DECISIONS FROM EXPERIENCE AND FROM DESCRIPTION: BELIEFS AND PROBABILITY WEIGHTING

ISBN: 978-90-3610-501-9

Cover design: Crasborn Graphic Designers bno, Valkenburg a.d. Geul

This book is no. 703 of the Tinbergen Institute Research Series, established through cooperation between Rozenberg Publishers and the Tinbergen Institute. A list of books which already appeared in the series can be found in the back.

Decisions from Experience and from Description: Beliefs and Probability Weighting

Besluiten op basis van ervaring en beschrijving: Kansopvattingen en het wegen van kansen

Thesis

to obtain the degree of Doctor from the
Erasmus University Rotterdam
by command of the
rector magnificus
Prof.dr. H.A.P. Pols

and in accordance with the decision of the Doctorate Board.
The public defence shall be held on
Thursday December 7, 2017 at 09:30 hours

by
ILKE AYDOGAN
born in Cankaya, Turkey

Erasmus University Rotterdam

Doctoral Committee:

Promotors:

Prof.dr. H. Bleichrodt

Prof.dr. A. Baillon

Other members:

Prof.dr. P.P. Wakker

Prof.dr. O. l'Haridon

Prof.dr. J. Qiu

Acknowledgment

I would first like to express my sincerest gratitude to Prof. Peter Wakker, Prof. Han Bleichrodt and Prof. Aurelien Baillon for their guidance and support during my PhD studies. Prof. Peter Wakker has been my mentor, and his lectures have made a deep and lasting influence on my research in many ways. Collaborating with Prof. Han Bleichrodt and Prof. Aurelien Baillon has been an invaluable experience, especially in the art of doing research. It has also been a privilege to work with Yu Gao, Chen Li, and Emmanuel Kemel. Collaborating with them has been very rewarding in my development as a researcher.

For four years, I have been lucky enough to be a part of the Behavioral Economics research group of Erasmus University Rotterdam. Besides benefiting from its intellectually stimulating environment, I also had the pleasure of being a part of a family outside of work. I would like to thank all my colleagues: Kirsten Rohde, Arthur Attema, Jan Stoop, Ning Liu, Vitalie Spinu, Amit Kothiyal, Rogier van Potter, Jan Heufer, Georg Granic, Xiao Yu, Ayşıl Emirmahmutoğlu, Yan Xu, and Benjamin Tereick.

During my PhD, I spent 2 months at Max Planck Institute for Human Development in Berlin as an academic visitor. It was a valuable research experience that enriched my research perspective. I want to thank Prof. Ralph Hertwig and Dr. Thorsten Pachur for hosting me in Berlin.

Lastly, I would like to thank those who made my 4 years in Rotterdam interesting and worthwhile. I would like to thank Uyanga Turmunkh, with whom we shared not only an apartment for 3 years but also many common leisure pursuits from movies and music to literature. I would like to thank Mehmet Kutluay, who has always been available for a drink after a good day’s work. I would like to thank my officemate Paul van Bruggen, with whom I enjoyed talking and smoking a good pipe tobacco. I would like to thank Jingni Yang, with whom I liked having a coffee break when I felt stuck with problems at work. I would like to thank Sait Ozturk, Tong Wang, and Violeta Misheva, with whom we had occasional movie nights and dinner parties. And I would like to thank my friends in Turkey: Aslihan Koyuncu, Tora Tospaci and Semih Ceyhan, who have always been reliable and supportive. After all the years I spent abroad, they never let me feel like a stranger.

Contents

Chapter 1

Introduction

More often than not, decisions that we make under uncertainty rely on our personal experiences with the contingencies. We often come to know about possible consequences of our actions from experience, or stay completely ignorant about them due to lack of experience, which determines our perception of the riskiness or the reliability of a source of uncertainty. In fact, a summary description of possible outcomes and probabilities of any prospect is hardly ever available to us, except in some cases such as weather reports.

Decisions from description (henceforth, DFD) typically concern *risk* in the literature on decision making. It is identified as a case where outcome probabilities are objectively known. Decisions from experience (henceforth, DFE), on the other hand, represent a case of *ambiguity*. Here, the outcome probabilities are not known objectively but they are subjectively inferred based on finite number of observations from a (finite or infinite) population of outcomes. As in many real life situations, probabilistic inference and information search are integral parts of DFE. Thus, they provide a realistic framework to understand decisions under uncertainty.

Two influential studies by Barron & Erev (2003) and Hertwig, Barron, Weber and Erev (2004) indicated that there is an intriguing behavioral gap between these two choice paradigms. In particular, these studies claimed that rare and extreme outcomes – so-called “black swans” – are overweighted under DFD, whereas they tend to be underweighted or even neglected under DFE. Both overweighting and underweighting (or neglect) stand as deviations from Expected Utility theory (henceforth, EU) under risk (Von Neumann & Morgenstern, 1944) and Subjective Expected Utility theory (henceforth, SEU) under uncertainty (Savage, 1954), which are considered the rