of caution. Examining more closely the effects that shape these curves is likely to shed more light on the behavioural implications of these two deviations.

Table 2.5 reports the P-values from MW tests for the same treatment comparisons as those summarized in Table 2.2. This time however, we can distinguish between different points of the probability interval. From the top panel of Table 2.5, we see that the largest and most significant DE gap stems from the Desc vs. E-Under comparison, where sampling bias is present in the form of under-representation of rare events. These findings corroborate our model-free conclusions.

Table 2.5: Statistical significance (P-values of MW tests)

DE gaps							
p	**Desc vs E-All**	**Desc vs E-Unamb**	**Desc vs E-Amb**	**Desc vs E-NR**	**Desc vs E-Res**	**Desc vs E-Over**	**Desc vs E-Under**
0.025	0.030**	0.109	0.246	0.011**	0.126	0.805	0.022**
0.050	0.078*	0.479	0.254	0.078*	0.0521*	0.303	0.022**
0.100	0.324	0.803	0.254	0.094*	0.591	0.958	0.196
0.250	0.987	0.409	0.485	0.685	0.735	0.527	0.994
0.500	0.967	0.555	0.887	0.491	0.705	0.723	0.515
0.750	0.644	0.996	0.736	0.175	0.811	1.000	0.697
0.900	0.762	0.841	0.643	0.361	0.518	0.878	0.024**
0.950	0.094*	0.432	0.581	0.127	0.014**	0.308	0.000***
0.975	0.316	0.307	0.650	0.443	0.065*	0.922	0.001***

Effects						
p	**Likelihood Repr.**	**Memory**	**Ambiguity**	**Sampling Bias**	**Over-representation**	**Under-representation**
0.025	0.109	0.239	0.844	0.811	0.232	0.099*
0.050	0.403	0.259	0.799	0.302	0.739	0.140
0.100	0.895	0.053*	0.285	0.743	0.472	0.629
0.250	0.437	0.247	0.125	0.345	0.306	0.567
0.500	0.622	0.239	0.601	0.542	0.783	0.343
0.750	0.965	0.180	0.670	0.529	0.635	0.591
0.900	0.751	0.316	0.549	0.275	0.733	0.022**
0.950	0.473	0.370	0.988	0.064*	0.790	0.000***
0.975	0.331	0.867	0.103	0.021**	0.554	0.000***

Note. Column 1 reports the list of objective probabilities for which decision weights were calculated. In the remaining columns we report P-values for the MW tests.

$^{***}P < 0.01$;** $P < 0.05$;* $P < 0.1$.

However, the RDEU mediated analysis offers some new insights too. We now can detect a significant DE gap in cases where there was no sampling bias. This version of the DE gap derives from the Desc vs. E-NR comparison for small values of p. Not surprisingly, this is the same region where E-NR exhibits apparent underweighting of rare events.

The Desc vs. E-NR version of the gap is the only one that features memory limitations, suggesting that this cognitive dimension of the gap is important. Nonetheless, we cannot attribute the entirety of this effect to the absence of the history table. As we discussed in Section 2.3, this treatment comparison is likely fostering effects that can be attributed to differences in likelihood representation and ambiguity.

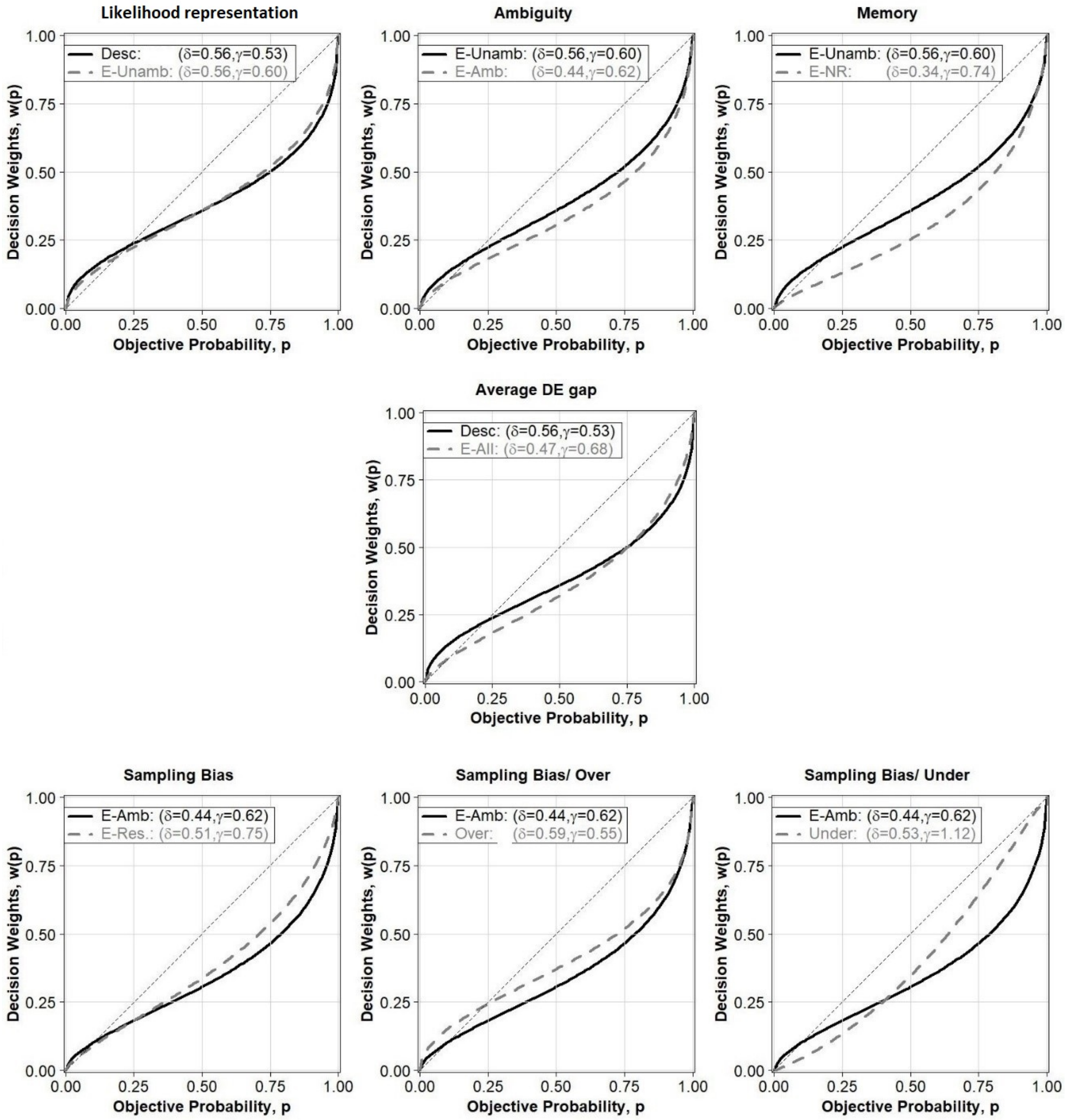
Result 5. *"Beyond sampling bias"*
When the role of memory limitations is coupled with that of presentation format and (potentially) that of ambiguity, we find a significant DE gap for desirable rare events, even when we control for sampling bias.

The fact that memory limitations cannot account entirely for the Desc vs. E-NR gap is also evident from the treatment comparisons that isolate each factor. In the bottom rows of Table 2.5, we see that *Memory* alone can produce a weakly significant effect only when $p = 0.10$. We interpret this as further support of the "stronger together" finding as summarized in Result 2. Our model free analysis findings are also echoed in our isolation of sampling bias. There, similarly to Result 3, we verify that sampling bias, stemming from under-representation of rare events is producing the widest effect.

Next, we proceed with providing a visual counterpart to the hitherto RDEU analysis. Figure 2.4 plots weighting curves across all treatments and treatment derivatives. These curves derive from the median decision weights reported in Table 2.4. To facilitate their presentation, we fitted a parametric specification of the weighting curve through a simple non-linear least squares minimisation, applied at the level of the median decision weight for each treatment at each probability level. We use the linear-in-log-odds specification of the weighting curve: $w(p) = \frac{\delta p^{\gamma}}{\delta p^{\gamma}+(1-p)^{\gamma}}$. This parametric function was introduced by Goldstein and Einhorn (1987) (see also Gonzalez and Wu, 1999) and has since gained popularity primarily due to its tractability. Parameter δ controls for the elevation of the curve while parameter γ for its curvature.

Figure 2.4 includes pairwise plots of the weighting curves, according to our isolation protocol (see Figure 2.1) as well as a comparison between Desc and E-All. Plotting the shape of each treatment's curve offers an easy to process, pictorial counterpart of Table 2.4. This makes it easy to compare the relevant position of each weighting curve pair so that we can draw qualitative inferences for the effects driving the two curves apart. For instance, focusing on

Figure 2.4: Weighting curves across treatments



Note. Top row: Comparisons without sampling bias. Bottom row: Comparisons with sampling bias. Centre: Desc vs. E-All. Legends report the parameters of the weighting function that was used to fit the curves.

the effect of *ambiguity*, we find that it exerts a downward force on the weighting curve (w_{E-Amb} tends to be below that of $w_{E-Unamb}$). This decline is in line with the suggestion that ambiguous information reduces the willingness to pay for a given gamble, independently of its winning probability (see Abdellaoui et al., 2011b).

The role of *likelihood representation* does not seem to leave a distinguishable footprint in our data by itself. Nonetheless, as we saw earlier, its effect could be interpreted as a type of "catalyst" that amplifies the impact of other effects when in contact, most noticeably, that of *memory*.

Right after Result 4, which stated that E-Res and E-NR are the only exceptions to the typical inverse S-shaped weighting pattern, we asked what where the behavioural implications from these deviations. We now return to this question by zeroing in on the weighting curve patterns that isolate the effects of *sampling bias* and *memory*. Notice that E-Res and E-NR are the conditions that drive the effects of sampling bias and memory. The former forcing participants to collect biased samples and the latter depriving them of a memory aid. Examining the rightmost panels of Figure 2.4 (top right: "Memory"; bottom right:"Sampling Bias/ Under") reveals that these two factors induce a tendency towards underweighting. Recall that over-weighting occurs when a weighting curve is above the diagonal for small values of p and below the diagonal for high values of p. This pattern is violated in the case of *sampling bias* (albeit less strikingly for high values of p) and in the case of *memory* for small values of p.

Regarding sampling bias, it might come as little surprise that overweighting of rare events is mitigated when these events are under-represented. From this perspective, we find the similarity between the effect of memory and that of sampling bias stemming from under-representation, potentially suggestive of the mechanism through which memory limitations affect risky decisions in *Experience*. If people can only recollect parts of the sequences they sampled, then their recollections are likely to suffer from the same under-representation effect that the binomial distribution "imposes" on small samples. Therefore, memory limitations can be viewed as a special case of cognitive sampling bias. The fact that this similarity is observed only at the leftmost end of the interval is likely related to the size of the stakes. The expected value of the lotteries that contain desirable rare events (small p) is on average £0.9 while that of the lotteries containing undesirable rare events (high p) £15.1. One possibility therefore is that in light of these higher stakes, participants exerted higher effort to retrieve previously encountered information, rendering cognitive under-representation discernible only at the low-stakes region, where desirable events are rare.

Another intriguing behavioural interpretation of these two exceptions stems from the statistical comparison of the $w_{E-Under}$ and w_{E-NR} with the diagonal in the domains of rare events. There, we observe that the apparent underweighting is (from a statistical perspective) indistinguishable from the diagonal. To the extent that non-linear probability weighting is a distortion of optimal behaviour[10] one could interpret this finding as the product of the clash of two opposing biases. The behavioural bias to overweight rare events, partially cancels out the statistical (and/or cognitive) bias to under-represent rare events in small - collected or recollected - samples. The idea of bias-complementarity, where an individual can be better off under two antago-

10 For a discussion on the normative superiority of Expected Utility Theory - which emerges as a special case of CPT with linear weighting - see Wakker (2010).

nistic biases rather than only one, is not novel in economics literature. For example, Waldman (1994) discusses how overconfidence and aversion to effort can harmoniously coexist. Although tentative, we find the remark that the behavioural tendency to overweight rare events is nature's antidote for our limited memory capacity or our tendency to under-represent rare phenomena in our brief and finite life-spans, a promising future direction.

Lastly, we turn to the central panel of Figure 2.4 where w_{Desc} is plotted next to w_{E-All}. Unlike the top and bottom row of this figure, this comparison does not isolate one effect. Rather, it displays the average DE gap we elicit in this study by juxtaposing Desc with an amalgamation of all *Experience* treatments. Although both curves are inverse S-shaped, we can see that w_{E-All} lies beneath (above) w_{Desc} for small (high) levels of p. It follows that:

Result 6. *"The Relative Underweighting Hypothesis"*
Rare events appear to be overweighted in Experience, but less so than in Description.

According to Result 6, the average DE gap we elicited is not of the type: "over- vs. under- weighting" but rather of the type: "over- vs. less over- weighting". Given how the magnitude of our average DE gap coincides with the literature's average, it is likely that the "relative underweighting hypothesis" is overall, an accurate picture of the DE gap.

2.5 CONCLUSION

We reported the results of a lab experiment that investigated the Description - Experience (DE) gap, a recent empirical phenomenon pointing to a discrepancy of risk attitudes when these are elicited from *Description* or from *Experience*. According to the most popular interpretation of the canonical finding in this literature, people in *Description* behave *as if* overweighting rare events, relative to their probability. Conversely, people in *Experience* tend to make decisions consistent with underweighting rare events.

We taxonomized what we believe are the key factors driving this empirical discrepancy in three broad categories by distinguishing between factors pertaining to: informational (sampling bias), preferential (ambiguity) or cognitive (likelihood representation and memory limitations) aspects of decision making. Then, we implemented a novel 5-treatment design comprising of one standard version of *Description* and 4 variations of *Experience*. Our treatment protocol was designed to isolate these factors through a series of pairwise comparisons. At the same time, this design allowed us to elicit a series of DE gap variants; one for each comparison of *Description* with a variant of *Experi-*

ence. Moreover, to address certain methodological concerns in this literature, we employed two measuring approaches.

First, we study the gap in the absence of any behavioural model assumptions, by focusing only on choice proportions from pair gamble questions. In doing so, we find that despite our adaptations to the "sampling paradigm" - the leading experimental framework for studying the phenomenon - our average elicited DE gap coincides in direction and size with the literature's average suggesting that the phenomenon is robust. Moreover, we find that each factor contributes positively to the phenomenon and therefore, comparisons entailing more than one factor, induce bigger effects. Among those factors, the most potent in isolation is sampling bias due to under-representation of rare events.

Second, assuming a rank dependent utility model, we compare decision-weighting functions across treatments. These are elicited semi-parametrically and at the individual level with the use of certainty equivalents. This level of analysis allowed us to examine probability weighting while controlling for other aspects of risky behaviour, as well as to explore a variety of probability regions separately.

We find the two levels of analysis to be complimentary. The model-mediated approach replicates the findings of the model-free analysis while the use of weighting functions and their shape allows us to shed some more light on aspects of behaviour that would otherwise be inaccessible. Most notably, we observe that a significant DE gap can be found even when controlling for sampling bias; a result that runs opposite to some previous claims. The role of memory limitations appears to be pivotal for this gap, especially when combined with the effect of likelihood representation and ambiguity. This finding has the following two implications.

On the one hand, it serves as a reminder that cognitive aspects of behaviour are not to be discounted when eliciting risky preferences nor when extrapolating lab findings to policy decisions. To this end, we discussed two intriguing interpretations of our finding. The first interpretation, parallels memory limitations to a type of cognitive sampling bias. The statistical property of the binomial distribution under which rare events are under-represented in small samples, might also be a property of the brain. Limited recollections of past events may also under-represent the rare ones. The second interpretation, draws from the evolutionary principle of bias complementarity. Specifically, we discuss the possibility that the behavioural bias to overweight rare events (as captured by several non - EU models such as CPT) may be countering the statistical (and/or cognitive) bias to under-represent them in small - collected or recollected - samples. As a result, the corresponding weighting curves are often statistically indistinguishable from the diagonal - and therefore from the normative EU benchmark - in the region of rare events.

On the other hand, despite finding that seeking for a DE gap beyond sampling bias is not a "fool's errand", our results recommend that the seeker be equipped with a magnifying glass, for the gap is small, and a compass, for it is not ubiquitous. Nonetheless, it is conceivable that outside the lab, these effects are likely to be amplified on the merit that the experiential elements we introduce in our *Experience* variations are not as potent as in every day life.

Lastly, except for when statistical or cognitive under-representations were present, we find that the standard inverse S-shaped probability weighting pattern is a good fit for *Description* as well as for *Experience*. We interpret this as evidence that CPT's behavioural tenet that rare events are overweighted, can be found beyond the narrow frame of described uncertainty. Indeed, our average DE gap is best summarized by a relative underweighting hypothesis, whereby rare events are overweighted in *Experience*, only less so than in *Description*. Given how our elicited DE gap is very similar to the literature's average, we suggest that the external validity of this hypothesis goes well beyond the context of this study.

BIBLIOGRAPHY

Abdellaoui, M. (2000). Parameter-free elicitation of utility and probability weighting functions. *Management Science*, 46(11):1497–1512.

Abdellaoui, M., Baillon, A., Placido, L., and Wakker, P. P. (2011a). The Rich Domain of Uncertainty: Source Functions and Their Experimental Implementation. *American Economic Review*, 101(2):695–723.

Abdellaoui, M., Bleichrodt, H., and L'Haridon, O. (2008). A tractable method to measure utility and loss aversion under prospect theory. *Journal of Risk and Uncertainty*, 36(3):245.

Abdellaoui, M., L'Haridon, O., and Paraschiv, C. (2011b). Experienced vs. described uncertainty: Do we need two prospect theory specifications? *Management Science*, 57(10):1879–1895.

Barberis, N. C. (2013). Thirty Years of Prospect Theory in Economics: A Review and Assessment. *Journal of Economic Perspectives*, 27(1):173–96.

Barron, G. and Erev, I. (2003). Small feedback-based decisions and their limited correspondence to description-based decisions. *Journal of Behavioral Decision Making*, 16(3):215–233.

Booij, A. S., Van Praag, B. M., and Van De Kuilen, G. (2010). A parametric analysis of prospect theory's functionals for the general population. *Theory and Decision*, 68(1-2):115–148.

Etchart-Vincent, N. (2004). Is Probability Weighting Sensitive to the Magnitude of Consequences? an Experimental Investigation on Losses. *Journal of Risk and Uncertainty*, 28(3):217–235.

Fehr-Duda, H., De Gennaro, M., and Schubert, R. (2006). Gender, financial risk, and probability weights. *Theory and Decision*, 60(2):283–313.

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10(2):171–178.

Fox, C. R. and Hadar, L. (2006). "Decisions from experience" = sampling error + prospect theory: Reconsidering Hertwig, Barron, Weber & Erev (2004). *Judgment and Decision Making*, 1(2):159–161.

Ghirardato, P. and Marinacci, M. (2001). Risk, ambiguity, and the separation of utility and beliefs. *Mathematics of Operations Research*, 26(4):864–890.

Gigerenzer, G. and Hoffrage, U. (1995). How to Improve bayesian Reasoning Without Instruction: Frequency Formats. *Psychological Review*, 102(4):684–704.

Glöckner, A., Hilbig, B. E., Henninger, F., and Fiedler, S. (2016). The reversed description-experience gap: Disentangling sources of presentation format effects in risky choice. *Journal of Experimental Psychology: General*, 145(4):486–508.

Goldstein, W. M. and Einhorn, H. J. (1987). Expression theory and the preference reversal phenomena. *Psychological Review*, 94(2):236–254.

Gonzalez, R. and Wu, G. (1999). On the Shape of the Probability Weighting Function. *Cognitive Psychology*, 38(1):129–166.

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with ORSEE. *Journal of the Economic Science Association*, 1(1):114–125.

Hau, R., Pleskac, T. J., and Hertwig, R. (2010). Decisions from experience and statistical probabilities: Why they trigger different choices than a priori probabilities. *Journal of Behavioral Decision Making*, 23(1):48–68.

Hau, R., Pleskac, T. J., Kiefer, J., and Hertwig, R. (2008). The description–experience gap in risky choice: The role of sample size and experienced probabilities. *Journal of Behavioral Decision Making*, 21(5):493–518.

Hertwig, R. (2012). The psychology and rationality of decisions from experience. *Synthese*, 187(1):269–292.

Hertwig, R., Barron, G., Weber, E. U., and Erev, I. (2004). Decisions from Experience and the Effect of Rare Events in Risky Choice. *Psychological Science*, 15(8):534–539.

Hills, T. T. and Hertwig, R. (2010). Information Search in Decisions From Experience: Do Our Patterns of Sampling Foreshadow Our Decisions? *Psychological Science*, 21(12):1787–1792.

Kopsacheilis, O. (2018). The role of information search and its influence on risk preferences. *Theory and Decision*, 84(3):311–339.

Luce, R. D. (1991). Rank- and Sign-Dependent Linear Utility Models for Binary Gambles. *Journal of Economic Theory*, 53(1):75–100.

Miyamoto, J. M. (1988). Generic utility theory: Measurement foundations and applications in multiattribute utility theory. *Journal of Mathematical Psychology*, 32(4):357–404.

Murad, Z., Sefton, M., and Starmer, C. (2016). How do risk attitudes affect measured confidence? *Journal of Risk and Uncertainty*, 52(1):21–46.

Patt, A. G. and Schröter, D. (2008). Perceptions of climate risk in Mozambique: implications for the success of adaptation strategies. *Global Environmental Change*, 18(3):458–467.

Quiggin, J. (1982). A theory of anticipated utility. *Journal of Economic Behavior & Organization*, 3(4):323–343.

Rakow, T., Demes, K. A., and Newell, B. R. (2008). Biased samples not mode of presentation: Re-examining the apparent underweighting of rare events in experience-based choice. *Organizational Behavior and Human Decision Processes*, 106(2):168–179.

Tversky, A. and Fox, C. R. (1995). Weighing risk and uncertainty. *Psychological Review*, 102(2):269–283.

Tversky, A. and Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323.

Ungemach, C., Chater, N., and Stewart, N. (2009). Are Probabilities Overweighted or Underweighted When Rare Outcomes Are Experienced (Rarely)? *Psychological Science*, 20(4):473–479.

Wakker, P. P. (2010). *Prospect Theory: For Risk and Ambiguity*. Cambridge University Press.

Waldman, M. (1994). Systematic errors and the theory of natural selection. *American Economic Review*, 84(3):482–497.

Weber, E. U., Shafir, S., and Blais, A.-R. (2004). Predicting risk sensitivity in humans and lower animals: risk as variance or coefficient of variation. *Psychological Review*, 111(2):430–445.

Wulff, D. U., Mergenthaler-Canseco, M., and Hertwig, R. (2018). A meta-analytic review of two modes of learning and the description-experience gap. *Psychological Bulletin*, 144(2):140–176.

3

THE ROLE OF INFORMATION SEARCH AND ITS INFLUENCE ON RISK PREFERENCES

*

'(...) the determination of the value of an item must not be based on its price, but rather on the utility it yields.'

— Daniel Bernoulli, *Specimen Theoriae Novae de Mensura Sortis* (1738)

3.1 INTRODUCTION

Uncertainty pervades almost every sphere of economic activity and understanding and predicting the choices people make under uncertain circumstances has been a central goal for decision theorists. Among the plethora of theories of risky behaviour, Cumulative Prospect Theory (henceforth CPT; Tversky and Kahneman, 1992), has emerged as the descriptive benchmark for laboratory experiments where lotteries' properties (list of all possible outcomes and associated probabilities) are fully described (Barberis, 2013). One of its key tenets is the claim that people tend to overweight low probability events. However, outside of the laboratory people do not often have access to such explicit numerical summaries of uncertainty.

To study more naturalistic situations, psychologists have recently revived the concept of 'Decisions From Experience' (DFE). Within this programme, the 'sampling paradigm' (Hertwig et al., 2004) has emerged as the most common lab-implementation of DFE. Unlike 'Decisions From Description' (DFD) where the properties of lotteries are explicitly described, subjects in DFE have to explore risky options by sampling from their content (in a computerised setting) prior to making a decision. On each screen, there are typically two such options, each with up to two different possible outcomes. Subjects can experience these outcomes and their relative frequency by clicking on each option. Sampling helps subjects decide which option they want to draw from in a final trial involving real monetary consequences. Unlike this final trial, none of the draws during sampling has any monetary effect. Comparing choices

* A version of Chapter 3 has been published in *Theory and Decision*:
Kopsacheilis, O. (2018). The role of information search and its influence on risk preferences. *Theory and Decision*, 84(3):311-339.

between DFD and DFE, a consistent discrepancy has emerged: in DFD - and in accord with CPT's tenets - people make choices as if they *overweight* rare[1] events; whereas, in DFE, it is as if they *underweight* them (Hertwig et al., 2004).

Several studies (e.g. Hau et al., 2008; Ungemach et al., 2009) have since replicated and explored the underpinnings of the 'Description - Experience gap' (DE gap), offering both a wealth of insights and some important open questions (see de Palma et al., 2014, for a recent review). In this study we address some of those questions by conducting a laboratory experiment with three treatments: a standard version of DFD and two variations of DFE. Our contribution to this literature is threefold.

First, we look at sampling patterns in DFE. One of the earliest and most robust findings is that subjects typically rely on small samples where rare events tend to be under-represented (Hertwig, 2012). We investigate how people adjust their search strategy as a function of the rarity of an event by looking at the correlation between sampling amount and a lottery's variance: low variance lotteries in our context contain rarer events.

Lejarraga et al. (2012) study the similar concept of 'experienced outcome variability' which occurs when a subject samples more than one outcome from a given option. The authors find that this variability correlates with higher levels of sampling and conclude that people are motivated to sample more from lotteries for which they have experienced more than one outcome. Mehlhorn et al. (2014) however, question the direction of this causality by pointing to an endogeneity concern: the likelihood of observing more than one outcome increases with the sampling amount. It is therefore possible that high levels of sampling are causing subjects to experience more than one outcome rather than the other way around and conclude that the driver of search effort is 'anticipated' rather than 'experienced' outcome variability. Studying the relationship between sampling amount and variance contributes to this dialogue in the following way. First, variance is a structural property of the lottery and therefore, unlike experienced variability, remains unaffected by sampling amount. Moreover high variance causes variability: a subject is more likely to experience more than one outcome from a '$50-50$' rather than from a '$99-1$' distribution. Therefore, if Lejarraga et al.'s (2012) thesis holds true, we would expect search effort to correlate positively both with experienced outcome variability and with variance. If however sampling amount is positively correlated with variability but negatively correlated with variance, the evidence would favour Mehlhorn et al.'s (2014) objection. Given the relation between variance and rare events, this is equivalent to asking whether subjects sample more from lotteries with rarer events.

1 It is a convention within this literature to refer to events occurring with $p \leq 0.20$ as 'rare'.

Another key novelty of our design is the introduction of a history table in one of our DFE variations: DFE-HT. This table records sampled events and displays them to subjects when they later evaluate the lottery. We examine how its presence influences search by comparing DFE-HT with a more standard version of DFE, DFE-NoHT where there is no such record[2]. One of the reasons we include this table relates to the role of memory constraints. If subjects rely significantly on memorisation during sampling then the history table will help them alleviate part of the associated cognitive load. If this is the case, we would expect to observe larger samples in DFE-HT than in DFE-NoHT. Because the role of memory is elusive to pinpoint (Wulff et al., 2018) we tackle it from two additional angles: by including a test of working memory and by examining whether sampling undertaken just before the moment of decision has more impact than sampling undertaken earlier ('recency effect').

Second, we search for potential differences on revealed preferences between these three ways of acquiring information: from description and from autonomous sampling with or without a history table. We record these preferences via a method of repeated choices between a risky and a safe option (see bisection method under Section 3.2.1). By comparing choice patterns across these three treatments we examine whether there is a DE gap in our data and if so, whether it is amplified or mitigated by the presence of the history table. Moreover, we elicit CPT's components (in the gains domain only) at the individual level. For this we rely on the methodology introduced by Abdellaoui, L'Haridon and Paraschiv (2011b), henceforth AHP, who recently applied the 'source method' (Abdellaoui et al., 2011a; Tversky and Fox, 1995) to study this gap. This method maps different sources of uncertainty (such as DFD and DFE) onto distinct probability weighting functions (weighting functions for short). By examining the shape of the elicited aggregate weighting functions we revisit an interesting tension in this literature: if subjects really underweight in DFE then CPT would prescribe a S-shaped weighting function as opposed to the standard inverse S-shaped curve assigned to DFD. We refer to this potential contrast between the weighting functions in DFD and DFE as the 'Underweighting Hypothesis'.

Recent papers were unsuccessful in validating this pattern. AHP for example report that CPT's standard inverse S-shaped weighting function fits both DFD and DFE well and find that the aggregate weighting function in DFE lies systematically below the function elicited in DFD. They attribute this pattern to a reduced willingness to bet in DFE which is induced by ambiguity aversion: subjects in DFE are less confident about the properties of the sampled options than subjects in DFD. We will refer to this pattern of the DE

2 Hau et al. (2010) use a similar recording device but in their Records-Treatment subjects were only allowed to sample a fixed amount of cards and hence its influence to search cannot be inferred. In our framework subjects can choose instead how much they want to sample.

gap where both weighting functions are inverse-S shaped but that of DFE lies beneath that of DFD as the 'Ambiguity Aversion Hypothesis'.

An attractive feature of AHP's methodology is that it allows the elicitation of decision weighting functions at the individual level both parametrically and non-parametrically. Additionally, this elicitation permits the manipulation of the degree and precision of the elicited curve. We follow this method and address the tension between the two hypotheses regarding the shape of weighting curves. Suspecting that rare events may hold the key to this investigation, we build on AHP's method by eliciting significantly more observations in the neighbourhood of rare events.

Third, we address an important methodological question that derives from AHP's adaptations of the sampling paradigm. There are four noticeable differences between the two approaches. First, if an event is never experienced in the sampling paradigm the subject is likely to remain ignorant about its existence. This is not the case with the AHP method where the list of outcomes is always eventually presented to the subject. Second, in the sampling paradigm sampled events reveal corresponding pecuniary outcomes. In contrast, sampled events are represented by different pairs of colours in the AHP method which are only later associated with monetary outcomes. Third, in the sampling paradigm subjects sample from two options at a time while in AHP only from one. Fourth and perhaps most importantly, there is a sharp distinction between the ways the two methods infer the DE gap. In the sampling paradigm this is done by comparing frequencies with which riskier options are chosen over safer ones between DFD and DFE. This comparison does not need to assume a preference model. In contrast, AHP elicit certainty equivalents (CEs) that make subjects indifferent between keeping or trading the lottery being evaluated. CEs are then used to estimate CPT's weighting functions and the DE gap is inferred by comparing their shape between DFD and DFE.

These differences raise the question of whether the sampling paradigm's DE gap is qualitatively similar to that reported by AHP or perhaps a different phenomenon altogether. We take a first step in answering this question by identifying the key DE gap properties inferred through choice proportion comparisons. We then examine how well they replicate under our valuation framework which is similar to the one AHP used to infer the DE gap in weighting. We do so by exploiting a feature of AHP's implementation of the bisection method: a hybrid between valuation and choice methods that elicits CEs by repeated choices between a risky and a safe option.

In what follows, Section 3.2.1 describes in detail our experimental and elicitation methods. Section 3.3 presents the ensuing results and Section 3.4 discusses their implications. Lastly, Section 3.5 concludes.

3.2 METHODS

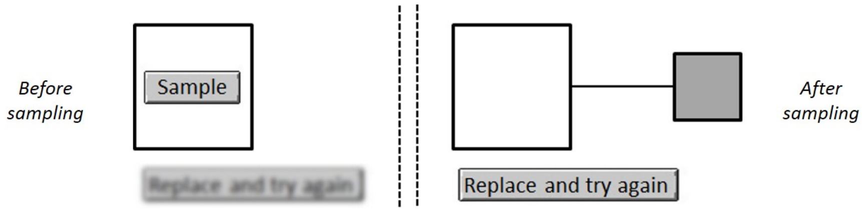
3.2.1 *Experimental design*

We conduct a laboratory experiment with three treatments using a between-subjects protocol[3]. These treatments are: a standard version of DFD and two variations of DFE, DFE-NoHT and DFE-HT.

Treatments consist of 19 time periods and in each period subjects evaluate a lottery. These lotteries are represented by virtual decks of cards, each containing two types of cards demarcated by different pairs of colours. In each period subjects first learn about the relative frequency of each colour. These colours are then linked with monetary outcomes and subjects are asked to evaluate the corresponding lottery.

The key difference between DFD and the two DFE treatments lies in the way subjects learn about these relative frequencies. In *DFD* subjects are informed via numerical descriptions, framed as one shot probabilities (E.g. '90% of the cards are blue and 10% are red'; see Appendix B.1 for an instance of this). In contrast, both DFE-treatments require that subjects find out about these likelihoods by sampling colours from the content of the deck in a separate sampling stage (Fig. 3.1).

Figure 3.1: Sampling stage



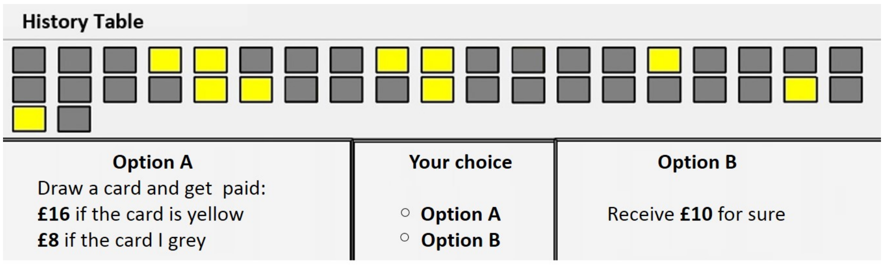
Note. Screen before (left) and after (right) a card is drawn. After drawing a card and seeing its colour, subjects can replace it in the deck where it gets re-shuffled. They can repeat this for as long as they want. This sampling process is identical in DFE-NoHT and DFE-HT and it appears on a separate screen from the evaluation part. Unlike most sampling technologies, there was no time delay between two consecutive draws. Subjects regulate the time the card remains on screen by pressing on the '*replace*' and '*sample*' buttons at their own discretion.

The first 7 periods correspond to lotteries with the same probability distribution (but differing outcomes). To communicate this, subjects in DFE go through only one sampling stage, linked to 7 evaluation parts. Therefore there were only 13 sampling stages in total in DFE. Lotteries and colour-pairs are randomized for each subject across periods. The first 7 lotteries are randomized only within that first cluster.

3 This is a difference with AHP's study which uses a within-subjects design where subjects always made description-based decisions prior to experience-based ones.

The only difference between the two DFE treatments is the presence (or absence) of the history table during the evaluation part. After subjects in DFE-HT finish sampling and proceed to the next screen associated with the evaluation part, they see a table that has recorded the colours of cards they encountered during sampling, in the order they saw them (see Fig. 3.2). This history table could only record up to a fixed number of cards. When during sampling this capacity was reached, a message appeared on screen informing subjects that they can continue sampling should they want to, but that their observations past this point would not be recorded. We chose a maximum capacity of 57 with the intention of avoiding a straight-forward calculation of a relative frequency in numeric form, resembling the information in DFD.

Figure 3.2: Evaluation part in *DFE-HT*



Note. Sampled events from the sampling stage are recorded and dispayed on the top of the evaluation screen in DFE-HT. This part of the screen remained blank in DFE-NoHT.

The evaluation protocol is common for all three treatments. In this section events (such as 'Drawing a yellow card') are associated with monetary consequences. We use the bisection method as applied by AHP to elicit CEs for each lottery. An instance of this can be seen at the bottom of Fig. 3.2. Every bisection process starts with a choice between a lottery and its expected value offered with certainty. In our experiment there were 5 such iterations for each lottery. Lotteries are presented under Option A while certain amounts under Option B. The method proceeds by updating Option B until a value close to indifference is reached. In Fig. 3.2's example, if the subject chooses Option B then the certain outcome will be updated to £9, the midpoint between the lowest outcome of the lottery and the certain outcome that was just chosen. If instead Option A is selected, then Option B will be updated to £13, the midpoint between the highest outcome of the lottery and the certain outcome that was just rejected.

This elicitation through iterative one-shot choices makes the bisection robust against the criticism that methods such as the multiple price list have received (see Erev et al. (2008) for such a criticism). Most importantly for our analysis is the fact that the very first choice in each new evaluation is always between a lottery and a monetary outcome of equivalent expected value (EV) offered with certainty. This is much like the setup that studies in the sampling paradigm have used to infer the DE gap in choice.

Finally, after all lotteries were evaluated, subjects go through a standard forward digit span task where they are asked to recall sequences of digits. Reporting correctly a digit awards the participant a point[4] and increases the sequence by one digit. After three errors the process is terminated. We use this task as a proxy for memory capacity.

Sessions were conducted in CeDEx's laboratory at the University of Nottingham. All treatments were programmed in z-Tree (Fischbacher, 2007). In total, 118 subjects took part in only one of these three treatments: 40 in *DFE-HT*, 39 *DFE-NoHT* and 39 in *DFD*. We used ORSEE (Greiner, 2015) for the recruitment process. At the end of the experiment one question was randomly selected and each subject would get paid according to their choice in that question. Average payment was £11, including a £3 participation fee, for approximately one-hour sessions.

3.2.2 *Elicitation of CPT in DFD and DFE*

3.2.2.1 *Preliminaries*

Let $x_{E_p}y$ stand for a binary lottery where x, y are non-negative outcomes[5] contingent on mutually exclusive events and $x > y$. E_p represents an event occuring with objective probability p and the high (or desirable) outcome x is always contingent to E_p. According to CPT, given a strictly increasing utility function: u and a weighting function W, subjects maximize:

$$x_{E_p}y \mapsto W(E_p)u(x) + (1 - W(E_p))u(y) \tag{3.1}$$

To make (3.1) operational we use the two-stage model idea proposed by Tversky and Fox (1995) and later developed into the 'source method' by Abdellaoui et al. (2011a). According to this model a decision maker first forms a subjective belief for an uncertain event ($P(E_p)$) and then transforms this value into willingness to bet via a probability weighting function:

$$W(E_p) = w_\sigma(P(E_p)) \tag{3.2}$$

In (3.2), $w_\sigma(\cdot)$ is the probability weighting function which depends on σ, the source of uncertainty. Applying (3.2) to (3.1) we get:

$$x_{E_p}y \mapsto w_\sigma(P(E_p))u(x) + (1 - w_\sigma(P(E_p)))u(y) \tag{3.3}$$

4 This task was not monetarily incentivised.
5 Restricting analysis to gains reduces CPT to the Rank Dependent Utility model (Quiggin, 1982).

We can break down (3.3) into: (i) utility over monetary outcomes, $u(\cdot)$, (ii) probability measure over outcome distribution, $P(\cdot)$ and (iii) source-dependent probability weighting function, $w_\sigma(\cdot)$. The source method adjusts this third component according to the environment where the risky choice takes place.

In DFD we are in an environment where probabilities are completely known and so $p = E_p$. When analysing DFE on the other hand, we are referring to an environment where probabilities cannot be calculated exactly but can instead be assessed in an empirical manner by the subject. To apply (3.3) in DFE, given that the belief $P(E_p)$ is essentially unobservable, we consider the following two proxies: objective (or true) probability (p) and experienced probability (f_p). The latter stands for the relative frequency with which an event has been observed in a sample.

Using true probabilities as proxies for beliefs, although convenient and widely used in this literature, can be problematic - especially in cases where sampling bias is prevalent. Therefore our analysis proceeds by reporting (mostly) experienced probabilities. Although this proxy might still not be perfect, there has been evidence for a high correlation between elicited beliefs and f_p (Fox and Hadar, 2006).

3.2.2.2 *Estimation*

Our approach is based on AHP's adaptation of the semi-parametric method developed by Abdellaoui et al. (2008). We use 16 lotteries (Table 3.1; lotteries 1-16) which we separate into two clusters. In the first cluster subjects evaluate 7 lotteries with a fixed probability $p = 0.25$. The reported CEs are then used for the estimation of a utility function. Assuming the power-function specification: $u(x) = x^\alpha$, we need only estimate $(W(E_{0.25}), \alpha)$ for each subject, where α captures the curvature of the utility function and $W(E_{0.25})$ the weight assigned to $E_{0.25}$. We do so by minimizing the non-linear least square function: $\|z - \hat{z}\|^2$, where z_i refers to the observed CE and $\hat{z}_i$:

$$\hat{z}_i = [W(E_{0.25})(x_i^\alpha - y_i^\alpha) + y_i^\alpha]^{\frac{1}{\alpha}} \tag{3.4}$$

Table 3.1: Lotteries

	Utility							Decision weights									Control		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
x	4	8	12	16	16	16	16	16	16	16	16	16	16	16	16	16	3	4	4
p	.25	.25	.25	.25	.25	.25	.25	.025	.05	.10	.25	.50	.75	.90	.95	.975	.25	.20	.80
y	0	0	0	0	4	8	12	0	0	0	0	0	0	0	0	0	0	0	0

Note. Lotteries 1-7 were used to estimate the utility function while lotteries 8-16 to elicit weighting functions (both parametrically and non-parametrically). Lotteries 17-19 are not relevant for the estimation and were included as control tasks due to their similarity with some of the commonly used lotteries in the early DE gap studies.

In the second cluster subjects evaluate a total of 9 lotteries with fixed high ($x = £16$) and low ($y = £0$) outcomes and varying p. Subsequently, using the estimated α from the first cluster of lotteries, we can control for risk curvature and calculate non-parametrically decision weights[6] for each level of p.

Let z_j' stand for the observed CE elicited from this second cluster of lotteries. Then from (3.4) we get that:

$$W(E_{p_j}) = \left(\frac{z_j'}{16}\right)^{\alpha}, forj = 1, ..., 9 \tag{3.5}$$

Finally, we used these decision weights in order to fit the following two-parameter, linear-in-log-odds weighting function introduced by Goldstein and Einhorn (1987).

$$w(p) = \frac{\delta p^{\gamma}}{\delta p^{\gamma} + (1-p)^{\gamma}} \tag{3.6}$$

This is the same weighting function that AHP used. Parameter γ controls curvature with $\gamma < 1$ indicating an inverse S-shaped weighting function while $\gamma > 1$ a S-shaped one (values close to 1 point to no curvature). Parameter δ controls elevation with $\delta < 1$, $\delta > 1$ and $\delta = 1$ pointing to 'low', 'high' and 'no' elevation respectively. Gonzalez and Wu (1999) offer an interesting psychophysical interpretation for these parameters according to which γ is interpreted as a measure of probabilistic sophistication while δ as a degree of optimism.

3.3 RESULTS

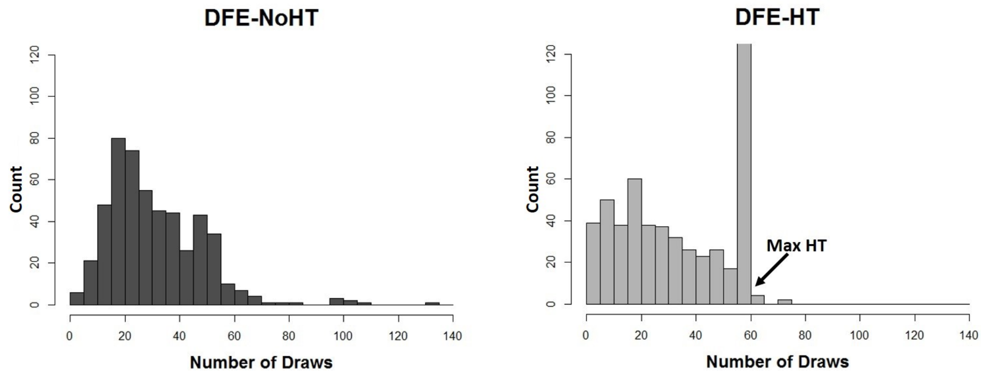
3.3.1 *Sampling*

We start by comparing sampling patterns between the two DFE treatments. Fig. 3.3 foreshadows the importance of the history table in influencing subjects' search.

In Fig. 3.3, sampling amounts for each subject and in each period are plotted across the two treatments. The spike in DFE-HT occurs right when the participant has filled this sampling-round's history table. We infer from this that the history table's maximum capacity (always set at 57 draws) was a very potent cue for search termination in *DFE-HT*. In its absence, participants' search-effort followed a more normal-like distribution.

6 Similarly to AHP, decision weights based on E_p for probability targets that were not represented in subjects' samples, were obtained by a linear interpolation of the weighting function at the individual level.

Figure 3.3: Distribution of draws across DFE-treatments



Note. 'Max HT' points to the maximum capacity of the history table (57 draws). Subjects could sample past that point but their observations would not be recorded in the history table.

Variance and experienced variability

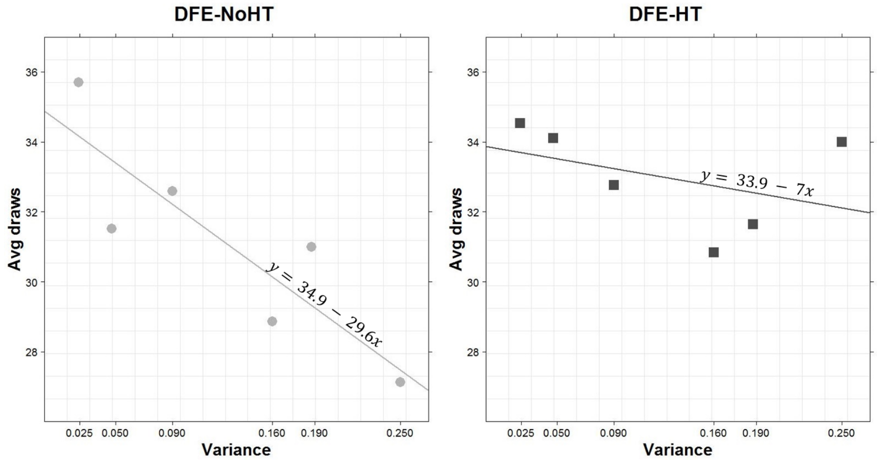
We first examine the effect of experienced-outcome[7] variability (variability for short). Following Lejarraga et al. (2012) we distinguish between: positive variability if someone sampled more than one type of cards in a deck and no variability otherwise. Comparing the means of these two groups we verify that experiencing positive variability correlates positively with higher amounts of sampling. Specifically, sampling amount for positive variability averaged 33.5 draws per lottery while that for no variability 19.5 (p-value < 0.01, MW test).

We turn next to the relation between sampling amount and a lottery's variance where we compute averages of sampling amount for each level of variance and examine how the two correlate in each sampling treatments. Subjects only sampled binary lotteries and hence their variance was always strictly positive. As mentioned earlier, low variance is associated with rare events. For example a binary lottery offering 1 with probability p and 0 otherwise has variance: $p(1-p)$ which is maximized when $p = 1/2$, i.e. when the rarity of the rarer event is minimized.

In both DFE treatments variance correlates negatively with search effort. Interestingly, this correlation is significant in DFE-NoHT (Spearman's $\rho = -0.89$, p-value $= 0.03$) but not in DFE-HT ($\rho = -0.6$, p-value $= 0.24$). Fig. 3.4 displays this information.

7 In our study this phenomenon is more accurately described as experienced *event* variability. Outcomes refer to monetary consequences while in our sampling stage subjects sampled events which were only later assigned to outcomes.

Figure 3.4: Average sampling amount over levels of variance



Note. Points represent average sampling - across all subjects - for different levels of variance in *DFE-NoHT* (left panel) and *DFE-HT* (right panel). The solid straight lines have been estimated by OLS at the aggregate level. Lotteries like: $(x, E_p; y)$ and $(x, E_{1-p}; y)$ are indistinguishable during sampling and were pooled together. Lotteries and hence levels of variance were randomized for each subject and so this effect is independent of time period.

Individual level analysis corroborates this finding. We estimate slopes for each subject from a simple linear regression, where average sampling over all rounds is regressed on levels of variance (a slope similar to the one in Fig. 3.4 but for each individual). Although average slopes are negative in both treatments (*DFE-NoHT*:-29.61 vs. *DFE-HT*:-6.98), only in DFE-NoHT this coefficient is significantly smaller than 0 (p-value < 0.01 for DFE-NoHT and p-value $= 0.146$ for DFE-HT, one-sided MW tests). Moreover, the slope is steeper in DFE-NoHT than in DFE-HT (p-value $= 0.043$, one-sided MW-test). Estimating rank correlation coefficients instead of slopes replicates this analysis. In both treatments the average correlation is negative (DFE-NoHT:-0.219, DFE-HT:-0.053) but only in DFE-NoHT this coefficient is significantly smaller than 0 (p-value < 0.01 for DFE-NoHT and p-value $= 0.259$ for DFE-HT, one-sided MW-tests). Recall from the introduction that lotteries with rarer events are associated with lower variance. With this in mind we can state the following:

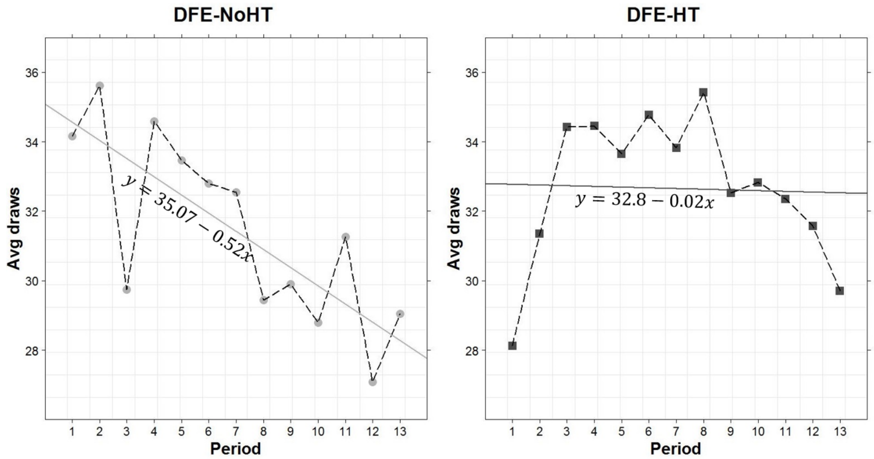
Result 1. *Decks containing rarer events instigate higher search-effort. The history table partially mitigates this effect.*

Result 1 runs opposite to Lejarraga et al.'s (2012) hypothesis that experienced variability causes higher amount of sampling. We return to this point in the Discussion.

Time periods

Fig. 3.5 plots average sampling amount over time. We see that in DFE-NoHT there is a clear negative trend: subjects possibly get tired of sampling over time. In DFE-HT the pattern is inverted U-shaped. It is possible that subjects realize the benefits of the history table after the end of the first sampling round and adjust their strategy to collecting larger samples. After this original upwards-adjustment, sampling amount stabilizes at a high level until it eventually decays in the last periods.

Figure 3.5: Average sampling amount over periods



Note. Points represent average sampling - across all subjects - for different time-periods in DFE-NoHT (left panel) and *DFE-HT* (right panel). Arguably the OLS at the aggregate level that is used to plot the solid straight lines is not informative for the DFE-HT treatment where the shape is inverted U.

We detect a significant negative time trend in search effort in *DFE-NoHT* ($\rho = -0.78$; p-value < 0.01). We found no significant such trend in *DFE-HT* ($\rho = -0.13$, p-value $= 0.66$). This is most likely due to the fact that with the exception of the first and last periods, sampling amount remained relatively unaffected by time in DFE-HT. Comparing the variances of average sampling amounts from periods 2 to 12, we find that the variance in DFE-HT (1.81) is smaller than the one in DFE-NoHT (6.97). Levene's test for variance equality shows that the two variances are significantly different (p-value $= 0.028$). When we look only at the second half of the time periods, we verify that eventually time affected subjects in DFE-HT too (Spearman's $\rho = -0.90$, p-value < 0.01). In summary:

Result 2. *Sampling amount diminishes over time. This effect is less prominent in DFE-HT.*

Slope and rank correlation analysis at the individual level verify this result. For brevity we report only rank correlation coefficients. For DFE-NoHT this coefficient was on average significantly smaller than 0 ($\rho = -0.14$, p-value

= 0.033, one-sided MW-test) and significantly smaller than the average for DFE-HT (p-value = 0.033, one-sided MW-test). The average rank correlation coefficient for DFE-HT is not significantly different than 0 ($rho = 0.04$, p-value = 0.492) but once again, when we focus on the second half of the periods, it becomes significantly (albeit weakly) negative ($\rho = -0.127$, p-value = 0.051, one-sided MW-test).

Memory

We examine whether the history table boosted search effort across the two treatments. First, we find that the median sampling amount across both treatments was 30, which is unusually high. This number was 7 ± 2 in most studies in the sampling paradigm (Hertwig and Pleskac, 2010) and between 15 and 21 in AHP. Consequently, in the current study subjects did not sample both types of cards in only 10% of the cases (9% in DFE-HT, 11% in DFE-NoHT). In Hertwig et al. (2004) that number is 44%. Nevertheless, sampling levels were not significantly different between the two DFE treatments. The median number of draws for *DFE-HT* was 30 while that for *DFE-NoHT* was 28 (p-value= 0.158, two-sided MW-test). Moreover, the forward digit span task, which served as our proxy for working memory, did not correlate with sampling amount in either treatment ($\rho = 0.13$, p-value = 0.41 and $\rho = 0.22$, p-value = 0.15 for *DFE-NoHT* and *DFE-HT* respectively).

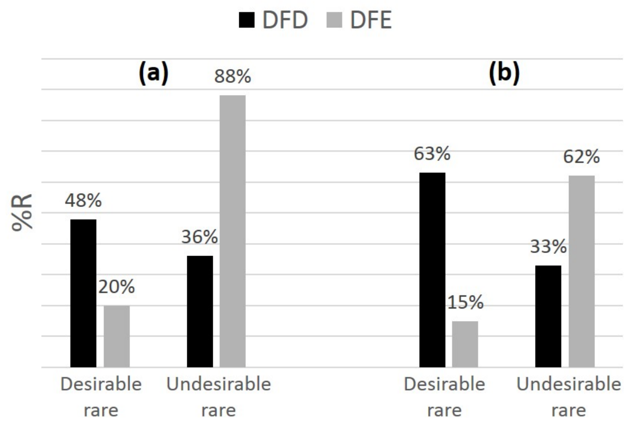
3.3.2 *Choices and preferences*

3.3.2.1 *The DE gap in choice*

In this section we examine the DE gap in choice over lotteries without the mediation of a preference model. We first look at the choice patterns reported by two important early studies in this literature: Hertwig et al. (2004) and Hau et al. (2008). These studies share a common set of decision problems where a subject is asked to choose between two options with similar EV but differing variance. We refer to the high variance option as 'Risky' and the low variance option as 'Safe'. To increase comparability with our study we consider only those decision problems that entail non-negative outcomes and where the 'Safe' option is a certain outcome (see Appendix A/ Table B.2 for the full list of decision problems). This restricts the analysis to 2 decision problems (from a total of 6) which we then characterize according to the desirability of the rare outcome of the 'Risky' option. Decision problems with a rare (un)desirable outcome are referred to as '(un)desirable rare'. Let '%*R*' stand for the percentage with which subjects chose 'Risky' over 'Safe'. Fig.

3.6 plots $\%R$ across treatments in these two studies for 'desirable rare' and 'undesirable rare'.

Figure 3.6: Choice patterns in early DE gap studies



Note. Percentage choosing 'Risky' over 'Safe' across studies ((a) and (b)), treatments (DFD and DFE) and decision problems ('desirable rare' and 'undesirable rare'). (a) Hertwig et al. (2004) (b) Hau et al. (2008). *Desirable rare*: Risky$=(32, E_{0.1}; 0)$ vs. Safe$=(3, E_{1.0})$. *Undesirable rare*: Risky$=(4, E_{0.8}; 0)$ vs. Safe$=(3, E_{1.0})$.

Table 3.2 lists the properties of the early DE gap according to the observed choice-patterns. Properties 1 and 2 derive from comparisons between DFD and DFE while Properties 3 and 4 from comparisons within each treatment. Property 1 is that people choose 'Risky' over 'Safe' more often in DFD than in DFE when the rare outcome is desirable while Property 2 is that the opposite holds true when the rare outcome is undesirable instead. Property 3 is that subjects in DFD choose 'Risky' over 'Safe' more often when the rare outcome is undesirable than when it is desirable while Property 4 is that this pattern is reversed when subjects make decisions in DFE.

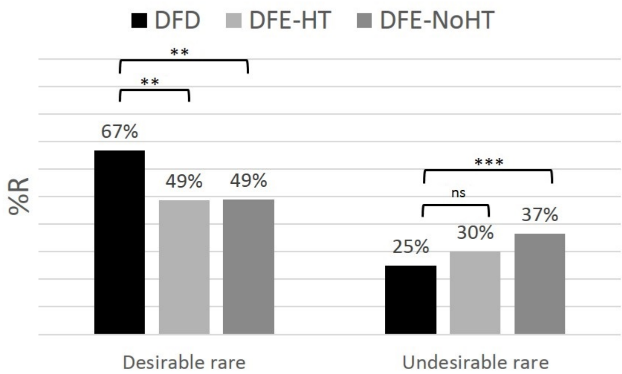
Table 3.2: Properties of the original DE gap in choice

#	Choice pattern	Condition
1.	$\%R_{DFD} > \%R_{DFE}$	Desirable rare
2.	$\%R_{DFD} < \%R_{DFE}$	Undesirable rare
3.	$\%R_{Desirable} > \%R_{Undesirable}$	DFD
4.	$\%R_{Desirable} < \%R_{Undesirable}$	DFE

Fig. 3.7 plots results from our study using an analysis similar to that summarized in Fig. 3.6. Recall that although our method relies on lottery-valuations, these valuations take place via repeated choices. For this analysis we use lotteries 8-16 from Table 3.1). These are the same lotteries which we later use in order to elicit weighting functions and therefore appropriate to compare the two types of DE gap: that inferred by choice-patterns (sampling paradigm) and that inferred by weighting patterns (AHP). We separate these lotteries

into two clusters: those with $p \leq 0.25$ and those with $p \geq 0.75$. Decision problems entailing a choice between a lottery with $p \leq 0.25$ and its EV are characterised as 'desirable rare' since the rare[8] event is associated with the high outcome (£16). Decision problems entailing a choice between a lottery with $p \geq 0.25$ and its EV are characterised as 'undesirable rare' since the rare event is associated with the low outcome (£0). This analysis leaves out only the lottery with the $50-50$ distribution where no event can be considered to be rarer than the other.

Figure 3.7: Choice patterns in this study



Note. Percentage choosing 'Risky' (%R) over 'Safe' in the current study across treatments (DFD, DFE-HT and DFE-NoHT) and types of decision problems (desirable and undesirable rare). 'Risky' refers always to the lottery and 'Safe' to its expected value. We consider lotteries 8-16 from Table 3.1 and cluster choices in the following way: *'Desirable rare'*: $Risky = (16, E_p; 0)$ for $p \leq 0.25$. *'Undesirable rare'*: $Risky = (16, E_p; 0)$ for $p \geq 0.75$. 'ns': not significant, '***': p-value< 0.01, '**': p-value< 0.05

According to Fig. 3.7, choice patterns in DFD are significantly different than in DFE-HT and DFE-NoHT for 'desirable rare' (p-value $= 0.016$ and p=value=0.018 for DFD vs. DFE-HT and DFD vs. DFE-NoHT respectively, two-sided MW-test[9]). For the 'undesirable rare' however, only the DFD vs. DFE-NoHT comparison is significant (p-value < 0.01 for DFD vs. DFE-NoHT and p-value $= 0.504$ for DFD vs. DFE-HT, two-sided MW-test). Result 3 summarizes this analysis.

Result 3. *Both versions of DFE generate a significant DE gap. This gap is smaller in the presence of the history table.*

Moreover, comparing the choice patterns in Fig. 3.7 with Table 3.2 we verify that 3 out of these 4 properties of the early DE gap hold in this analysis. However, the fact that %R in DFE is higher in the 'desirable rare' than in the 'undesirable rare' violates Property 4. With this in mind, we claim that:

8 We replicate our results when considering stricter thresholds for rare events such as $p < 0.25$ or $p < 0.10$.

9 To control for repeated observations, we first compute the average of %R for each individual and then compare these averages across treatments

Result 4. *The DE gap we capture in this study is qualitatively similar but not identical to the original phenomenon.*

We examine two hypotheses for the low level of $\%R_{Undesirable}$ in DFE. First we consider the possibility that this is due to the asymmetry in the EV of the risky option between early DE gap studies (3.2) and the current one (14.3 on average). Second, we conjecture that the difference is driven by information-asymmetries between the two paradigms: unlike the sampling paradigm, subjects in our study were always informed about the existence of the second outcome. Moreover, due to the higher levels of sampling we recorded, rare events were under-represented less often than in earlier studies.

With respect to the first hypothesis, we examine choices from control lottery: $(4, E_{0.8}; 0)$ and observe that the pattern is very similar to that in Fig. 3.7 ($\%R_{DFD} = 26\%, \%R_{DFE-HT} = 30\%, \%R_{DFE-NoHT} = 31\%$; see Appendix A/ Table B.3 for details on the choice patterns of all 'control' lotteries). For the second hypothesis we repeated the analysis in Fig. 3.7 but considering only cases in which the probability of the rare event has been under-represented. We see that in this case all 4 properties of the early DE gap hold for the comparison between DFD and DFE-NoHT (but still not for that between DFD and DFE-HT; see Appendix A/ Fig. B.2) and therefore conclude that the second hypothesis is more likely to be the explanation behind the violation of Property 4.

One last thing to notice about Fig. 3.7 is that risk aversion (as inferred by $\%R$) is probability dependent. In DFD subjects seem to be overall risk seeking ($\%R > 50\%$) for small gain probabilities (i.e. when the rare event is desirable) but risk averse ($\%R < 50\%$) for high gain probabilities (i.e. when the rare event is undesirable). This is in accord with CPT's fourfold pattern. In DFE, subjects seem to be overall risk neutral ($\%R \simeq 50\%$) for small gain probabilities but risk averse (albeit comparatively less so than in DFD) for high gain probabilities.

3.3.2.2 *The DE gap in preferences*

We proceed by incorporating in the analysis all iterations of the bisection and extracting a CE for each lottery. We use these CEs in order to estimate CPT's components as described under Section 3.2.2.2. We start by comparing utility curvature (α) across treatments. Median values in all treatments suggest a near linear utility curvature (Table 3.3). These values are higher than those reported by AHP ($\alpha = 0.79$ for DFD and $\alpha = 0.82$ for DFE) as well as than the usual values reported by studies with medium to low awards (slightly less than 1; see Booij et al., 2010). They are nevertheless within the typically reported range (see Epper et al., 2011; Murad et al., 2016 for values of α slightly higher than 1). By classifying subjects according to utility curvature ($\alpha < 0.9$

as concave, $\alpha \in [0.9, 1.1]$ as linear and $\alpha > 1.1$ as convex), we see that overall most of the subjects (57%) are best characterized by a utility function that is either concave or linear rather than convex (see Appendix A/ Table B.4 for more details). There were no significant differences between α's across treatments (p-value= 0.77, Kruskal-Wallis).

Table 3.3: Median estimates of $(\alpha, \delta, \gamma,)$

Treatment	Utility Curvature (α)	Weighting Elevation (δ)	Weighting Curvature (γ)
DFD	1.06	0.53	0.49
	(0.37)	(0.13)	(0.12)
DFE-HT	1.06	0.48	0.52
	(0.35)	(0.10)	(0.07)
DFE-NoHT	1.02	0.44	0.67
	(0.35)	(0.11)	(0.08)

Note. For DFE treatments, the δ's and γ's are estimated according to experienced probabilities. Median standard errors from the estimation procedure are reported in parentheses. Overall, parameters were equally dispersed across treatments; equality of variance was never rejected (p-value=0.199 for α, 0.722 for γ and 0.804 for δ, Levene's tests.) Interquartile ranges were: $[0.83 - 1.49]$ for α, $[0.20 - 0.91]$ for δ and $[0.37 - 0.87]$ for γ.

Having estimated α, we can use Equation 3.5 to calculate decision weights for each individual. Treatment-level weighting functions can be obtained either by aggregating weights across subjects for each level of probability (non-parametric analysis) or by fitting the parameters from Equation 3.6 for each subject and aggregating (γ, δ) across all subjects (parametric-analysis)[10]. We begin with the latter.

Parametric analysis

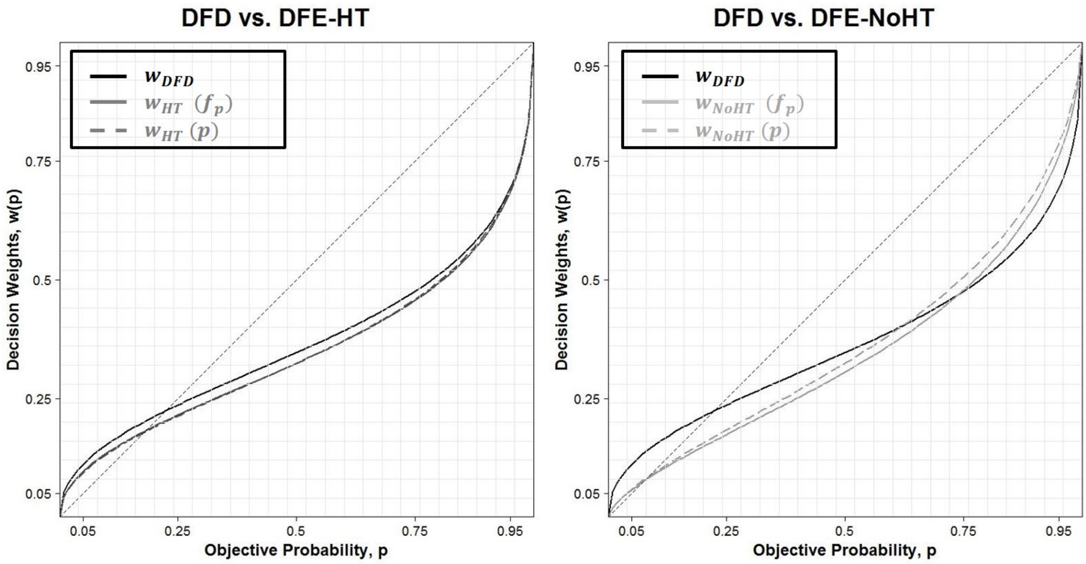
Kruskal-Wallis tests detect significant differences between and γ-values across the three treatments (p-value $= 0.038$) but not for δ-values (p-value $= 0.501$). Focusing on γ's, the difference between γ_{DFD} and γ_{NoHT} is significant (p-value= 0.015 ,two-sided MW-test) while that between γ_{DFD} and γ_{HT} only weakly so (p-value= 0.065, two-sided MW-test). Moreover the hypothesis that $\gamma_{NoHT} = \gamma_{HT}$ cannot be rejected (p-value $= 0.485$, two-sided MW test).

Fig. 3.8 plots differences in weighting between description and the two versions of experience: with (left panel) and without (right panel) a history table. The proximity between the experienced-based parameter estimates (solid lines) and objective-based such estimates (dashed lines), holds testament to

10 See Appendix A/ Fig. B.3 for a demonstration of this process at the individual level.

the high amount of sampling which brought experienced and objective probabilities very close. In fact, with the exception of $p = 0.975$ for DFE-HT, we were never able to reject the hypothesis that $f_p = p$ (see Appendix A/ Table B.1 for details). A corollary to this is that the role of sampling bias was - at least at the aggregate level - quite limited.

Figure 3.8: Comparison of parametric weighting functions between DFE and DFD



Note. For DFE, dashed lines are estimated according to true probabilities (p) while solid lines are based on experienced probabilities (f_p).

Unlike what the 'underweighting hypothesis' would have predicted, Fig. 3.8 suggests that the common inverse S-shaped weighting function accommodates well DFD as well as both DFE treatments. Moreover, the relation between w_{DFD} and both versions of w_{DFE} provides little support for the 'ambiguity aversion hypothesis' according to which w_{DFE} should lie beneath w_{DFD} throughout the probability interval. Although this is true for small to medium values of p, the pattern reverses for high values of p (this is arguably clearer in the case of *DFE-NoHT* where the turning point occurs somewhere in $p \in [0.6, 0.8]$). Keeping in mind that rare events are located near the edges of the probability interval (desirable rare events close to $p = 0$ and undesirable rare events close to $p = 1$), we can summarize Fig. 3.8's pattern as follows:

Result 5. *The 'Relative Underweighting Hypothesis': Subjects overweight rare events in DFD and in DFE; this overweighting is less pronounced in DFE.*

At the individual level, we categorize the curvature of weighting functions as 'inverse-S' when $\gamma < 0.9$, as 'S-shaped' when $\gamma > 1.1$ and as 'no curvature' when $\gamma \in [0.9, 1.1]$. There are approximately twice as many subjects compatible with an 'S-shaped' weighting function in DFE (DFE-HT: 10 subjects, DFE-NoHT: 9 subjects) as in DFD (5 subjects). Interestingly, most of these S-shaped curves come from subjects who sampled less than the median amount of that treatment: 60% in DFE-HT and 89% in DFE-NoHT (see more details of this classification in Appendix A/ Table B.4). A rank correlation test between sampling behaviour (1 if someone sampled less or equal to the median amount and 0 if more) and curvature of the weighting function (1 if $\gamma > 1.1$ and 0 otherwise) verifies that there is a significant correlation between the two ($\rho = 0.318, p - value < 0.01$)[11]. No such correlation was detected for similar classifications of δ ($\rho = -0.038, p - value = 0.736$).

Result 6. *S-shaped weighting curves are more common to subjects who sample less.*

Result 6 may be very useful in explaining why we find so little support of the 'underweighting hypothesis'; we return to this point in the discussion section.

Non-parametric analysis

Table 3.4 reports average decision weights - computed according to experienced probabilities (f_p) - across individuals according to probability level and treatment[12].

Table 3.4: Non-parametric decision weights (averages)

Probability p	DFD w_{DFD}	DFE-HT w_{HT}	DFE-NoHT w_{NoHT}	DFD vs. p t_{39}	DFE-HT vs. p t_{40}	DFE-NoHT vs. p t_{39}
0.025	0.16	0.11	0.09	4.58***	2.38**	2.35**
0.05	0.18	0.16	0.12	4.24***	2.99***	2.82***
0.10	0.20	0.18	0.15	3.23***	2.36**	2.12**
0.25	0.25	0.21	0.21	-0.09^{ns}	-1.43^{ns}	1.44^{ns}
0.50	0.36	0.30	0.31	-3.64***	-5.52***	-5.52***
0.75	0.49	0.47	0.49	-6.41***	-6.31***	-6.31***
0.90	0.63	0.65	0.65	-5.48***	-5.20***	-5.19***
0.95	0.60	0.69	0.69	-7.11***	-5.07***	-5.02***
0.975	0.71	0.72	0.74	-5.43***	-4.07***	-4.07***

Columns 2-4: Average decision weights for each level of probability (p). For brevity, only weights that have been estimated according to experienced probabilities (f_p) are reported. Columns 5-7: two-sided t-statistics for the comparison with the identity line. With the exception of $p = 0.25$, low probabilities are significantly overweighted (see '$+$' sign on t-statistic) and medium to high probabilities significantly underweighted (see '$-$' sign on t-statistic). Two sided MW-tests confirm this analysis.

Qualitatively the non-parametric analysis corroborates Result 5: aggregate decision weights point to inverse S-shaped weighting functions in all treat-

11 This result is robust for different classifications of S-shaped curves such as with $\gamma > 1$. We used γ's based on decision weights from objective probabilities. The reason we did not use weights corrected for f_p was so that we capture the effect of mis-representing objective probabilities (p) in low sampling cases. When we perform the same analysis adjusted for f_p we find that the correlation is reduced but still significant ($\rho = 0.245$, p-value $= 0.029$).

12 See Appendix A/ Table B.5 for a comparison of the median decision weights between this study and AHP.

ments with a cross-over point in the vicinity of $p = 0.25$. This is supported by statistical analysis comparing decision weights with the diagonal (see last three columns of Table 3.4). Moreover, this overweighting appears to be partially mitigated for rare events: $w_{DFE} < w_{DFD}$ for $p < 0.25$ and $(1 - w_{DFE}) < (1 - w_{DFD})$ for $p > 0.75$. Statistical analysis however warrants a note of caution regarding the last assertion. A 3×9 ANOVA does not detect any significant differences between the 3 treatments (p-value=0.412)[13].

3.3.2.3 *Recency effects*

We explore whether events experienced towards the end of the sampling process influenced choices more than events that were sampled in the beginning. Incorporating a similar idea with AHP, we first compute decision weights corresponding to the experienced probability of the first (f^1) and second (f^2) half of the sampling process.[14] We then compare those decision weights with that corresponding to the experienced probability of the entire sampling process (f). Had recency effects been present, we would expect the final decision weight to be closer to the decision weight from the second half so that: $|w_\sigma(f) - w_\sigma(f^1)| > |w_\sigma(f) - w_\sigma(f^2)|$.

Notwithstanding, parametric (2×9 ANOVA with repeated measures for the first and second half) and non parametric (two-sided MW-tests for each level of p with Bonferroni corrections) did not detect significant asymmetries between the early and the later observations of the sampling process. We thus conclude that there were no recency effects.

3.4 DISCUSSION

Variance vs. Variability

We began by exploring the effect on sampling amount of two related concepts: experienced event variability and a lottery's variance. We verify that experienced variability correlates with higher levels of sampling. Does that mean however that experiencing variability *causes* subjects to sample more as Lejarraga et al. (2012) have claimed? Or is it rather that high levels of sampling lead subjects to sample more than one event? To clarify the direction

13 Similarly, conducting two-sided MW-tests with Bonferroni corrections, we can never reject the hypothesis that decision weights are equal between DFD and DFE-HT nor between DFD and DFE-NoHT for any level of p.

14 AHP use a similar approach but by comparing absolute differences between revealed and experienced probabilities. Revealed probabilities are estimates of $P(E_p)$, the likelihood assigned by the subject to event E_p (see expression (3.2) in Section 3.2.2.1). For more details see AHP pp. 1890.

of causality we examined the role of variance which is a proxy for experienced variability: lotteries with higher variance are more likely to generate experienced variability. At the same time, unlike experienced variability, variance is a structural property of the lottery and thus cannot be affected by the amount of sampling. In our setting, low variance is associated with rarer events. Therefore, if experienced variability causes higher levels of sampling, we would expect high-variance lotteries to be associated with higher levels of sampling. Instead, Fig. 3.4 and Result 1 point to the opposite: subjects sample more from lotteries with low variance, or equivalently, lotteries containing rarer events. According to a property of the binomial distribution, rare events tend to be revealed later on during search. Consequently, Result 1 has more in common with Mehlhorn et al.'s (2014) suggestion that it is anticipated rather than experienced variability that instigates higher levels of sampling.

Does the History Table crowd out attention from the sampling process?

As Result 1 suggests, the increased sensitivity towards rare events was attenuated in the anticipation of the history table. Result 2, highlights another such search-policy rigidity in *DFE-HT*. Unlike the clear negative time-trend in DFE-NoHT, average sampling in DFE-HT has a significantly less steep decline. In fact, excluding first and last periods, average sampling remained relatively stable in DFE-HT (we observed significantly lower variance of average sampling compared to DFE-NoHT during these periods). One possible overarching explanation for these results is that the anticipation of the history table makes cues unrelated to it less salient. Fig. 3.3 can perhaps be interpreted along these lines. The frequency with which subjects in *DFE-HT* chose to collect a sample just equal to the table's maximum capacity, corroborates the hypothesis that cues such as time and variance were overriden by that of filling up the history table.

Memory limits

Taking into account their elusive nature we chose to approach the potential effects of memory bounds from three different angles. First, we asked whether alleviating the cognitive load of memorizing via the history table can boost search effort. Second, we examined whether individual idiosyncratic memory capacity correlates with the size of drawn samples. Finally, we examined whether later observations exert more influence on final decisions when compared to earlier ones. Despite this multidimensional approach we were unable to detect a clear effect in all three accounts. Subjects' sample size did not vary significantly between DFE-HT and DFE-NoHT nor did it correlate with the forward digit span task. Lastly, we find no evidence for recency effects.

Given the intuitive appeal of the role of memory bounds this absence of effects may seem counter-intuitive. This impression is only strengthened by the fact that in our study samples were unusually high, which should have amplified the impact of the role of memory. However, these results add to an increasing amount of evidence that challenges the importance of memory bounds (e.g. Rakow et al., 2008; Wulff et al., 2018, for a relevant discussion). To this end we welcome studies that seek to understand how decisions are informed by exploring mechanisms beyond plain memorisation.

Why so much sampling?

Subjects in both versions of our DFE treatments, were much more eager to explore options than what has commonly been reported. One explanation for this search 'explosion' relates to the absence of waiting time between two consecutive draws. In our experiment subjects were able to regulate the time the card remains on their screen. On the one hand, this feature increased clicking effort as subjects had to click twice -instead of only once which is more typical- before observing a new card: first to replace the previously drawn card and then to sample a new one. On the other hand, this adaptation made subjects' role during exploration more active and potentially faster (should subjects choose to make it so). It has been argued that in DFE, subjects are the 'masters of their information search' (Hills and Hertwig, 2010) and in this sense this study's framework takes this exploration-ownership one step further. Perhaps the more subjects relate to the role of an actor instead of that of an observer, the more encouraged they feel to explore further. A more prosaic explanation would be that the cost of clicking twice is a small price to pay for removing waiting time and therefore our intervention simply reduced the opportunity cost of sampling.

The DE gap across different elicitation methods

The differences between AHP's methodology (which this study adopts) with that of the sampling paradigm in inferring a DE gap, have raised concerns regarding the compatibility of the findings within these two approaches. Results 3 and 4 are reassuring in that respect. Result 3 shows that our method can detect a significant DE gap even without the mediation of a preference model, by focusing only on choice patterns. These choices are elicited from the first iteration of the bisection method which entails a choice between a risky and a safe option of equal EV; a setting very similar to that in early DE gap studies. Moreover, according to Result 4 this DE gap is qualitatively similar to that elicited in the sampling paradigm. Just as in Hertwig et al. (2004), subjects in our study chose the risky option more frequently in DFD than in DFE when rare events were associated with desirable outcomes while the opposite was true when the outcomes were undesirable. However, unlike in the early DE gap studies, subjects in our DFE treatments were overly hesitant in choosing 'Risky' in 'undesirable rare' decision problems. We discuss two possible explanations for this.

First, the fact that subjects knew about the existence of the (rare) undesirable outcome might have contributed to their hesitation of choosing 'Risky'. This is in accord with the 'mere presentation effect' discussed in Erev et al. (2008). Unlike the sampling paradigm where if this outcome was never sampled subjects might had never inferred its existence, AHP's method requires that subjects eventually found out about this outcome. Moreover, the fact that subjects in our study sampled a lot and were overall very well informed about the likelihood of the undesirable outcome might have amplified this effect. Indeed, when we look only in samples where this probability was underrepresented we see that subjects become more willing to take the risky option in such 'undesirable rare' decision problems. Second, we consider the discrepancy between the EV of lotteries under consideration. In earlier studies, subjects typically faced lotteries with an EV of approximately £3 (or less). In our study that EV was somewhere between £12 and £15.6 which could have made subjects more hesitant to reject the safe option. Given however that our analysis of the control lottery: $(4, E_{0.8}; 0)$ replicated this unusually high hesitation we believe that our first hypothesis is more likely to be the case.

The relative underweighting hypothesis

Our elicited weighting patterns provided little support for both the 'underweighting' and the 'ambiguity aversion' hypotheses. With respect to the first, our data in all treatments reveal -at the aggregate level- an inverse S-shaped weighting function which prescribes overweighting instead of underweighting of rare events. Moreover, unlike the second hypothesis, DFE-elicited weighting curves do not lie entirely beneath that elicited in DFD. Instead, our pattern seems to fit best under a third hypothesis that can be interpreted as a modest version of the underweighting one. The 'relative underweighting hypothesis' as summarized by Result 5 posits that although subjects overweight rare events in DFE, they do it less so than in DFD.

Regarding the discord with the 'underweighting hypothesis', Result 5 is not entirely surprising. Over the last few years, an increasing amount of studies have also failed to detect a S-shaped weighting curve, irrespective of the elicitation method they followed (e.g. AHP; Aydogan and Gao, 2016; Glöckner et al., 2016). One possible explanation for the absence of a S-shaped pattern in our DFE treatments is related to the high levels of sampling amount we recorded. Indeed, Result 6 seems to point in that direction as S-shaped weighting functions are prevalent among subjects who sample less. This is not surprising: subjects who do not sample enough are more likely to under-represent and thus underweight rare events. It is therefore plausible that if our levels of sampling had been significantly lower, we might have seen more evidence for the 'underweighting hypothesis'.

With respect to the disagreement with the 'ambiguity aversion hypothesis' we suggest the following explanation. The fact that subjects in our study collected larger samples than those in AHP might have affected their confidence during the evaluation of the lotteries. It is true that subjects in DFE can never be entirely certain regarding the underlying probability distribution. Nevertheless, richer information sets - such as the ones collected in our study - could have increased their confidence about those likelihoods and consequently reduced the associated ambiguity aversion.

Does the history table bridge the DE gap?

Lastly, we turn to a comparison of the DE gap between the two versions of experience that caused it. Result 3 suggests that although the gap is significant in both cases, its size is not symmetric. Specifically, our choice patterns reveal a bigger DE gap between DFD and DFE-NoHT. This asymmetry is corroborated by the weighting function comparison -w_{DFD} and w_{HT} are 'closer' than w_{DFD} and w_{NoHT}- as well as by the resistance of DFE-HT to 'conform' to all 4 properties of the original DE gap, even when we focus on under-represented probabilities.

To the extent that the analogical display of previously sampled events in DFE-HT has a similar 'descriptive' effect to the numerical summaries of uncertainty in DFD, this result should not come as a surprise. We interpret this 'bridging' of the gap as evidence that the DE gap should not be seen as a dichotomy but rather as a continuum over different levels of uncertainty.

3.5 CONCLUSION

We conduct a lab-experiment and examine how people search for information about uncertainty and how this influences their ensuing risky choices. We find that besides the properties of the risky options at hand, the environment in which these options are presented and evaluated is also important. With respect to search patterns in DFE, we show that a lottery's variance is negatively correlated with sampling amount which in this context means that people sample more from options with rarer events. We also find that sampling amount decreases over time periods. Both of these findings become less salient after the introduction of a history table which records and displays previously sampled outcomes at the time of the lottery evaluation. The cue that stands out in that case is the maximum capacity of that table. Moreover, our examination of the role of memory in sampling suggests that memory bounds were not very influential on search policies.

With respect to choices and preferences we compare responses between two variations of DFE: with (DFE-HT) and without a history table (DFE-NoHT) and compare them with those elicited from a standard version of DFD. Both of these comparisons generate a significant DE gap which is mitigated, however, by the inclusion of the history table. We interpret these choices through the CPT preference model by eliciting risk curvature (parametrically) and weighting functions (both parametrically and non-parametrically) at the individual level. Although utility curvature does not differ across treatments, the shape of decision weighting functions does. In our version of the DE gap in weighting, subjects in DFE overweight rare events but less so than in DFD. We show that the absence of under-weighting in DFE can partially be explained by the unusually high levels of sampling observed in our study.

Lastly, we report a measure that allows us to compare the type of gap found in studies using valuation methods -like this one- with the type of gap elicited in studies that use choice methods. We show that the phenomenon is qualitatively similar but not identical between the two methods.

BIBLIOGRAPHY

Abdellaoui, M., Baillon, A., Placido, L., and Wakker, P. P. (2011a). The Rich Domain of Uncertainty: Source Functions and Their Experimental Implementation. *American Economic Review*, 101(2):695–723.

Abdellaoui, M., Bleichrodt, H., and L'Haridon, O. (2008). A tractable method to measure utility and loss aversion under prospect theory. *Journal of Risk and Uncertainty*, 36(3):245.

Abdellaoui, M., L'Haridon, O., and Paraschiv, C. (2011b). Experienced vs. described uncertainty: Do we need two prospect theory specifications? *Management Science*, 57(10):1879–1895.

Aydogan, I. and Gao, Y. (2016). Are Black Swans Really Ignored? Reexamining Decisions from Experience. *Mimeo*.

Barberis, N. C. (2013). Thirty Years of Prospect Theory in Economics: A Review and Assessment. *Journal of Economic Perspectives*, 27(1):173–96.

Booij, A. S., Van Praag, B. M., and Van De Kuilen, G. (2010). A parametric analysis of prospect theory's functionals for the general population. *Theory and Decision*, 68(1-2):115–148.

de Palma, A., Abdellaoui, M., Attanasi, G., Ben-Akiva, M., Erev, I., Fehr-Duda, H., Fok, D., Fox, C. R., Hertwig, R., Picard, N., Wakker, P. P., Walker, J. L., and Weber, M. (2014). Beware of black swans: Taking stock of the description–experience gap in decision under uncertainty. *Marketing Letters*, 25(3):269–280.

Epper, T., Fehr-Duda, H., and Bruhin, A. (2011). Viewing the future through a warped lens: Why uncertainty generates hyperbolic discounting. *Journal of Risk and Uncertainty*, 43(3):169–203.

Erev, I., Glozman, I., and Hertwig, R. (2008). What impacts the impact of rare events. *Journal of Risk and Uncertainty*, 36(2):153–177.

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10(2):171–178.

Fox, C. R. and Hadar, L. (2006). "Decisions from experience" = sampling error + prospect theory: Reconsidering Hertwig, Barron, Weber & Erev (2004).

Judgment and Decision Making, 1(2):159–161.

Glöckner, A., Hilbig, B. E., Henninger, F., and Fiedler, S. (2016). The reversed description-experience gap: Disentangling sources of presentation format effects in risky choice. *Journal of Experimental Psychology: General*, 145(4):486–508.

Goldstein, W. M. and Einhorn, H. J. (1987). Expression theory and the preference reversal phenomena. *Psychological Review*, 94(2):236–254.

Gonzalez, R. and Wu, G. (1999). On the Shape of the Probability Weighting Function. *Cognitive Psychology*, 38(1):129–166.

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with ORSEE. *Journal of the Economic Science Association*, 1(1):114–125.

Hau, R., Pleskac, T. J., and Hertwig, R. (2010). Decisions from experience and statistical probabilities: Why they trigger different choices than a priori probabilities. *Journal of Behavioral Decision Making*, 23(1):48–68.

Hau, R., Pleskac, T. J., Kiefer, J., and Hertwig, R. (2008). The description–experience gap in risky choice: The role of sample size and experienced probabilities. *Journal of Behavioral Decision Making*, 21(5):493–518.

Hertwig, R. (2012). The psychology and rationality of decisions from experience. *Synthese*, 187(1):269–292.

Hertwig, R., Barron, G., Weber, E. U., and Erev, I. (2004). Decisions from Experience and the Effect of Rare Events in Risky Choice. *Psychological Science*, 15(8):534–539.

Hertwig, R. and Pleskac, T. J. (2010). Decisions from experience: Why small samples? *Cognition*, 115(2):225–237.

Hills, T. T. and Hertwig, R. (2010). Information Search in Decisions From Experience: Do Our Patterns of Sampling Foreshadow Our Decisions? *Psychological Science*, 21(12):1787–1792.

Lejarraga, T., Hertwig, R., and Gonzalez, C. (2012). How choice ecology influences search in decisions from experience. *Cognition*, 124(3):334–342.

Mehlhorn, K., Ben-Asher, N., Dutt, V., and Gonzalez, C. (2014). Observed Variability and Values Matter: Toward a Better Understanding of Information Search and Decisions from Experience. *Journal of Behavioral Decision Making*, 27(4):328–339.

Murad, Z., Sefton, M., and Starmer, C. (2016). How do risk attitudes affect measured confidence? *Journal of Risk and Uncertainty*, 52(1):21–46.

Quiggin, J. (1982). A theory of anticipated utility. *Journal of Economic Behavior & Organization*, 3(4):323–343.

Rakow, T., Demes, K. A., and Newell, B. R. (2008). Biased samples not mode of presentation: Re-examining the apparent underweighting of rare events in experience-based choice. *Organizational Behavior and Human Decision Processes*, 106(2):168–179.

Tversky, A. and Fox, C. R. (1995). Weighing risk and uncertainty. *Psychological Review*, 102(2):269–283.

Tversky, A. and Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323.

Ungemach, C., Chater, N., and Stewart, N. (2009). Are Probabilities Overweighted or Underweighted When Rare Outcomes Are Experienced (Rarely)? *Psychological Science*, 20(4):473–479.

Wulff, D. U., Mergenthaler-Canseco, M., and Hertwig, R. (2018). A meta-analytic review of two modes of learning and the description-experience gap. *Psychological Bulletin*, 144(2):140–176.

4

THE DESCRIPTION - EXPERIENCE GAP IN COOPERATION

*

> *"In critical moments men sometimes see exactly what they wish to see"*
>
> — Mr. Spock, *Star Trek - The Tholian Web'* (1968)

4.1 INTRODUCTION

Social dilemmas are situations where individually rational actions lead to undesirable social outcomes. The human tendency to neglect their immediate material interests and cooperate in anonymous, one-shot such situations -like the one-shot Prisoner's Dilemma (PD) game- although puzzling, is well-documented (Sally, 1995). One prominent explanation argues that people are conditionally cooperative (Fehr and Fischbacher, 2003; Fehr and Gächter, 2002; Fischbacher and Gächter, 2010). They are willing to cooperate if they believe others will do the same. In many circumstances however, the parties involved in a social interaction are uncertain about others' intentions. In these cases, conditional cooperation necessitates the formation of expectations about the likelihood of cooperation by others. Although the importance of such expectations in social dilemmas has been highlighted before (Dijkstra and van Assen, 2017), the question of how cooperation depends on the format in which uncertain information is obtained remains relatively unexplored.

As recent evidence from individual risky decisions suggests, the distinction can be important. According to a series of lab and online experiments, summarized hereon under the umbrella term: the 'Description – Experience (DE) gap', risky decisions might differ markedly as a function of how information about uncertainty was obtained (Barron and Erev, 2003; Hertwig et al., 2004; Weber et al., 2004). In these experiments, participants make a series of choices between risky options. In the Description condition, the properties of these options are communicated through explicit and numerical formats. Conversely, in the Experience condition, participants are asked to discover these properties by sampling from these options' distribution of outcomes.

* This chapter is based on joint work with Ozan Isler and Dennie van Dolder.

The most prevalent interpretation of the canonical finding in this literature posits that when people make decisions from description –like when consulting statistics on house fires prior to deciding whether to purchase the corresponding insurance— they tend to behave as if overweighting the impact of rare events. Conversely, when decisions are informed from past experiences –like contemplating to buy a high security bike lock based on experience with bike theft— people tend to behave as if underweighting the impact of rare events.

Sampling bias seems to be (at least) part of the explanation for this empirical disparity. Participants in Experience usually collect small samples that tend to underrepresent rare events. However, the gap has been found to persist even when sampling bias is eliminated (Hau et al., 2008). One of the leading explanations for the existence of a gap beyond sampling bias is attributed to attitudes towards ambiguity (Abdellaoui et al., 2011). When people make decisions from description they are fully aware of outcomes and outcome-probabilities. In decisions from experience however, this information is ambiguous. Ever since the famous urn-experiments by Ellsberg (1961), we know that the two domains can give rise to very different choices.

Despite voluminous research on the DE gap in individual risky decision making (see Wulff et al., 2018, for a recent review), its potential implications for social decision making remain largely unexplored. In this study, we ask whether the format in which social information is obtained -descriptive or experiential- influences cooperation in social dilemmas. According to the insights drawn from the canonical finding in the DE gap literature, we might expect that conditional cooperators would put more emphasis on rare events when these are described rather than experienced. Consequently, we hypothesize that conditional cooperators in Description will cooperate more when cooperation is rare and defect more when defection is rare than conditional cooperators in Experience. On the other hand, it is also possible that drivers of the DE gap in individual risky decisions, influence behaviour differentially in a social context. For example, Bolton and Ockenfels (2010) show that people's risk attitudes are affected by social comparisons .

To examine these hypotheses, we conduct an online experiment where participants play versions of the one-shot PD. In standard implementations of this game, people infer others' actions from the payoff matrix. In our study, participants have an additional source to consult. Specifically, they can learn something about the likelihood that their match cooperates. This modification allows us to systematically vary the likelihood of being matched with a cooperative agent. The information about this cooperation likelihood, is acquired by each participant through one of three conditions: Description or one of two Experience conditions. We introduce two variations of Experience to examine the properties of a potential DE gap in this social context beyond sampling bias. One in which participants can make as many observations as they wish and one in which the sample size of observations is fixed to match the relative frequency that was communicated to participants in Description.

This study's contribution is twofold. First, we explore if and how the format of acquiring information about the cooperation rates in one's environment affects ensuing cooperative behaviour. Information formats have been found to systematically influence behaviour in individual decision making. Given that the vast literature on social dilemma experiments does not make this distinction explicitly, it is important to know if information format affects behaviour in such social contexts. Second, we develop an experimental protocol that allows systematic manipulations of expectations regarding the likelihood of cooperation. Our protocol has the added methodological advantage of making the provision of false information unnecessary. Previous experimental designs achieve similar manipulations by telling people that they interact with other humans when in fact their interaction is with a computer algorithm (Santa et al., 2018). In our experiment actions and payoffs are interlinked among participants and our instructions do not contain untrue statements.

In what follows, Section 4.2 reviews the relevant background in greater detail, Section 4.3 describes the experimental design while Section 4.4 presents our results. Finally, Section 4.5 discusses these results and Section 4.6 concludes.

4.2 BACKGROUND

There is sizeable evidence that people often behave cooperatively even when the selfish action offers a higher material payoff. This finding holds even in one-shot interactions where reputation effects and other factors are meticulously muted. A relevant meta-analysis finds that approximately half (47%) of participants cooperate in such situations (Sally, 1995).

The principal driver of this cooperative behaviour has been argued to rest on people who cooperate as long as others are doing the same (Fehr and Fischbacher, 2003). We refer to these people as conditional cooperators. As Dijkstra and van Assen (2017) point out, under uncertainty, when the action of one's match is unknown, conditional cooperators will cooperate as long as they believe that there is a good enough chance of a mutual cooperative outcome. What constitutes a 'good enough chance' varies across people with some maintaining lower cooperation thresholds than others. On the other hand, there are also people who do not condition their behaviour on others' cooperative actions (or intentions thereof). We refer to people who cooperate unconditionally as unconditional cooperators while people who would always defect as free riders.[1]

In this study our goal is to examine whether the way people infer the likelihood of cooperation can influence their cooperative behaviour. To this end, we need a framework where such cooperative expectations can be exogenously manipulated. Popular designs in this literature typically do not much these criteria precisely. For example, in the repeated PD game with feedback, one can find out over time about the likelihood of her match cooperating (Andreoni and Miller, 1993). But this probability is endogenous as choosing to cooperate or defect in time t affects behaviour in $t + 1$.

One-shot games have the advantage of controlling for these strategic concerns. One way to manipulate expectations in a one-shot game is to allow participants to play the game repeatedly but employ a matching protocol that randomly assigns them to a new match. If participants receive anonymous feedback for the actions selected at the end of each round (Cooper et al., 1996) then in the long run, they could estimate the likelihood of cooperation in their environment and by extension in future encounters. The limitation of this approach, however, is that the experimenter does not have control over this likelihood. Moreover, it would be highly time consuming to try and extracting a cooperative profile across a variety of such likelihoods.

The manipulation used by Santa et al. (2018) is the closest to the one we are interested in. In their study, the authors develop an experimental design that manipulates beliefs about the likelihood of cooperation in a one-shot PD in order to elicit response times and test the intuitiveness of the cooperative action. Participant i is matched with participant j and together they play a simultaneous, one-shot PD. Prior to playing the game, i receives a summary statistic for the likelihood that j will cooperate. Although i is told that j is a fellow participant, in fact, j is always a computer algorithm that cooperates with a predetermined frequency.

We develop a framework that achieves the same exogenous and flexible manipulation of expectations regarding the likelihood of cooperation as in Santa et al. (2018) without the need to give false information to participants. Based on the one-shot PD, our framework manipulates i's expectations through varying the probability that her match j cooperates. However, in our experiment, both i and j are real participants that have participated in the same experiment. Using a variation of the strategy method (Selten, 1967), we can vary the cooperation likelihood exogenously for a variety of different probability scenarios.

1 Technically, there is a fourth category that includes people who will cooperate when others defect and defect when others cooperate. Most studies, however, tend to not warrant too much attention to this type and treat as an error of understanding from the participant's perspective.

Our design is also flexible. Since all tasks in our experiment are one-shot, they did not require giving immediate feedback to participants. This alleviates our experimental protocol from the complexity of real-time interactions.

This framework allows us to investigate our main research goal. Namely, to examine if and how the way in which people receive information about the likelihood of cooperation - from description or from experience - affects ensuing cooperative behaviour.

The literature on the DE gap in individual risky decisions is a good source to inform our hypotheses. The most common experimental setup in this literature is the sampling paradigm (Hertwig et al., 2004). There, participants are allocated either in Description or in Experience and are asked to make a series of choices between two gambles: a risky and a safe one. The risky option typically contains a rare event which can be either desirable or undesirable while the safe one usually consists of an outcome offered with certainty. The difference between the two treatments pertains to how the properties of these gambles are discovered. In Description, participants learn about these properties through numerical and explicit formats (e.g. there is a 10% chance that event A will occur). In Experience on the other hand, they discover this information by drawing independent observations sequentially from a source of uncertainty. This source has been tallied to generate events with the same relative frequency as the objective probability in Description. Whether participants observe this information accurately, however, rests upon their willingness to collect sufficiently large samples.

The canonical finding in this literature is that people in Description prefer the risky over the safe option when the rare event is desirable and the safe option over the risky one, when the rare event is undesirable. The opposite holds true when people make this decision in Experience. The most common interpretation of this pattern is that people in Description assign more weight to rare events than their objective probabilities warrant, whereas in Experience, these events are underweighted (Hertwig and Erev, 2009).

Exporting the intuitions from the DE gap in the individual context to a social dilemma, one may predict higher sensitivity to information about rare events when this information is obtained descriptively rather than experientially. In the context of the PD, this would imply that when cooperation (defection) is rare, people in Description will cooperate (defect) more than those in Experience. We refer to this as the 'canonical hypothesis' as it is grounded in the expectation that the canonical finding -as well as its most common interpretation- from the DE gap literature in the individual context, also applies to the social context.

Notwithstanding, the 'canonical hypothesis' is not the only candidate. Glöckner et al. (2016) report that sufficiently complex choice environments can lead to reversals of the canonical gap, so that rare events can appear to be overweighted in Experience more than in Description. Examples of such complex environments include cases where all available gambles are non-reduced (i.e. containing each at least two possible outcomes). According to the authors, noisy probability evaluations induce a 'regression towards the mean' effect to probability estimates and by extension to the weights people assign to those probabilities. Similar to Glöckner et al., participants in our Experience treatments are likely to believe that there were always facing two possibilities: their match could have cooperated or defected. Therefore, high enough probabilistic distortions can lead to more overweighting of rare events in Experience compared to Description.

As Glöckner et al. report, eliminating sampling bias is essential for this 'reversal' of the canonical gap. Sampling bias refers to the fact that participants in Experience often collect small samples that tend to under-represent rare events. It comes therefore as little surprise that under-represented rare events are underweighted. Indeed, as we show in Chapter 3 and as a recent meta-analysis demonstrates (Wulff et al., 2018), sampling bias is the most significant contributor to the gap in individual choices. Other studies that have also controlled for sampling bias, however, find somewhat different patterns. For example Abdellaoui et al. (2011), find an overall reduction in the willingness to gamble in Experience which they attribute to ambiguity aversion. This pattern distinguishes between rare events according to their desirability. Desirable rare events are overweighted more in Description but undesirable rare events are overweighted more in Experience.

To explore this aspect of the DE gap, we employ two variations of Experience, one in which participants can sample freely (E-Free) and one in which sampling has been fixed so that it always matches the objective probability (E-Fixed). To the extent that sampling bias is as pivotal to the gap as indicated by studies in the individual context, we might expect the canonical hypothesis to hold in the comparison between Description with E-Free but not between Description and E-Fixed.

Despite the voluminous research on the DE gap in individual risky decisions, to the best of our knowledge there have been only two studies that examine the gap in a social context. Fleischhut et al. (2014) investigate the DE gap in the context of an Ultimatum Game. Proposers, prior to making their offers, are informed —either through descriptions or by sampling— about the relative frequency with which each allocation had been accepted or rejected in previous experiments. The authors report that people tend to prefer the riskier option less often in Experience than in Description (independently of whether that rare event was desirable or undesirable) but treatment differences are almost never significant. Artinger et al. (2012) use a public good game for their social context with a stochastic payoff from cooperation. The payoff is low with probability p and high with $1 - p$. Similarly to Fleishhut et al., they find little evidence for significant differences between Description and Experience.

Taken together, the evidence in favour of a significant gap in the social context is weak. In this study we revisit the question by taking steps towards making the social component more salient. To this end, unlike Fleishhut et al., the actions taken from participants in our experiment influence the payoff of fellow participants. Becoming aware of payoff interdependence has been shown to stimulate the social component of behaviour (Martin et al., 2014). Moreover, unlike Artinger et al., we manipulate the likelihood of cooperation rather than the payoff structure of the game. We suspect that placing the locus of uncertainty on the action of a fellow decision maker highlights the social character of the decision.

Lastly, as a robustness check for the interpretation of our findings, we would like to distinguish between people according to their cooperative preferences. One would expect for example, that if a DE gap does exist in this social dilemma, then it ought to be driven by conditional cooperators rather than unconditional types (free riders or unconditional cooperators), as they are the ones who by definition care for social information. For the same reason, we expect that conditional cooperators are more eager to search longer in E-Free. Eliciting these types in the two Experience treatments is problematic as beliefs regarding the underlying probability of cooperation may not coincide with observed relative frequencies. This can be true even if people have collected representative samples.[2] We therefore include a separate task that allows us to elicit such preferences at the end of the experiment. This task is inspired from Fischbacher et al. (2001) who use a variant of the strategy method to elicit a profile of cooperative responses to a variety of scenarios that differ with respect to others' contributions. They then use these profiles to elicit cooperative types (such as conditional cooperators or free riders). We adopt their design for the PD game.

4.3 METHODS

This is a between-subjects experiment. Participants are randomly assigned to one of three treatments: decisions from description (Description), decisions from experience with free sampling (E-Free) and decisions from experience with fixed sampling (E-Fixed). Our experimental design comprises of three

2 This is less of an issue in Description where it is generally accepted that described information coincides with participants' beliefs.

stages.[3] All stages comprise of PD games with the same underlying payoff structure as the one in Table 4.1. In our framing of the game, 'Keep' is the non-cooperative action while 'Share' is the cooperative one. As all games are one-shot and there was no feedback between stages, the data collection process does not rely on real-time interactions.

Table 4.1: Payoff matrix for the Prisoner's Dilemma game

If you choose	& If your match chooses:	Then you get:	& Your match gets:
Keep	Keep	50 p	50 p
Share	Share	100 p	100 p
Keep	Share	150 p	0 p
Share	Keep	0 p	150 p

Note. This payoff structure was common for the tasks in all 3 stages. This table appeared in every stage as a reminder.

In Stage 1, all participants are told that they are matched with a fellow participant and are asked to play the simultaneous one-shot PD game shown in Table 4.1.

Stage 2 consists of the main task for this study. This is a modified PD game that systematically and exogenously manipulates the likelihood of cooperation in a social environment through a variant of the strategy method. Participants play the one-shot PD game in Table 4.1 in seven different scenarios. These scenarios differ with respect to the likelihood of cooperation.

Specifically, in every scenario, participant i is assigned to a sub-population of participants. She is then informed that she has been matched with another participant, j, from this sub-population. A crucial aspect of our manipulation is the ability to provide i with information about the likelihood that j will cooperate. To this end, we inform i that she will play a one-shot PD with j's action from Stage 1.[4] Although i does not know j's Stage 1-action, she can form an expectation about it by accessing information about this scenario's sub-population probability of cooperation (SPoC , measured in %). SPoC summarizes the proportion of Stage 1 cooperative actions of all participants except i in a given sub-population. Therefore, it represents the objective probability

3 Participants know this from the start but instructions for each stage are provided only at the beginning of each stage.

4 From a temporal perspective, this makes the PD game in Stage 2 a sequential PD game with imperfect information about the other player's action. Note that this situation is conceptually equivalent to a simultaneous PD game with partial information about the other player's intention. There has been some evidence that despite this logical equivalence, people are less prone to cooperate if they can infer that their match has already decided on her action, even when they have no way of knowing what this action was (Shafir and Tversky, 1992). Since this sequential feature is constant across treatments, any treatment effect should remain unaffected by it.

that i is facing a cooperative agent j in that scenario. There are seven levels of SPoC: $\{0, 10, 30, 50, 70, 90, 100\}$ and their order is randomized for each participant. Similarly to the strategy method, participants are informed that only one of these seven scenarios is real and that only the real scenario can be relevant for their final payment.[5] Importantly however, there is no way of making sure which scenario that is.

The way i accesses information about SPoC is treatment dependent. In Description she learns about it through numerical descriptions. When SPoC=50 for example, she will read on her screen that '50% of your group chose to Keep and 50% of your group chose to Share'.[6] In E-Free, they obtained this information through a sampling process. Every time they pressed a button on the screen they observed the Stage 1 action taken by a randomly chosen member of their sub-population. This sampling process was with replacement. Therefore, given enough draws, the observed relative frequency of cooperation converges to the objective SPoC. In E-Fixed, participants also sampled Stage 1 actions but unlike in E-Free, their sampling was tallied so that the sample distribution of cooperative actions coincides with the objective SPoC. To achieve this, participants in E-Fixed were required to sample exactly ten times in each scenario.

Finally, in Stage 3 we ask i to state her responses in a sequential PD game (with the same payoff structure as in Table 4.1) where her match has: a) Cooperated and b) Defected. Similarly to Stage 2, the actions of i's match are drawn from Stage 1. This is a variation -adapted for a PD game- of the task that Fischbacher et al. (2001) use to elicit cooperative types in the context of a public goods game. In the absence of uncertainty regarding the other person's intentions, preferences are separated from beliefs. We therefore categorize participants into four types: 'conditional cooperators' who match the other player's action, 'free riders' who always defect, 'unconditional cooperators' who always cooperate, and 'others'.

For our online experiment, we recruited 1094 participants from Prolific (Peer et al., 2017) and collected data using Qualtrics. On average, sessions lasted approximately 20 minutes. Average payment was about £2, which includes a £1.25 participation fee and a variable payment that was determined by their payoff in a randomly selected task. For more details on how payments were executed according to the matching protocol see Appendix C.2. Although allocation between the three information format conditions was random, we chose to collect approximately 1.5 more participants in E-Free (n = 473) than in Description (308) or E-Fixed (317). This was due to a need for

5 The real sub-population according to which payments were carried out comprised of three participants. Therefore, the only potentially real scenarios are SPoC=$\{0, 50, 100\}$.

6 Although the term 'sub-population' is technically more accurate, the term 'group' was preferred as it was deemed easier to internalize by participants.

higher statistical power in E-Free, which was associated with more hypotheses (e.g. regarding sampling behaviour). Approximately 9.8% of participants that were recruited were excluded either due to incomplete submissions or due to failure to answer control questions in the beginning of the experiment. These questions were designed to assess their understanding of the instructions. Participants who repeatedly failed to answer these questions correctly, were given a participation fee but were not allowed to proceed with the experiment. The number of excluded participants did not vary across treatments ($P = 0.872, \chi^2 - test$). In our final analysis, we use data from 990 participants (279; 435; 276, for Description, E-Free and E-Fixed respectively). Instructions of the experiment can be found in Appendix C.1.

4.4 RESULTS

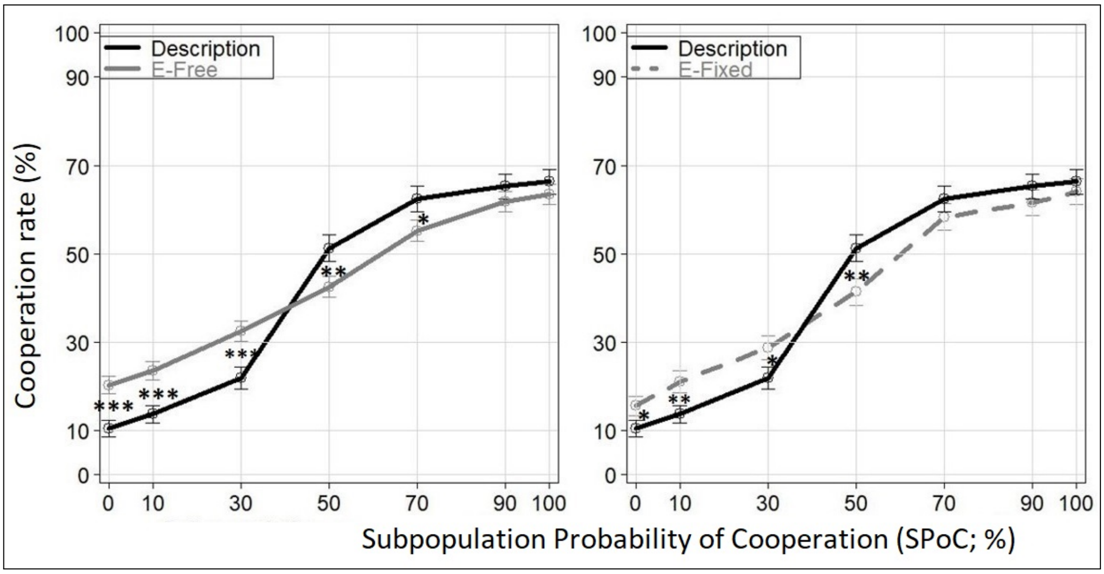
In Stage 1, 57.9% of our overall sample cooperated (chose to "Share") and this measure was not significantly different across treatments ($P = 0.141, \chi^2 test$). This is a high cooperation rate which nonetheless falls well within the range of commonly reported values (Dawes and Thaler, 1988; Sally, 1995).

Next, we turn to the main task in our study: the modified PD game in Stage 2. Figure 4.1 summarizes our main findings. Since individual responses were binary (Share=1, Keep=0), each point in Figure 4.1 represents the percentage of people who chose to cooperate in the corresponding treatment and scenario of SPoC. We refer to the line that connects these cooperation rates for a given condition as the 'response function'. Each panel juxtaposes the Description treatment with one of the Experience treatments. These comparisons point towards a significant gap between Description and E-Free (left panel) and between Description and E-Fixed (right panel).

Overall, cooperation rate increases with SPoC, suggesting that many people care about the cooperativeness of the environment and are in fact more likely to cooperate when the probability of being matched to another cooperator is higher. This relation is indicative of conditional cooperation. Nonetheless, we can also infer tendencies for unconditional behaviour as there is a sizeable portion of people who cooperate even when SPoC=0 and defect even when SPoC=100.

Average cooperation differs significantly between Description and E-Free for SPoC$\leq$ 70 and between Description and E-Fixed for SPoC$\leq$ 50. Overall, we see that both Experience curves lie above Description for low SPoC values (when cooperation is rare) and below Description for medium to high SPoC values (when defection is rare). Interestingly, and to some perhaps surpris-

Figure 4.1: Response functions



Note. Pearson's χ^2 tests. '***': $P < 0.001$; '**':$P < 0.05$;'*':$P < 0.1$. Error bars represent standard errors.

ingly, this pattern is the opposite of what the 'canonical hypothesis' would have predicted.

Another interesting feature of this gap is that it does not appear to be affected by sampling bias, the leading driver of the gap in individual risky decisions. If sampling bias were an important moderator in this social context, we would expect the pattern between Description and E-Free (where sampling bias is present) to be markedly different than that between Description and E-Fixed (where sampling bias is ruled out by design). In fact, average cooperation rates do not differ between E-Free and E-Fixed in any scenario. This lack of difference occured despite significant sampling bias in E-Free. Participants in E-Free sampled relatively little, the median being 4 cards per round. As a result, in 63% of all cases where a sample was obtained, the relative observed frequency misrepresented SPoC by 10 percentage points or more. Hence, our first two results can be summarized as follows

Result 1. *There is a significant DE gap in cooperation. When the likelihood of cooperation (defection) is low, people in Experience cooperate (defect) more than those in Description.*

Result 2. *Sampling bias does not drive the DE gap in cooperation.*

To better understand these results, we introduce two indexes: cooperativeness and conditionality. The cooperativeness index is constructed by calculating the average cooperation rate across all levels of SPoC for each treatment. Intuitively, this score represents the overall cooperative tendency in a treatment. Conditionality is constructed by calculating the difference of cooperation rate at SPoC= 100 and at SPoC= 0. It captures, thus, the overall elevation of the response curve. The closest to 1 this score is for a given treatment, the higher the propensity to conditionally cooperate.

Table 4.2: Cooperativeness and conditionality indexes across treatments

	COOPERATIVENESS	CONDITIONALITY
DESCRIPTION	0.416	0.558
	(0.016)	(0.034)
E-FREE	0.427	0.432
	(0.015)	(0.026)
E-FIXED	0.415	0.486
	(0.019)	(0.033)
P-VALUE	0.909	<0.01***

Note. Standard errors in parentheses. P-values derive from Kruskal-Wallis tests across treatments.

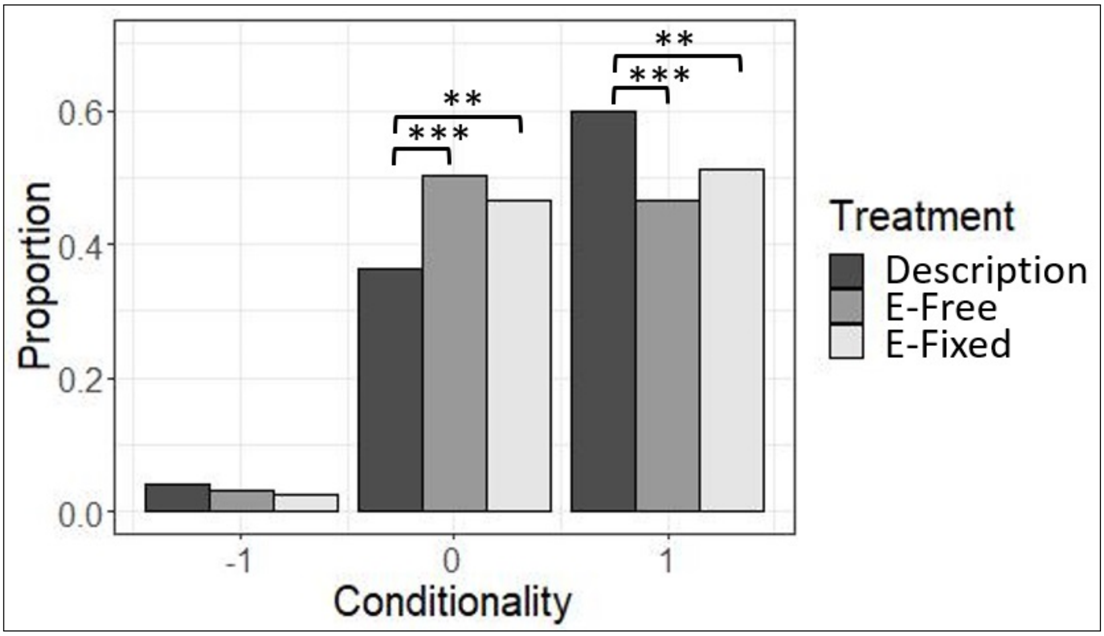
Table 4.2 reports values for these two indexes across treatments. Cooperativeness does not differ across the three treatments ($P = 0.909$, Kruskal-Wallis test) but conditionality does ($P = 0.005$, Kruskal-Wallis test). The effect on conditionality originates from differences between Description and E-Free and between Description and E-Fixed and not between the two experience treatments (P=0.001, P=0.071 and P=0.218, Wilcoxon rank sum tests for each binary comparison).

We examine this asymmetry in conditionality further by looking at behaviour at the individual level. Specifically, we distinguish between people who scored '1' (cooperated at SPoC= 100 but not at SPoC= 0), '0' (did not change their behaviour from one scenario to the other) and '-1' (people who cooperated at SPoC= 0 and defected at SPoC= 100). Figure 4.2 plots the distribution of these scores across treatments.

As we see in Figure 4.2, the percentage of people who do not update their action between SPoC= 0 and SPoC= 100 is significantly higher in the Experience treatments. In contrast, there are significantly more people in Description who switch from defection at SPoC= 0 to cooperation at SPoC= 100.[7] There is no statistical difference between the two Experience treatments. A small percentage of people (less than 4% overall) chose to cooperate when SPoC=0 but defect when SPoC=100. We are agnostic about the interpretation of this behaviour: it can reflect a type of 'reverse conditionality' or simply be due to

7 It is worth noting however that there is a sizeable proportion of people (approximately a third of the total sample) who would switch their action more than once if we were to consider intermediate SPoC scenarios. Although we could have imposed a single switching protocol, we decided that it is best to let participants' choices unmediated. Appendix C.2.2 delves into this issue further. Nonetheless, the conditionality index is unaffected by these inconsistencies as it ignores such intermediate switches.

Figure 4.2: Conditionality at the individual level



Note. Pearson's χ^2 tests. '***': $P < 0.001$; '**':$P < 0.05$;'*':$P < 0.1$.

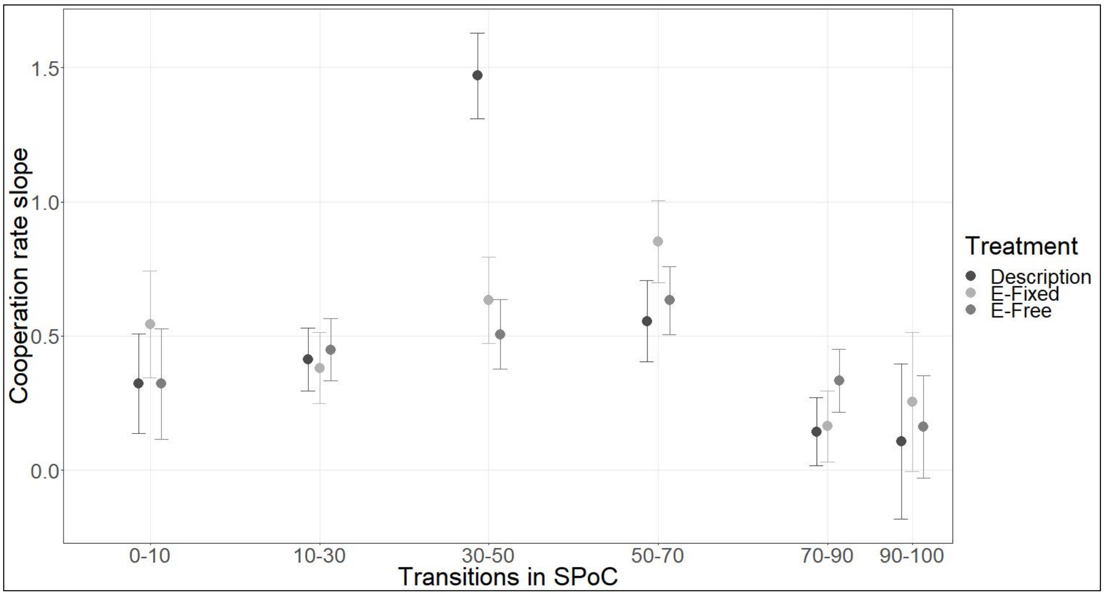
people who misunderstood the task. In either case, this behaviour is very rare and its existence does not significantly differ across treatments.

Result 3. *People in Experience are less prone to condition their behaviour on social information compared to those in Description.*

The insensitivity to social information in Experience is also evidenced by difference in cooperation at SPoC=50 in Figure 4.1, where we observe a steep 'jump' in cooperation responses in Description but not in Experience. One interpretation of this asymmetry is that a sizeable fraction of participants want to follow a simple heuristic to cooperate if the chance that their match will do the same is at least 50%, but participants in the Experience treatment cannot discern this probability with sufficient certainty. To examine this pattern more closely, we plot in Figure 4.3 the slope of the response function at each transition from SPoC(x) to SPoC(x+1). Intuitively, the steeper the slope, the sharper the reaction to a given level of SPoC.

Figure 4.3 verifies that participants in Description display a distinctly sharper reaction when they encounter the signal SPoC$=$ 50. No other signal in Description, and no signal in general in Experience, is nearly as impactful. Most people who chose to switch from cooperation to defection in Description, did so when receiving the signal that there was at least a 50% chance of a cooperative outcome. On the other hand, when this signal was ambiguous, people were significantly less sensitive to it. Statistical analysis corroborates this. The difference in slopes is significant across treatments only at SPoC$=$ 50 (P$<$ 0.01, Kruskal-Wallis test).

Figure 4.3: Response function slopes



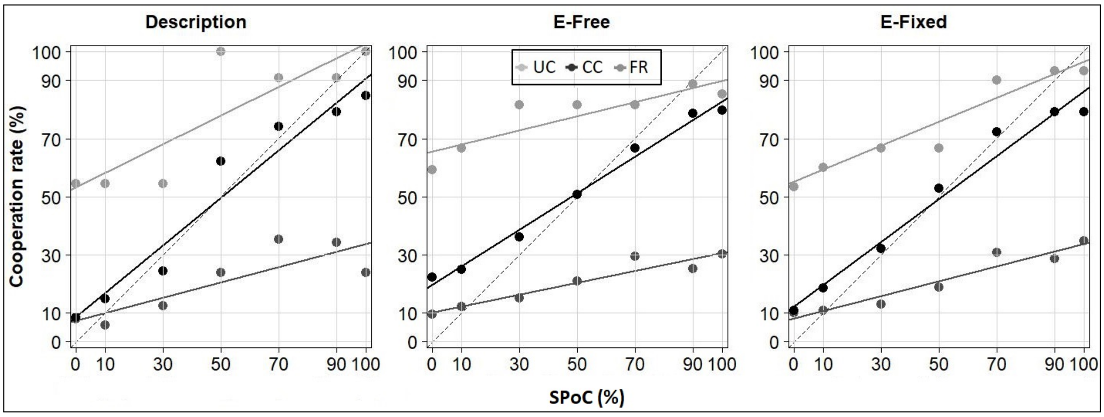
Note. '0-10': Slope of the cooperation function between SPoC=0 and SPoC=10. Whiskers represent standard errors that derive from the calculation of this slope for each individual.

Lastly, we look at differences in behaviour between different types of players, as revealed from choices in Stage 3. Overall, the majority of participants was categorized as conditional cooperators (Description: 63.4%; E-Free: 59.3%; E-Fixed: 50.7%). The second most frequent category was that of 'free riders' (Description: 31.5%; E-Free: 32.0%; E-Fixed: 36.6%) while 'unconditional cooperators' were in the minority (3.9%; 6.2%; 10.9%).[8]

Although there is no difference in the distribution of 'free riders' ($P = 0.352$; χ^2-test) and 'Others' ($P = 0.380$; χ^2-test) across treatments, we do observe a significant asymmetry in the distribution of 'conditional cooperators' ($P < 0.01$; χ^2-test) and 'unconditional cooperators ($P < 0.01$; Chi^2-test). This asymmetry stems from E-Fixed, where there is a significant increase of 'unconditional cooperators' compared to Description and E-Free ($P < 0.01$ and $P = 0.036$ respectively; χ^2-tests) and a significant drop in 'conditional cooperators' compared to Description ($P < 0.01$ and $P = 0.030$, respectively; χ^2-tests). There is no statistical difference in the distribution of types between Description and E-Free ($P = 0.973$, $P = 0.305$, $P = 0.252$ and $P = 0.275$ for 'free riders', 'conditional cooperators', 'unconditional cooperators' and 'others', respectively; χ^2- tests). We return to this asymmetry in E-Fixed in the Discussion section.

8 There was also an almost negligible percentage of people that do not fit in any of these three categories (Description: 1.0%; E-Free: 2.6%; E-Fixed 1.8%). These are people who would cooperate when their match defects and who would defect when their match cooperates (similar to the people who score '-1' in our conditionality index). We refer to those people as 'Others'.

Figure 4.4: Aggregate cooperation rates conditioned on types



Note. 'UC': Unconditional cooperators; 'CC': Conditional Cooperators; 'FR': Free riders. Linear models between SPoC and cooperation rate have been fitted for each type in each treatment as a visual aid.

Table 4.3: Cooperation indexes conditioned on types

	COOPERATIVENESS			CONDITIONALITY		
	CC	FR	UC	CC	FR	UC
DESCRIPTION	49.6	20.4	77.9	76.1	16.0	45.4
	(1.7)	(2.2)	(8.2)	(3.8)	(5.4)	(16.5)
E-FREE	51.2	20.3	77.8	57.8	20.8	26.0
	(1.8)	(2.1)	(4.9)	(3.8)	(4.5)	(8.8)
E-FIXED	49.2	20.9	74.8	68.5	24.8	40.0
	(2.2)	(2.7)	(3.8)	(4.3)	(4.6)	(10.5)
P-VALUE	0.768	0.877	0.788	P<0.01***	0.528	0.402

Note. Standard errors in parentheses. P-values derive from Kruskal-Wallis tests across treatments.

Figure 4.4 depicts Stage 2 cooperative behaviour, conditioned on Stage 3 typology while Table 4.3 reports cooperation indexes for each type across treatments. Aggregate behaviour in Stage 2 is overall consistent with type categorization.[9] Those who are categorized as conditional cooperators in Stage 3 are indeed more prone to condition their behaviour on social information in Stage 2. Likewise, those categorized as unconditional cooperators score the highest in cooperativeness and score low on conditionality, while those who are categorized as free riders have the lowest cooperative score and are also scoring low on conditionality.

Statistical comparisons in Table 4.3 provide some robustness tests for earlier intuitions. Specifically, we find that all treatment differences are driven from conditional cooperators who exhibit different degrees of conditionality (and

9 In fact, as we show in Appendix C.2.2, these types are highly predictive of Stage 2 behaviour at the individual level as well.

not cooperativeness) across treatments. There are no significant treatment differences for free riders or conditional cooperators. This is in line with the expectation that the DE gap in cooperation can only be driven by people who care about social information. Moreover, in support of Result 3, conditional cooperators are significantly more sensitive to SPoC in Description compared to E-Free and E-Fixed ($P < 0.01, P = 0.090$; Wilcoxon rank sum tests).

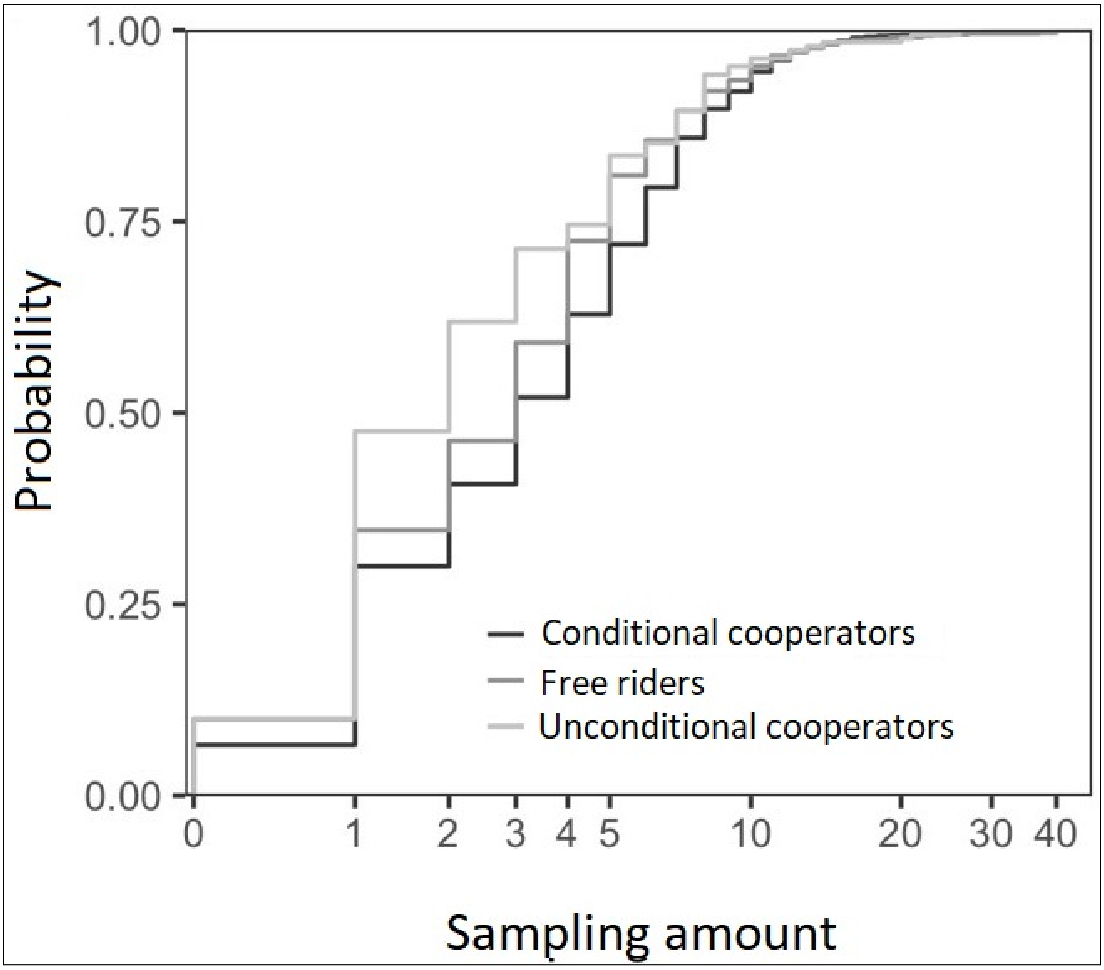
One noticeable difference with our previous analysis, however, is that now we also observe a significant difference between E-Free and E-Fixed ($P = 0.0395$; Wilcoxon rank sum test). Taking this result at face value, it suggests that people in E-Free were less likely to condition their behaviour on the true underlying SPoC. This could be the case either because of sampling bias in E-Free, or because the low amounts of sampling made the information in E-Free more ambiguous than the information received in E-Fixed. This difference, however, should be taken with a grain of salt due to the fact that the type distribution in E-Fixed is significantly different from that in the other two treatments. This suggest that Stage 2 affected the Stage 3 elicitation. We will return to this issue in the discussion.

In the E-Free treatment, participants could decide how much information they would sample. Although such sampling did not entail any monetary cost, it does involve exerting more effort and spending more time on the task. Theoretically, we would expect that different types have a different willingness to spend this time and effort in order to get the social information. In particular, conditional cooperators should be more interested in others' behaviour than unconditional cooperators or free-riders. As a result, we would expect them to collect bigger samples.

Figure 4.5 plots the cumulative distribution of sampling amount across different types. Although the overall sampling amount is low, we do observe that conditional cooperators sample more than the other types. Sampling amount for conditional cooperators was indeed the highest, with an average of 4.1 draws per round vs. 3.7 for free riders and 3.1 for unconditional cooperators. These differences were statistically significant ($P = 0.032$ for the first comparison and $P = 0.019$ for the second; Wilcoxon test for clustered data).[10]

10 See Rosner et al. (2006) for more details on this test.

Figure 4.5: Cumulative distributions of sampling amount across types



4.5 DISCUSSION

4.5.1 *The DE gap in individual and social context*

Although we find a significant DE gap in cooperation, this gap displays the opposite pattern relative to the canonical finding in individual risky decisions (Result 1). At face value, this finding suggests that people process information differently when facing social decisions, than when they make individual choices. However, recently Glöckner et al. (2016) have observed reversals similar to ours in the individual domain. In the next two subsections we discuss two possible explanations for this reversal. The first provides a link with the literature of the DE gap in individual decisions while the latter introduces an explanation unique to our experiment.

4.5.1.1 *Noisy evaluation*

Glöckner et al. suggested that environments that entail a choice between two non-reduced gambles, are likely to induce noise in the probabilistic estimates of participants in Experience. This noise will regress probabilities and by extension decision weights to the mean, accounting thus for increased overweighting in Experience relatively to Description. As we explained in Section 4.2, the similarities of our design with that of Glöckner et al., render the 'noisy evaluation' hypothesis a viable explanation for the reversal of the DE gap we observed.[11] Indeed, this could explain why the social cue of SPoC=50 that was revealed so important to participants in Description, did not register any considerable response with participants in Experience. If noise makes every probability seem like 50%, then it should come as little surprise if the actual 50% signal becomes less discernible. However, the fact that in our experiment this reversal occurs even in the absence of sampling bias (Result 2) suggests that there may be more to the story.

4.5.1.2 *Strong priors*

Previous studies in the DE gap in individual risky decisions, provide participants with gambles framed in an abstract format where uncertainty is ultimately arbitrated by a randomness-generating machine. In our setting on the other hand, uncertainty relates to cooperation which is arbitrated by the actions and intentions of fellow humans. This is a much less abstract and emotionally more loaded framework. People will likely hold priors regarding the likelihood of cooperation of a potential partner.

We argue that priors regarding the likelihood of cooperation in a social dilemma, are stronger than any priors regarding the likelihood of winning in an abstract gamble. As a result, players will update their prior less in light of new information in the former than in the latter frame. For example, a person with high prior expectations regarding the likelihood of cooperation, might be more reluctant to update her belief down to 0% when in the scenario of SPoC=0. This is especially true when this information is ambiguous (such as in Experience) since uncertain information is likely to be perceived as less trustworthy and therefore assigned less weight. Assuming that people's priors are heterogeneous and span the entire domain, response curves in Experience are likely to be more distorted near the edges (where rare events are) than in Description producing thus the reversed DE gap we observed. In contrast,

11 It is true that unlike in Glöckner et al., scenarios with SPoC=0 and SPoC=100 were technically 'reduced gambles' in the sense that the match could have behaved only in one way. Nonetheless, participants in E-Free and E-Fixed could not have been certain of it, due to the inherent ambiguity which is associated with information in Experience.

such reversals are less likely to occur in versions of Experience where the underlying uncertainty is framed in abstract terms. In these cases, people will be less resistant to update their beliefs and the role of priors will be less influential.

4.5.1.3 *The verdict*

It is worth noticing that the two explanations ('noisy evaluation' and 'strong priors') can very well act synergistically towards the reversed DE gap in cooperation. In fact, together they can explain why the effect of ambiguity in this social context is different than that described by Abdellaoui et al. (2011) and by the findings in Chapter 3. Both of these studies (Abdellaoui et al., and Chapter 3) were conducted in an abstract frame where priors are malleable. Moreover, the choice environment in those two studies was simpler as it always included one reduced gamble.

However, although the 'noisy evaluation' explanation has been successfully tested before (Wulff et al., 2018), the 'strong priors' one remains currently tentative. Fleischhut et al. (2014) provide some evidence in favour of this potential asymmetry between the rigidity of priors in the individual and in the social context. They find that participants tend to search information significantly less in the social compared to the individual setting and attribute this to the strength of priors: participants in social contexts have strong beliefs regarding others' behaviour and do not need to sample from their actions too much to inform their decision. Some evidence of this can be found in the low amounts of sampling – relative to past evidence — we observe overall in this study, but we cannot be sure as we have no control treatment for this.

4.5.2 *An effect on beliefs, not preferences*

Notice how both explanations ('noisy evaluation' and 'strong priors') for the reversal of the canonical DE gap, point to effects taking place in the domain of beliefs rather than of preferences. More precisely, Glöckner et al. do not argue that the treatment manipulation (described or experienced information) transforms a risk averse individual to a risk seeking one (or vice versa). Rather, the change in behaviour is due to participants' inability in Experience to discern signals (probabilities) as accurately as in Description. This inability regresses their weighting functions closer to the mean, inducing a behaviour consistent with more overweighting of rare events.

Similarly, our finding that people in Experience are less likely to condition their behaviour on social information (Result 3), should not be interpreted as a claim that Experience turns conditional cooperators into unconditional types (free riders or uncondtional cooperators). Instead, we argue that the gap in social cooperation is due to the coexistence of ambiguous information with strong prior beliefs. Conditional cooperators who erroneously infer that there is a chance for cooperation (when there is none), appear as unconditional cooperators. Symmetrically, conditional cooperators that see the chance of defection (when there is none) are likely to appear as free riders.

4.5.3 *Limitations and steps forward*

One potential limitation of our design can be the fact that the elicitation of cooperation types always succeeded Stage 2 instead of having its order of occurrence randomised. The reason for this was to avoid biasing behaviour in Stage 2 which was this experiment's main task. Although this decision was deliberate, it does leave the possibility open that the typology is affected from potential treatment spill-over effects.

We find it unlikely that such spill-over effects have affected Description and E-Free as despite significant differences between the two treatments in Stage 2, there was no asymmetry in the distribution of Stage 3 types. This is not the case with E-Fixed, however, where we observed significantly more people being revealed as free riders and significantly less people being revealed as conditional cooperators, relatively to E-Free and Description. This asymmetry is likely behind the fact that when we try to replicate the DE gap in Result 1 with people who have been revealed as conditional cooperators by our Stage 3 task, we find a weaker version of the gap between Description and E-Fixed than between Description and E-Free. Since Result 1 can only be driven by conditional cooperators (by definition unconditional types should not care about social information irrespective if this is described or experienced) this should not have been the case.

One explanation may be that some participants in E-Fixed were feeling somewhat anxious to finish the experiment. Their impatience might be justified in that they had to spend a larger amount of time (relatively to E-Free or Description) in Stage 2, as they were forced to sample 10 times in each SPoC scenario. Therefore, these participants might have rushed through the task by clicking away their preferred action twice without reading the instructions carefully. This would inflate the rates of revealed unconditional types in this treatment as we observed.

With this foresight in mind, an extension of this study could entail a version of E-Fixed where the forced sampling amount is reduced to a smaller threshold, so that time spent in the Stage 2 is comparable to E-Free, and test whether the asymmetry persists.

Other extensions could entail a set of treatments similar to Description, E-Free and E-Fixed but this time applied to individual risky decisions. This would give us a more direct test of the underlying differences between the two domains (social and individual). One particularly interesting aspect of this investigation is related to the 'strong priors' explanation we put forward in this study. Specifically, it would be interesting to see a study with two different Description treatment. One in which risk is expressed in its typical abstract frame and one in which it is framed in terms that are likely to evoke strong predispositions (for example 'risks of drinking and driving' or 'risks of travelling by plane vs. risk of travelling by car'). According to the 'strong priors' hypothesis, the 'loaded' version of Description will exhibit more overweighting of rare events than the abstract one.

4.6 CONCLUSION

People often choose to cooperate even when the selfish action is the material payoff dominating one. Conditional cooperation – the willingness to cooperate if others do the same – is the leading explanation for this phenomenon. Facing uncertainty, conditional cooperators act upon their expectations of others' intentions. To inform these expectations, people rely sometimes on descriptive statistics and other times on personal experiences.

Literature from individual risky decisions suggests that the distinction between descriptive and experiential formats is important. The canonical interpretation of this 'Description - Experience gap', is that people in Description behave as if overweighting rare events whereas, in Experience, they behave as if underweighting these events. The most decisive driver of this disparity owns to sampling bias: a tendency of people in Experience to collect small samples that under-represent rare events. We know very little, however, whether a DE gap persists in the context of social uncertainty.

The goal of our study was to examine if the way cooperation expectations are informed, through descriptions or through experiential sampling, influences ensuing cooperative behaviour. In a large online experiment we manipulated the likelihood of cooperation in a Prisoner's Dilemma game by providing participants with aggregate information on cooperation. In Description this information is provided explicitly in numerical presentation while in Experience (E-Free), participants find out autonomously through a sequential sampling process. We then observe how people choose to cooperate as a function of differing cooperation likelihoods.

In accord with past evidence in favour of conditional cooperation, we verify that a majority of participants cooperate more if their partner is expected to do so as well. Interestingly, we find evidence for a gap in social decision making, but in the opposite direction than what the canoncial finding in the individual context would have predicted. Rare events (of cooperation or defection) appear to be more overweighted in Experience rather than in Description. In a variation of Experience (E-Fixed) where sampling is fixed so that participants are forced to experience mathematically equivalent information with that in Description, we verified that sampling bias does not affect this 'reversal'.

Moreover, using a separate task to elicit cooperative types and provide some robustness tests related to these findings findings. We verify that cooperative behaviour maps intuitively into these types. Moreover, we confirm that the gap in cooperation is driven by conditional cooperators. These are by definition the people who would care about social information. Similarly, we show how conditional cooperators tend to collect significantly more social information than free riders or unconditional cooperators.

To interpret our key findings, we put forward a simple way to derive indexes that analyze behaviour in two dimensions. Cooperativeness captures the overall propensity of people to cooperate. Conditionality refers to their tendency to cooperate conditionally. We find that cooperativeness does not differ across the three treatments but conditionality does. Moreover, we show how participants in Description were more responsive to information regarding others' behaviour. On the other hand, people in Experience were significantly less sensitive to this information. We connect Experience's insensitivity to a 'regression to the mean' effect which can explain the reversal of the gap in cooperation. We propose that a key reason why similar reversals do not happen more often in the individual context is an asymmetry in the strength of priors (stronger in the social context) and discuss how this hypothesis can be addressed in future research.

BIBLIOGRAPHY

Abdellaoui, M., L'Haridon, O., and Paraschiv, C. (2011). Experienced vs. described uncertainty: Do we need two prospect theory specifications? *Management Science*, 57(10):1879–1895.

Andreoni, J. and Miller, J. H. (1993). Rational cooperation in the finitely repeated prisoner's dilemma: Experimental evidence. *The Economic Journal*, 103(418):570–585.

Artinger, F., Fleischhut, N., Levati, M. V., and Stevens, J. R. (2012). Cooperation in a risky environment: Decisions from experience in a stochastic social dilemma. Technical report, Jena Economic Research Papers.

Barron, G. and Erev, I. (2003). Small feedback-based decisions and their limited correspondence to description-based decisions. *Journal of Behavioral Decision Making*, 16(3):215–233.

Bolton, G. E. and Ockenfels, A. (2010). Betrayal aversion: Evidence from Brazil, China, Oman, Switzerland, Turkey, and the United States: Comment. *American Economic Review*, 100(1):628–33.

Cooper, R., DeJong, D. V., Forsythe, R., and Ross, T. W. (1996). Cooperation without reputation: experimental evidence from prisoner's dilemma games. *Games and Economic Behavior*, 12(2):187–218.

Dawes, R. M. and Thaler, R. H. (1988). Anomalies: cooperation. *Journal of Economic Perspectives*, 2(3):187–197.

Dijkstra, J. and van Assen, M. A. (2017). Explaining cooperation in the finitely repeated simultaneous and sequential prisoner's dilemma game under incomplete and complete information. *The Journal of Mathematical Sociology*, 41(1):1–25.

Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *The Quarterly Journal of Economics*, 75(4):643–669.

Fehr, E. and Fischbacher, U. (2003). The nature of human altruism. *Nature*, 425(6960):785–791.

Fehr, E. and Gächter, S. (2002). Altruistic punishment in humans. *Nature*, 415(6868):137–140.

Fischbacher, U. and Gächter, S. (2010). Social preferences, beliefs, and the dynamics of free riding in public goods experiments. *American Economic Review*, 100(1):541–56.

Fischbacher, U., Gächter, S., and Fehr, E. (2001). Are people conditionally cooperative? evidence from a public goods experiment. *Economics Letters*, 71(3):397–404.

Fleischhut, N., Artinger, F., Olschweski, S., Volz, K. G., and Hertwig, R. (2014). Sampling of social information: decisions from experience in bargaining. In *36th Annual Meeting of the Cognitive Science Society*, pages 1048–1053. Cognitive Science Society.

Glöckner, A., Hilbig, B. E., Henninger, F., and Fiedler, S. (2016). The reversed description-experience gap: Disentangling sources of presentation format effects in risky choice. *Journal of Experimental Psychology: General*, 145(4):486–508.

Hau, R., Pleskac, T. J., Kiefer, J., and Hertwig, R. (2008). The description–experience gap in risky choice: The role of sample size and experienced probabilities. *Journal of Behavioral Decision Making*, 21(5):493–518.

Hertwig, R., Barron, G., Weber, E. U., and Erev, I. (2004). Decisions from Experience and the Effect of Rare Events in Risky Choice. *Psychological Science*, 15(8):534–539.

Hertwig, R. and Erev, I. (2009). The description–experience gap in risky choice. *Trends in Cognitive Sciences*, 13(12):517–523.

Martin, J. M., Gonzalez, C., Juvina, I., and Lebiere, C. (2014). A description–experience gap in social interactions: Information about interdependence and its effects on cooperation. *Journal of Behavioral Decision Making*, 27(4):349–362.

Peer, E., Brandimarte, L., Samat, S., and Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70:153–163.

Rosner, B., Glynn, R. J., and Lee, M.-L. T. (2006). The Wilcoxon signed rank test for paired comparisons of clustered data. *Biometrics*, 62(1):185–192.

Sally, D. (1995). Conversation and cooperation in social dilemmas: A meta-analysis of experiments from 1958 to 1992. *Rationality and Society*, 7(1):58–92.

Santa, J. C., Exadaktylos, F., and Soto-Faraco, S. (2018). Beliefs about others' intentions determine whether cooperation is the faster choice. *Scientific*

Reports, 8(1):7509.

Selten, R. (1967). Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperiments. In *Sauermann, H. (Ed.), Beiträge zur experimentellen Wirtschaftsforschung*, pages 136–168. J.C.B. Mohr, Tübingen.

Shafir, E. and Tversky, A. (1992). Thinking through uncertainty: Nonconsequential reasoning and choice. *Cognitive Psychology*, 24(4):449–474.

Weber, E. U., Shafir, S., and Blais, A.-R. (2004). Predicting risk sensitivity in humans and lower animals: risk as variance or coefficient of variation. *Psychological Review*, 111(2):430–445.

Wulff, D. U., Mergenthaler-Canseco, M., and Hertwig, R. (2018). A meta-analytic review of two modes of learning and the description-experience gap. *Psychological Bulletin*, 144(2):140–176.

5

GENERAL CONCLUSIONS

"I may not have gone where I intended to go,
but I think I have ended up where I needed to be."

— Douglas Adams, *The Long Dark Tea-Time of the Soul* (1988)

This thesis contributes to the understanding of the 'Description - Experience (DE) gap' (Barron and Erev, 2003; Hertwig et al., 2004; Weber et al., 2004), which posits that people's risky choices are - at least partly - dependent on the way information about uncertainty is communicated: from description or from experience.

Over the past years, this discrepancy has been established in experimental studies where participants are assigned to one of two conditions. In the *Description* condition, they make choices between gambles whose properties (outcomes and outcome-probabilities) are explicitly and completely described in numerical form. Conversely, in the *Experience* condition, this information is inferred through a sequential sampling process. The key finding relates to the role of rare (low probability) events. According to the canonical interpretation of this finding, people in Description tend to make decisions *as if* overweighting such events, relative to their probability, whereas in Experience they tend to make decisions consistent with underweighting them (Hertwig et al., 2004).

In a series of lab and online experiments, we try to address some of the most important open questions in this literature. Briefly, these questions can be summarized as: 'What are the underpinnings of the DE gap?' (Chapter 2), 'How do people search in Experience and how do different search patterns influence ensuing risky choices?' (Chapter 3) and 'Is there a DE gap in social domain and if so, what are its characteristics?' (Chapter 4).

More specifically, in Chapter 2 we taxonomized what we believe are the key factors driving this empirical discrepancy in three broad categories by distinguishing between factors pertaining to: informational (sampling bias), preferential (ambiguity) or cognitive (likelihood representation and memory limitations) aspects of decision making. Then, we implemented a novel 5-treatment design comprising of one standard version of *Description* and 4 variations of *Experience*. Our treatment protocol was designed to isolate these factors through a series of pairwise comparisons. Moreover, to address certain methodological concerns in this literature, we employed two measuring approaches.

First, we study the gap in the absence of any behavioural model assumptions, by focusing only on choice proportions from pair gamble questions. We find an overall significant DE gap which is compatible in direction and equal in size with the literature's average. Among the driving forces, the most potent in isolation is sampling bias due to under-representation of rare events. Second, assuming a rank dependent utility model (Quiggin, 1982) and using the 'source method' (Abdellaoui et al., 2011a,b), we elicit and compare decision-weighting functions across treatments. This level of analysis allowed us to examine an integral component of the DE gap (probability weighting) while controlling for other aspects of risky behaviour (such as utility curvature).

We find the two levels of analysis to be complimentary. The model-mediated approach replicates the findings of the model-free analysis while the use of weighting functions and their shape allows us to shed some more light on aspects of behaviour that would otherwise be inaccessible. Most notably, we observe that a significant DE gap can be found even when controlling for sampling bias; a result that runs opposite to some previous claims (Fox and Hadar, 2006). In the absence of sampling bias, the role of memory limitations appears to be important, especially when combined with the effect of likelihood representation and ambiguity.

This finding serves as a reminder that cognitive aspects of behaviour are not to be discounted when eliciting risky preferences nor when extrapolating lab findings to policy decisions. To this end, we discussed two intriguing interpretations of our finding. The first interpretation parallels memory limitations to a type of cognitive sampling bias. The statistical property of the binomial distribution under which rare events are under-represented in small samples, might also be a property of the brain. Limited recollections of past events may also under-represent the rare ones. The second interpretation draws from the evolutionary principle of bias complementarity (Waldman, 1994). Specifically, we discuss the possibility that the behavioural bias to overweight rare events (Tversky and Kahneman, 1992) may be countering the statistical (and/or cognitive) bias to under-represent them in small - collected or recollected - samples.

Nonetheless, the size of the gap in the absence of sampling bias is relatively small. Therefore, although seeking for a DE gap beyond sampling bias is not a "fool's errand", our results recommend that the seeker be equipped with a magnifying glass, for the gap is small, and a compass, for it is not ubiquitous.

Lastly, except for when statistical or cognitive under-representations were present, we find that the standard inverse S-shaped probability weighting pattern is a good fit for Description as well as for Experience. We interpret this as evidence that Cumulative Prospect Theory's (Tversky and Kahneman, 1992) behavioural tenet that rare events are overweighted, can be found beyond the narrow frame of described uncertainty.

One of the key strengths of Chapter 2 lies in the design of an experimental protocol that controls strictly which aspects of Experience are effective. For example, to control the accuracy of the information that participants receive (and therefore determine whether sampling bias is effective or not), we do not allow participants to collect samples autonomously. This could also be interpreted, however, as a potential limitation. The freedom to search for information without restrictions may be an important aspect of the DE gap.

In Chapter 3, we build on this idea by focusing primarily on the aspect of search in decisions from experience. In this chapter, we conduct a lab-experiment and examine how people search for information in Experience as well as how do different search patterns influence ensuing risky choices. With respect to search patterns, we show that a lottery's variance is negatively correlated with sampling amount which in this context means that people sample more from options with rarer events. We also find that sampling amount decreases over time periods. Both of these findings become less salient after the introduction of a history table which records and displays previously sampled outcomes at the time of the lottery evaluation. The cue which stands out in that case is the maximum capacity of that table.

With respect to choices, we elicit preferences over gambles from two variations of Experience: with and without a history table and compare them with those elicited from a standard version of Description. Both of these comparisons generate a significant DE gap which is mitigated, however, by the inclusion of the history table. Similarly to Chapter 2, we interpret these choices through a preference model that accounts for probability weighting. Analyzing the weighting patterns between Description and Experience, we propose a variant of the canonical interpretation of the DE gap. We refer to it as the 'relative underweighting hypothesis', which states that rare events in individual risky decisions are overweighted in Experience too, but less so than in Description.

Extensions of this work could explore search strategies in more complex environments. The DE gap framework addresses situations where people have to choose between only two alternatives. In many realistic scenarios however, decision makers are faced with choice between several alternatives (e.g. buying a house, looking for the right job). This latter choice architecture scenario, falls under a broad category of problems which is commonly referred to as 'optimal stopping problems' (see Weitzman, 1979). Over the last 30 years, significant theoretical progress has been made on identifying optimal strategies of search in a wide variety of such problems. However, the empirical question of how people fare with respect to these theoretical benchmarks, remains largely unexplored (see Bearden et al., 2006, for one such exception).

Lastly, in Chapter 4, we transfer the DE gap from its usual individual-decision setting, to a social setting. Our primary goal is to investigate whether the format in which social information is obtained -descriptive or experiential-influences cooperation in social dilemmas and if so, how. To address these questions, we develop a variation of the prisoner's dilemma game, where we observe cooperative responses over a range of likelihoods of cooperation. The likelihoods are communicated either in descriptive or experiential formats. In line with previous findings in the literature of social dilemmas (Fischbacher and Gächter, 2010), we find that conditional cooperation -the willingness to cooperate if others do the same- is prevalent in our study. Cooperation rates across treatments increase with the probability of cooperation. Nonetheless, there are significant differences in the cooperation patterns between Description and Experience. Interestingly, we find evidence that this gap in social decision making, is in the opposite direction from what the canonical finding in the individual context would have predicted. Rare events (of cooperation or defection) appear to be more overweighted in Experience rather than in Description. Another asymmetry with the individual domain is that sampling bias, the predominant driver of the DE gap in risky choices, does not affect the gap in the social domain.

Further analysis on these, surprising at face value, results leads us to conclude that the reversal of the canonical DE gap in the social context is due to a decrease in sensitivity towards social information in Experience. We attribute this disparity between individual and social context to an asymmetry in the strength of priors. Specifically, we argue that priors regarding the likelihood of cooperation in a social dilemma, are stronger than any priors regarding the likelihood of winning in an abstract gamble. Furthermore, we show how people who are reluctant to update their beliefs in light of new information, are (on average) more likely to overweight rare events.

An extension of this study could entail a set of treatments that provide an empirical test of this explanation. An example in this direction could involve two different Description treatments. One in which risk is expressed in its typical abstract frame and one in which it is framed in terms that are likely to evoke strong predispositions (for example 'risks of drinking and driving' or 'risks of travelling by plane vs. risk of travelling by car'). According to the 'strong priors' hypothesis, the 'loaded' version of Description will exhibit more overweighting of rare events than the abstract one.

Overall, this investigation surfaced a series of gaps between Description and Experience, highlighting that besides the content of the information, the format in which this is communicated is also important. Nonetheless, we also discovered that the size and direction of this gap can be sensitive to a number of parameters. We take this as evidence that the DE gap should not be seen as a dichotomy between Description and Experience. Instead, we recommend that it is viewed as a continuum over different levels of uncertainty. From this perspective, the DE gap can work as a malleable experimental framework that can help researchers navigate through the rich domain of uncertainty, while at the same time, bring canonical experimental designs closer to real consumer decision problems.

BIBLIOGRAPHY

Abdellaoui, M., Baillon, A., Placido, L., and Wakker, P. P. (2011a). The Rich Domain of Uncertainty: Source Functions and Their Experimental Implementation. *American Economic Review*, 101(2):695–723.

Abdellaoui, M., L'Haridon, O., and Paraschiv, C. (2011b). Experienced vs. described uncertainty: Do we need two prospect theory specifications? *Management Science*, 57(10):1879–1895.

Barron, G. and Erev, I. (2003). Small feedback-based decisions and their limited correspondence to description-based decisions. *Journal of Behavioral Decision Making*, 16(3):215–233.

Bearden, J. N., Rapoport, A., and Murphy, R. O. (2006). Sequential observation and selection with rank-dependent payoffs: An experimental study. *Management Science*, 52(9):1437–1449.

Fischbacher, U. and Gächter, S. (2010). Social preferences, beliefs, and the dynamics of free riding in public goods experiments. *American Economic Review*, 100(1):541–56.

Fox, C. R. and Hadar, L. (2006). "Decisions from experience" = sampling error + prospect theory: Reconsidering Hertwig, Barron, Weber & Erev (2004). *Judgment and Decision Making*, 1(2):159–161.

Hertwig, R., Barron, G., Weber, E. U., and Erev, I. (2004). Decisions from Experience and the Effect of Rare Events in Risky Choice. *Psychological Science*, 15(8):534–539.

Quiggin, J. (1982). A theory of anticipated utility. *Journal of Economic Behavior & Organization*, 3(4):323–343.

Tversky, A. and Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323.

Waldman, M. (1994). Systematic errors and the theory of natural selection. *American Economic Review*, 84(3):482–497.

Weber, E. U., Shafir, S., and Blais, A.-R. (2004). Predicting risk sensitivity in humans and lower animals: risk as variance or coefficient of variation.

Psychological Review, 111(2):430–445.

Weitzman, M. L. (1979). Optimal search for the best alternative. *Econometrica*, 47(3):641–654.

A

APPENDIX FOR CHAPTER 2

A.1 INSTRUCTIONS

Instructions were handed to participants in printed form and were read out loud by the experimenter prior to the start of the experiment. Before the start of the experiment and after the instructions had been read out loud, subjects played one trial round.

A.1.1 *Instructions for Description*

In this study you are asked to make choices that involve lotteries. For each choice, just pick the option you prefer as there are no 'right' or 'wrong' answers. Overall you are going to consider a total of 19 lotteries which are described by virtual decks of cards. Each deck contains exactly two types of cards represented by two different colours. Each deck has its own mix of these two types of cards.

The information about the relative frequency and the monetary value of each type of card will be provided to you (in the form of percentages) prior to making a choice. This information is seen on the bottom of the screen.

The first 7 lotteries are all associated with the same deck of cards. This guarantees that the relative frequency of each colour is the same for Lotteries 1 to 7. Notice however that the rewards associated with each outcome will differ from one lottery to another.

Later in the experiment, you may have the opportunity to 'play' a lottery. That would mean drawing once more from a deck you have sampled and receiving the sum of money assigned to the colour of the drawn card.

Your task is to choose each time between playing the Lottery and receiving the Certain Outcome. Each Lottery entails 5 such choices between the Lottery (Option A) which remains constant across these 5 Choice-Rounds and a Certain Outcome (Option B) that will be changing from each choice to the next.

Payoff Stage

At the end of the experiment one choice is going to be randomly selected to be played out for real. All choices are equally likely to be drawn so each choice you make has equal chances of affecting your final payment. There are two cases:

Case 1: If in the randomly selected choice you chose Option B (the Certain Outcome) then the monetary value of this choice is going to be added directly to your final payment.

Case 2: If in the randomly selected choice you chose Option A (the Lottery) then the deck of cards corresponding to that choice will reappear on the screen. You will then be asked to draw one card from it. Then the monetary value assigned to the colour of the card you just drew will be added to your final payment.

A.1.2 *Instructions for E-Unamb*

In this study you are asked to make choices that involve lotteries. For each choice, just pick the option you prefer as there are no 'right' or 'wrong' answers. Overall you are going to consider a total of 19 lotteries which are described by virtual decks of cards. Each deck contains **exactly two** types of cards, represented by two different colours. Each deck has its own mix of these two types of cards.

For every lottery you go through two stages:
Stage 1: the 'Sampling Stage'
Stage 2: the 'Choice Stage'
Exception: The first 7 Lotteries all share the same 'Sampling Stage' because they relate to the same deck. This means that you will only sample once for the first seven lotteries. Each of the lotteries 8 - 24 has its own Sampling Stage (because it relates to its own deck).

Stage 1: Sampling Stage

In each Sampling Stage you go through a particular computerized deck and explore one by one all of their cards. The information about the relative frequency of each type of card is unknown to you prior to the start of the sampling process. However by the end of the process, this information will be completely revealed to you as you will have seen every card in the deck exactly once. As mentioned earlier, the first 7 lotteries relate to the same deck. This guarantees that the relative frequency of each colour is the same for Lotteries 1 to 7. We recommend that you pay attention during this sampling process as this information is relevant for your decisions later on and hence your final payment.

Every time you click on the '*Draw*' button you will observe a new card from the deck. Once you observe its colour click on '*Proceed with the next card*' for the '*Draw*' button to reappear. You will repeat this process until you go exactly once through all the cards in each deck. Once you have done so, a message will appear on the screen verifying that you have seen all the cards in this deck and a button that reads: '*Go to the Choice Stage*' will become accessible at the bottom of the screen. Once you click on that button you will move on to the 'Choice Stage'.

Stage 2: 'Choice Stage'

At this stage a monetary value is assigned to the colour of each card. This information is seen on the bottom of the screen. Later in the experiment, you may have the opportunity to "play" a lottery. That would mean drawing once more from a deck you have sampled and receiving the sum of money assigned to the colour of the drawn card.

On the top of the screen you will observe a 'History Table' where you can track your sampling history from each lottery's 'Sampling Stage'. As mentioned earlier, the first 7 lotteries are all associated with the same deck of cards and hence share the same "History Table". Notice however that although the relative frequency of each colour of card is the same for lotteries 1 to 7, the rewards associated with each outcome will differ from one lottery to another.

Your task in this stage is to choose each time between playing the Lottery and receiving the Certain Outcome. Each Lottery entails 5 such choices between the Lottery (Option A) which remains constant across these 5 Choice-Rounds and a Certain Outcome (Option B) that will be changing from each choice to the next.

Payoff Stage

At the end of the experiment one choice is going to be randomly selected to be played out for real. All choices are equally likely to be drawn so each choice you make has equal chances of affecting your final payment. There are two cases:

Case 1: If in the randomly selected choice you chose Option B (the Certain Outcome) then the monetary value of this choice is going to be added directly to your final payment.

Case 2: If in the randomly selected choice you chose Option A (the Lottery) then the deck of cards corresponding to that choice will reappear on the screen. You will then be asked to draw one card from it. Then the monetary value assigned to the colour of the card you just drew will be added to your final payment.

A.1.3 *Instructions for E-NR*

Instructions for this treatment were identical to those for E-Unamb, except that there was no reference to a history table.

A.1.4 *Instructions for E-Amb*

For every lottery you go through two stages: Stage 1: the "Sampling Stage" Stage 2: the "Choice Stage" Exception: The first 7 Lotteries all share the same "Sampling Stage" because they relate to the same deck. This means that you will only sample once for the first seven lotteries. Each of the lotteries 8 - 24 has its own Sampling Stage (because it relates to its own deck).

Stage 1: Sampling Stage

In each Sampling Stage you draw a sample of cards from a particular computerized deck. The information about the relative frequency of each type of card is unknown to you prior to the start of the sampling process. However by the end of this process you will have discovered something more about this mix because you will have seen a selection of draws from that deck. As mentioned earlier, the first 7 lotteries relate to the same deck. This guarantees that the relative frequency of each colour is the same for Lotteries 1 to 7. We recommend that you pay attention during this sampling process as this information is relevant for your decisions later on and hence your final payment.

Every time you click on the "Draw" button one of the two types of cards will appear. Once you observe its colour click on "Draw again" for the "Draw" button to reappear. You will repeat this process until the "Go to the Choice Stage" button becomes available at the bottom of the screen. Once you click on that button you will move on to the "Choice Stage".

Stage 2: 'Choice Stage'

At this stage a monetary value is assigned to the colour of each card. This information is seen on the bottom of the screen. Later in the experiment, you may have the opportunity to "play" a lottery. That would mean drawing once more from a deck you have sampled and receiving the sum of money assigned to the colour of the drawn card.

On the top of the screen you will observe a 'History Table' where you can track your sampling history from each lottery's 'Sampling Stage'. As mentioned earlier, the first 7 lotteries are all associated with the same deck of cards and hence share the same "History Table". Notice however that although the relative frequency of each colour of card is the same for lotteries 1 to 7, the rewards associated with each outcome will differ from one lottery to another.

Your task in this stage is to choose each time between playing the Lottery and receiving the Certain Outcome. Each Lottery entails 5 such choices between the Lottery (Option A) which remains constant across these 5 Choice-Rounds and a Certain Outcome (Option B) that will be changing from each choice to the next.

Payoff Stage

At the end of the experiment one choice is going to be randomly selected to be played out for real. All choices are equally likely to be drawn so each choice you make has equal chances of affecting your final payment. There are two cases:

Case 1: If in the randomly selected choice you chose Option B (the Certain Outcome) then the monetary value of this choice is going to be added directly to your final payment.

Case 2: If in the randomly selected choice you chose Option A (the Lottery) then the deck of cards corresponding to that choice will reappear on the screen. You will then be asked to draw one card from it. Then the monetary value assigned to the colour of the card you just drew will be added to your final payment.

A.1.5 *Instructions for E-Restr*

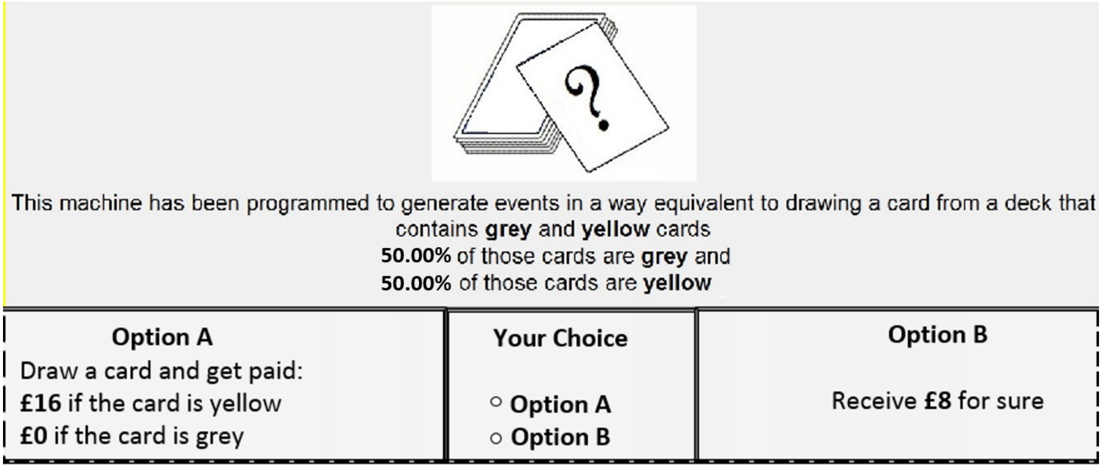
Instructions for E-Restr were identical to those in E-Amb.

B

APPENDIX FOR CHAPTER 3

B.1 INTERFACE FOR DESCRIPTION

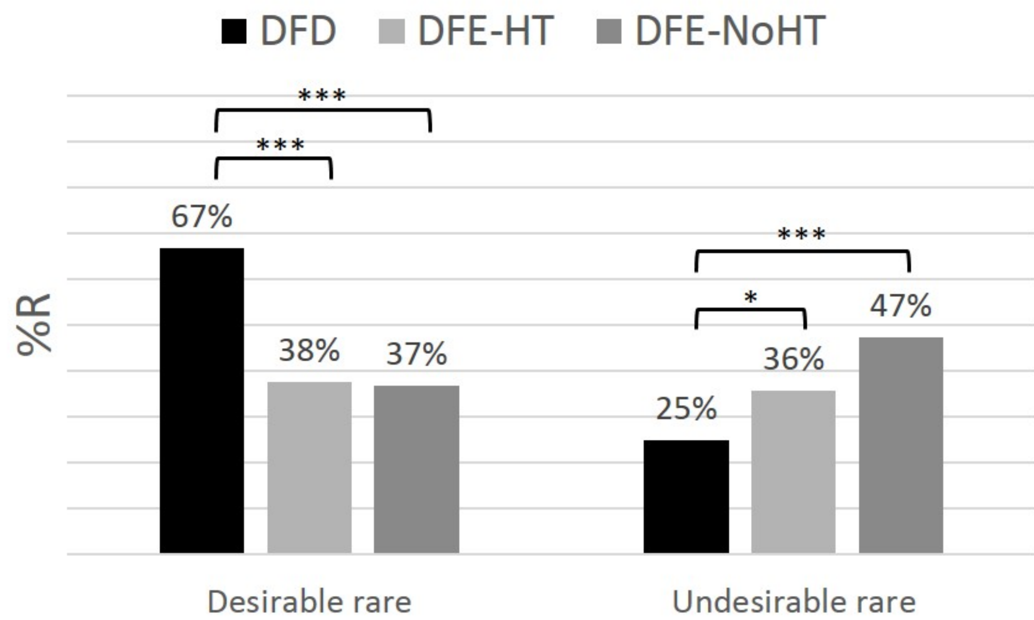
Figure B.1: Instance from DFD's information-evaluation stage



Note. Information about the deck and lottery evaluation take place in the same screen in DFD.

B.2 CHOICE PATTERNS

Figure B.2: Choice patterns in cases where rare events have been under-represented



Note. '***': p-value< 0.01, '*': p-value< 0.10

Table B.1: Median experienced probabilities for DFE-HT and DFE-NoHT

	DFE-HT	DFE-NoHT
p	E_p	
0.025	0.030	0.025
0.050	0.069	0.052
0.100	0.089	0.103
0.250	0.231	0.230
0.500	0.521	0.504
0.750	0.753	0.752
0.900	0.915	0.889
0.950	0.951	0.947
0.975	0.983	0.977

Note. With the exception of p=0.975 for DFE-HT (p-value< 0.01, two-sided MW-test), we are never able to reject the hypothesis that $p = E_p$.

Table B.2: Decision set from Hertwig et al.(2004)

Decision problem	Lotteries Risky	Safe
1	$(4, E_{0.8}; 0)$	$(3, E_{1.0})$
2	$(4, E_{0.2}; 0)$	$(3, E_{0.25}; 0)$
3	$(-3, E_{1.0})$	$(-32, E_{0.1}; 0)$
4	$(-3, E_{1.0})$	$(-4, E_{0.8}; 0)$
5	$(32, E_{0.1}; 0)$	$(3, E_{1.0})$
6	$(32, E_{0.025}; 0)$	$(3, E_{0.25}; 0)$

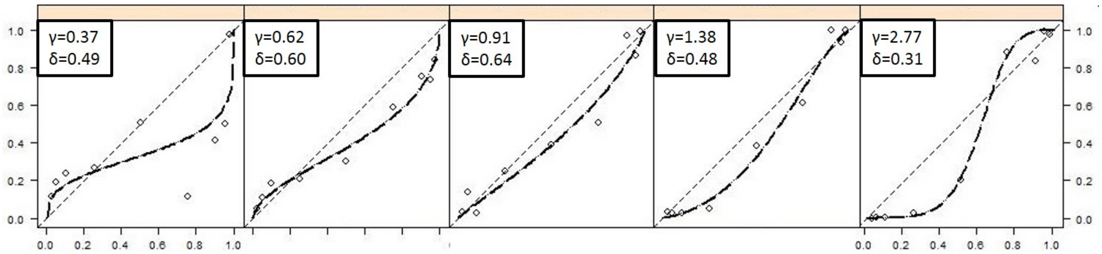
Table B.3: Choice patterns in 'control' lotteries

Decision	Lotteries		$\%R$		
problem	Risky	Safe	DFD	DFE-HT	DFE-NoHT
A	$(4, E_{0.8}; 0)$	$(3.2, E_{1.0})$	26%	30%	31%
B	$(4, E_{0.2}; 0)$	$(0.8, E_{1.0})$	69%	73%	53%
C	$(3, E_{0.25}; 0)$	$(0.75, E_{1.0})$	67%	65%	67%

Note. Risky options in this table were included as 'control' tasks due to their similarity with some of the commonly used problems in the sampling paradigm (see Table B.2). For example decision problem 2 in Table B.2 corresponds to a choice between the risky option in B and the risky option in C from this table. Since these lotteries could only be evaluated separately in this study, we can compare choice patterns only indirectly by comparing $\%R$ across problems B and C. According to early DE gap, $\%R$ should be higher in B than in C for DFD while the opposite must be true for DFE. This pattern is verified in the comparison between DFD and DFE-NoHT but not between DFD and DFE-HT. Moreover, $\%R$ should be higher in DFE than in DFD for problem A. This is indeed the case for both DFE-NoHT and DFE-HT. All of the aforementioned differences are relatively small and not statistically significant.

B.3 individual analysis

Figure B.3: Examples of decision weights elicitation



Note. Examples of non-parametric (circles) and parametric (curves) weighting functions. From left to right: increasing values of γ for a relatively small range of δ - values. There are two ways of constructing an aggregate weighting function from these five examples. The parametric approach entails aggregating across those five $\gamma's$ and $\delta's$ while according to the non-parametric one, we would aggregate across each level of probability.

Table B.4: Classification of subjects according to the curvature of utlity and weighting functions

	Utility function			Weighting function		
	Concave $\alpha < 0.9$	Linear $\alpha \in [0.9, 1.1]$	Convex $\alpha > 1.1$	Inverse S-shaped $\gamma < 0.9$	No curvature $\gamma \in [0.9, 1.1]$	S-shaped $\gamma > 1.1$
DFD	33.3%	25.6%	41.0 %	82.1 %	5.1 %	12.8 %
DFE-HT	35.0%	20.0%	45.0 %	62.5 %	12.5 %	25.0 %
DFE-NoHT	35.9%	20.5%	43.6 %	69.2 %	7.7 %	23.1 %

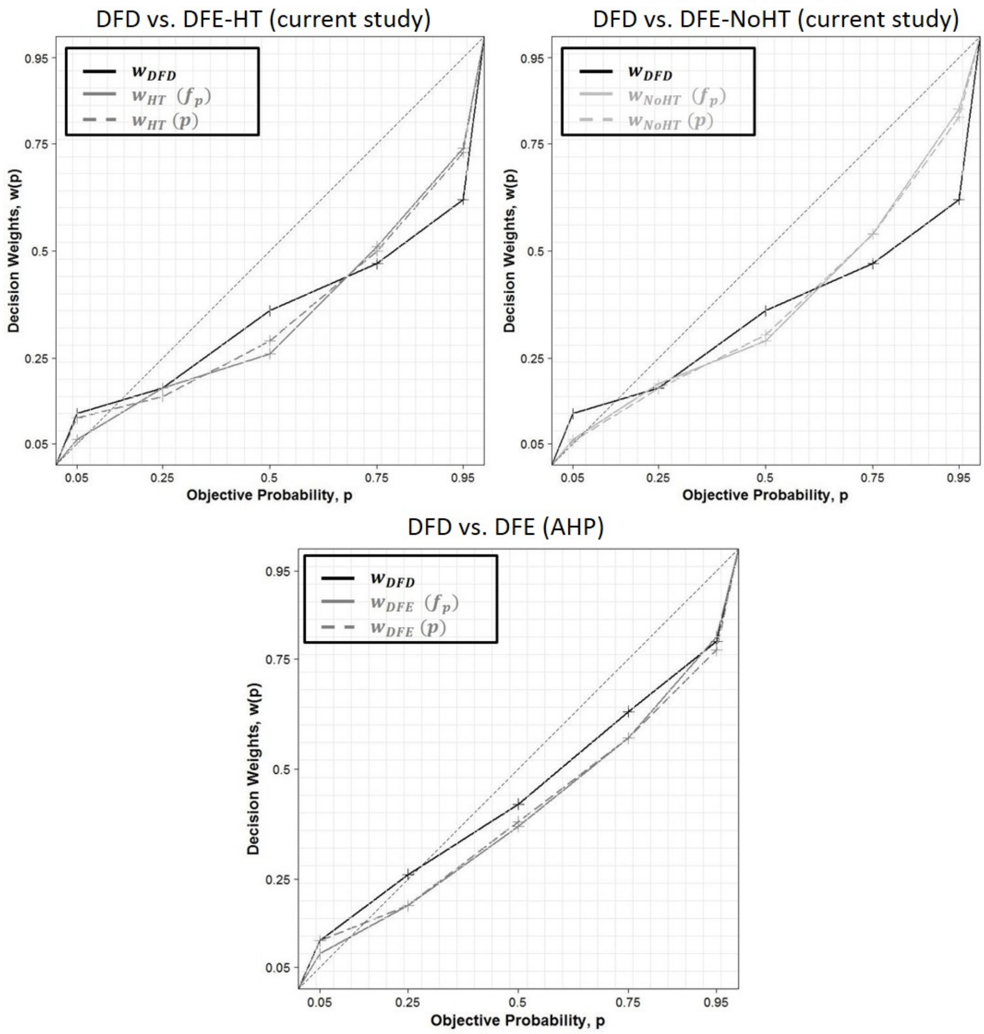
Note. Classification of γ's was made according to decision weights calculated based on objective probabilities.

B.4 NON-PARAMETRIC ANALYSIS

Table B.5: Median decision weights in this study and in AHP

	Current			AHP	
p	DFD	DFE-HT	DFE-NoHT	DFD	DFE
0.05	0.12	0.06	0.06	0.11	0.08
0.25	0.18	0.18	0.18	0.26	0.19
0.50	0.36	0.26	0.30	0.42	0.37
0.75	0.47	0.51	0.54	0.63	0.57
0.95	0.62	0.74	0.81	0.79	0.80

Figure B.4: Non-parametric weighting functions in the current study and in AHP



Note. Plotting values from Table B.5. Top row: current study; bottom row: AHP. For DFE, dashed lines are estimated according to objective probabilities (p) while solid lines according to experienced probabilities (f_p). Only probability targets included in AHP are plotted. This excludes observations at $p \in \{0.025, 0.10, 0.90, 0.975\}$ from this study.

C

APPENDIX FOR CHAPTER 4

C.1 INSTRUCTIONS AND INTERFACE

Unless specified otherwise, all screens were encountered by participants across all three treatments.

Figure C.1: Welcome screen

Welcome

Thank you for accepting to participate in this study. You will be awarded £1.25 for your participation. You also have the opportunity to earn additional money, depending on your choices and the choices of other Prolific participants.

The study consists of three stages. In each stage, you will be provided 50 pence to interact with other participants. After you complete the study, one of your choices will be randomly selected in order to compute your additional earnings.

You cannot participate in this study more than once. You will receive your total payment within one week after the experiment.

Figure C.2: Stage 1/A

Stage 1

You have been randomly matched with another participant which we refer to as your "match".

You and your match are each provided 50 pence for this stage.

You each have to decide whether to ***Keep*** or ***Share*** your 50 pence.

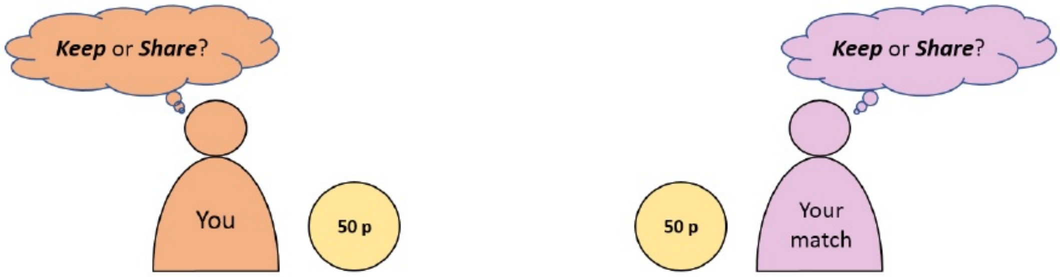
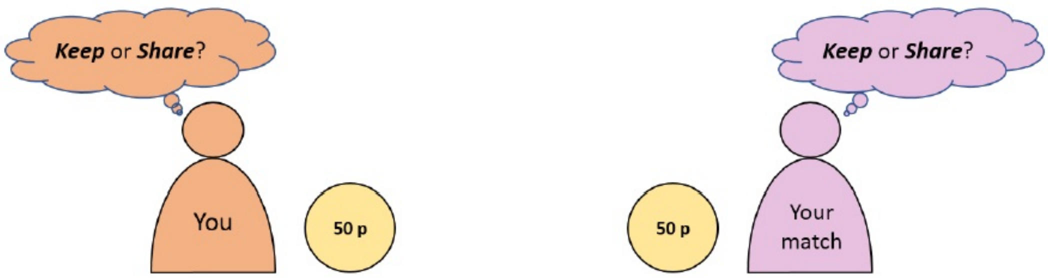


Figure C.3: Stage 1/B

Stage 1

There are **four possible outcomes** depending on the choices made by you and your match:



CHOICES			OUTCOMES		
If you choose:	&	If your match chooses:	Then you get:	&	Your match gets:
Keep		***Keep***	50 p		50 p
Share		***Share***	100 p		100p
Keep		***Share***	150 p		0 p
Share		***Keep***	0 p		150p

Figure C.4: Stage 1: Decision

Your Decision for Stage 1

Now please make your decision:

○ KEEP

○ SHARE

If you choose:	&	If your match chooses:	Then you get:	&	Your match gets:
Keep		***Keep***	50 p		50 p
Share		***Share***	100 p		100p
Keep		***Share***	150 p		0 p
Share		***Keep***	0 p		150p

Note. This is the decision interface for Stage 1.

Figure C.5: Stage 2/A

Stage 2

In Stage 2, you are assigned to **a group** of participants who have also completed Stage 1. One participant from this group will be randomly chosen as your **new match** for Stage 2.

You and your new match will participate in the **same task** as in Stage 1, with one exception: **The decision of your new match will be exactly the same as their choice in Stage 1.**

Next screen will explain how you can learn more about your new match's choice in Stage 1.

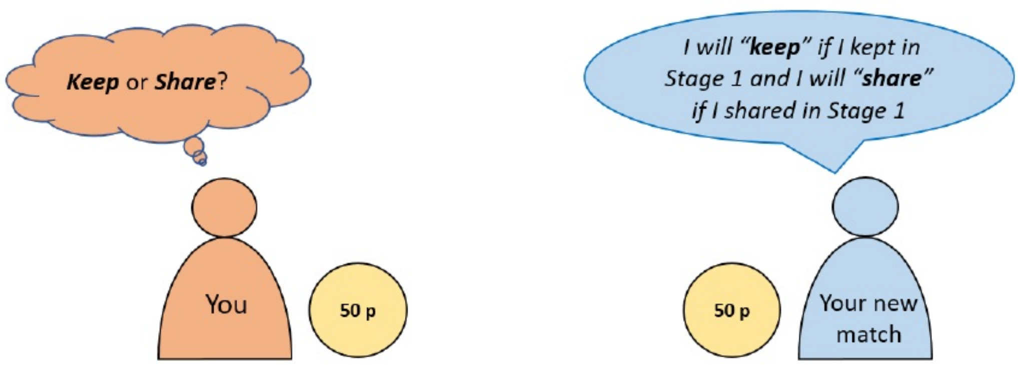
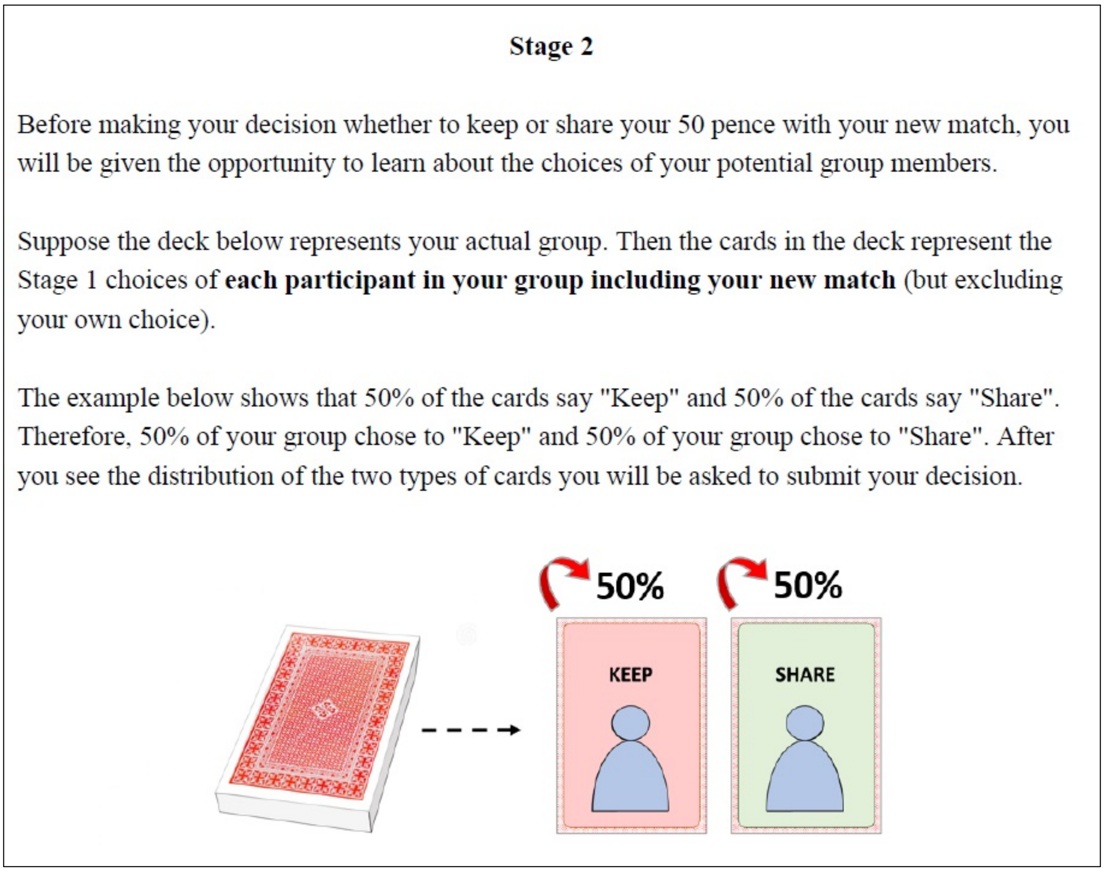
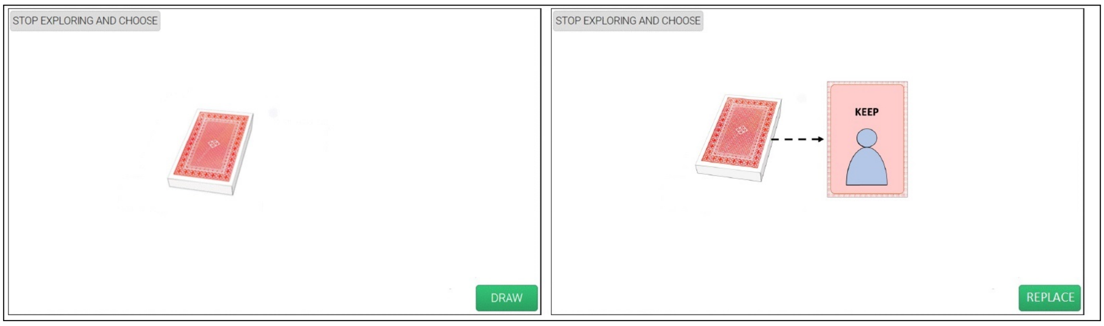


Figure C.6: Stage 2/B: Description only



Note. This screen was encountered only by participants in Description. Participants in Experience saw instead the screen in the next Figure.

Figure C.7: Stage 2/B: Experience only



Note. These two screens were encountered only by participants in E-Free. Participants in E-Fixed saw a similar demonstration but there was no 'STOP EXPLORING AND CHOOSE' button on the top-left of the screen. Moreover, the 'REPLACE' button was replaced with one that read 'NEXT CARD' as in E-Fixed sampling was without replacement.

Figure C.8: Stage 2/C

Stage 2

You will see **seven different decks of cards**.

Only **one** of the seven decks corresponds to the **actual group** you have been assigned to. The other six decks present hypothetical situations.

You do not know which deck describes your actual group. Therefore, you will make seven independent decisions to keep or share, one for each deck.

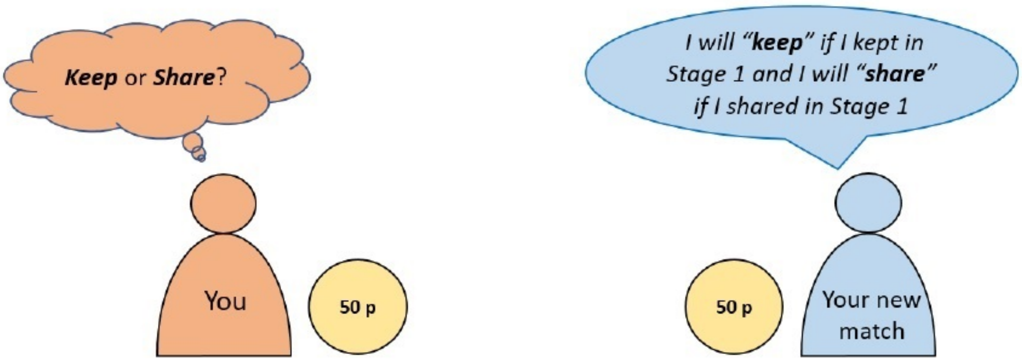
Only one of your seven decisions will be used to determine the outcome of Stage 2, which will be your decision for the deck that represents **the actual group** you are assigned to.

Hence, you should consider each deck independently of the other decks and make your decision assuming that the deck that you learn about in fact describes your actual group.

Figure C.9: Stage 2/D

Stage 2

The outcome of Stage 2 depends on what you choose for the deck representing your actual group, and what your new match from the actual group has chosen in Stage 1. The relationship between actions and earnings is exactly the same as in Stage 1, as presented below:



CHOICES			OUTCOMES		
If you choose:	&	If your new match chose:	Then you get:	&	Your new match gets:
Keep		***Keep***	50 p		50 p
Share		***Share***	100 p		100p
Keep		***Share***	150 p		0 p
Share		***Keep***	0 p		150p

Figure C.10: Stage 2: Decision

Your Decision

Please decide. Your decision will determine your earnings if your assigned group is actually represented by the deck that you just explored.

- SHARE
- KEEP

If you choose:	&	If your new match chose:	Then you get:	&	Your new match gets:
Keep		***Keep***	50 p		50 p
Share		***Share***	100 p		100p
Keep		***Share***	150 p		0 p
Share		***Keep***	0 p		150p

Note. This screen follows the screen where participants learn about the distribution of each scenario. Examples of how this is information is obtained for each scenario can be seen in Figure C.6 for Description and Figure C.7 for Experience.

Figure C.11: Stage 3/A

Stage 3

You have been randomly matched with another participant which we refer to as your "Stage 3 match".

You and your Stage 3 match are each provided 50 pence for this stage.

You and your Stage 3 match will participate in the **same task** as in the previous two stages. Similar to Stage 2, **the decision of your Stage 3 match will be exactly the same as their choice in Stage 1**.

You will make a new choice whether to ***Keep*** or ***Share*** your 50 pence **for each possible decision that your Stage 3 match could have made in Stage 1**.

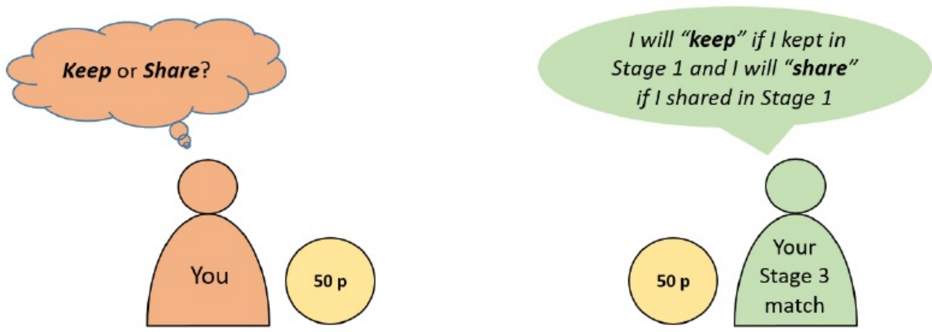
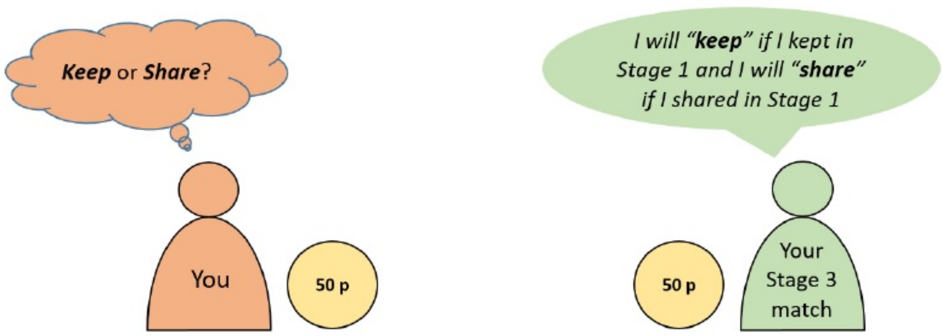


Figure C.12: Stage 3/B

Stage 3

The outcome of Stage 3 depends on what you choose in Stage 3 and what your Stage 3 match has chosen in Stage 1. The relationship between actions and earnings is exactly the same as in the previous two stages, as presented below:

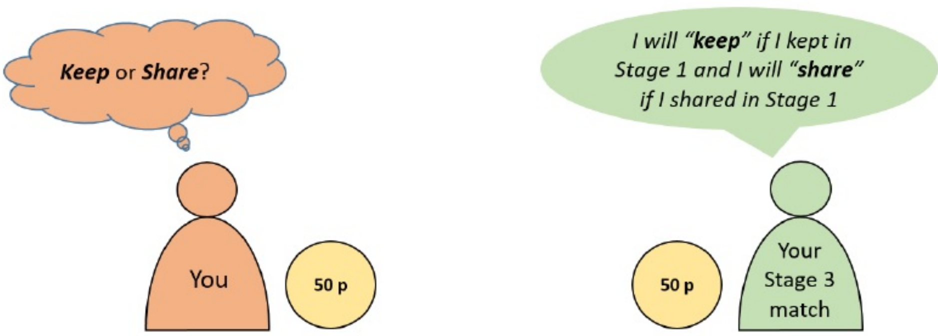


CHOICES		OUTCOMES	
If you choose: &	If your Stage 3 match has chosen:	Then you get: &	Your Stage 3 match gets:
Keep	***Keep***	50 p	50 p
Share	***Share***	100 p	100p
Keep	***Share***	150 p	0 p
Share	***Keep***	0 p	150p

Figure C.13: Stage 3/C

Stage 3

The outcome of Stage 3 depends on what you choose in Stage 3 and what your Stage 3 match has chosen in Stage 1. The relationship between actions and earnings is exactly the same as in the previous two stages, as presented below:



CHOICES			OUTCOMES		
If you choose:	&	If your Stage 3 match has chosen:	Then you get:	&	Your Stage 3 match gets:
Keep		***Keep***	50 p		50 p
Share		***Share***	100 p		100p
Keep		***Share***	150 p		0 p
Share		***Keep***	0 p		150p

C.2 MATCHING PROTOCOL AND PAYMENT

Participants were randomly assigned to a treatment and given an index number within this treatment (e.g. 1,2,3...). They were then assigned to sub-populations of 3 members. To illustrate this process assume that there are 6 participants in a treatment and that they are assigned to two sub-populations - A and B - such that $A = \{1, 2, 3\}$ and $B = \{4, 5, 6\}$. The matching algorithm consists of the following steps:

- 1 and 2 are matched for Stage 2
- 4 and 5 are matched for Stage 3
- 3 and 6 get matched across sub-populations and get paid for Stage 1
- if there is a number that is indivisible by 3:
  - if two players are left then they are matched for Stage 1
  - otherwise, if one player is left unmatched, she gets the maximum payoff (£3.75)

Matchings for Stage 2 and Stage 3 require that one of the two players acts according to her Stage 2 action while the other according to her Stage 1 action. This is randomly decided for each pair.

C.2.1 *Cooperation rate*

Table C.1: Cooperation rates across treatment

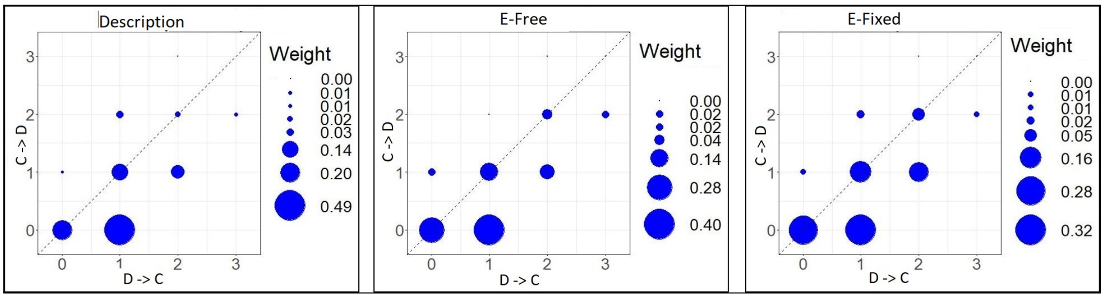
SPoC	Description	E-Free	E-Fixed
0	10.4	20.2	15.6
10	13.6	23.4	21.0
30	21.9	32.4	28.6
50	51.6	42.5	41.3
70	62.4	55.1	58.3
90	65.2	61.8	61.6
100	66.3	63.4	64.1

Note. The Sub-population Probability of Cooperation (SPoC) is the probability of being matched to a cooperative agent in a given scenario. These values correspond to the coordinates of Figure 4.1

C.2.2 *Consistent behaviour in Stage 2*

As we do not restrict behaviour in Stage 2, participants can switch multiple times between cooperation and defection. A preference profile is a selection of responses for every scenario of SPoC. We define as consistent behaviour, preference profiles that switch at most once from cooperation to defection as SPoC values increase from 0 to 100. In Figure C.14, consistent behaviour can be found in the following coordinates: $(0,0)$ (people that never switch) and $(0,1)$ (people who switch once, from defecting to cooperating). We see that approximately a third of the participants in each treatment behaved in an inconsistent manner (31%,%32,%40; for Description, E-Free and E-Fixed respectively).

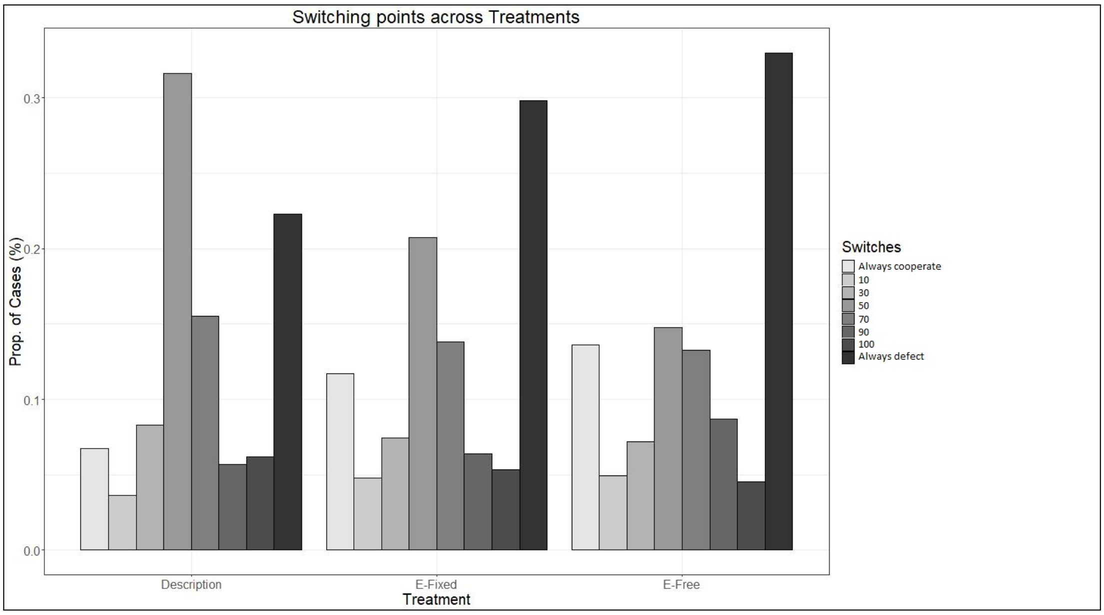
Figure C.14: Transitions of cooperative behaviour



Note. x-axis: counts number of transitions from Defection to Cooperation ('C→D'). y-axis: counts number of transitions from Cooperation to Defection ('D→C').

Building on this analysis, Figure C.15 displays the switching points of these consistent participants. In accord with our main findings we see a spike of people in Description switching when SPoC=50. Moreover, we see that the proportion of people who never switch is higher in Experience compared to Description.

Figure C.15: Switch



Note. Values on the legend (10,30...) correspond to the value of SPoC where people chose to switch from defecting to cooperating.

We can use Figure C.15 to elicit types according to cooperative preferences. This can only be achieved in Description under the assumption that people's beliefs about the likelihood of cooperation coincides with the objective probability that was given to them by the experimenter. Notice that this exercise cannot be performed for participants in Experience. The presence of ambiguity (in E-Free and E-Fixed) as well as sampling bias (in E-Free) does not allow a clear separation between beliefs and preferences.

Looking at the left cluster of barplots in Figure C.15 we can characterize as conditional cooperators those who switch at some point. Those who always cooperate can be characterized as unconditional cooperators while those who never cooperate as free riders. We can then test the correlation between Stage 2 and Stage 3 typology. Each entry in the diagonal of Table C.2 reports the percentage of people that were characterized in the same way between Stage 2 and Stage 3. This calculation is performed from the subset of participants who exhibited a consistent behaviour in Stage 2 and were not characterized as 'Others' in Stage 3. This subset's size corresponds to 64.3% of the total subject pool. Summing the diagonal elements we observe that 84% of participants in this subset are characterized in the same way between the two methods.

Table C.2: Correlation between Stages 2 and 3 for Description

S2|S3	UC	CC	FR
UC	**0.03**	0.03	0.00
CC	0.02	**0.62**	0.07
FR	0.00	0.04	**0.19**

Note. Rows: types according to Stage 2. Columns: types according to Stage 3. 'UC': Unconditional cooperators. 'CC':Conditional cooperators. 'FR': Free riders. Only participants who exhibited consistent behaviour in Stage 2 and were not characterized as 'Others' in Stage 3 are included.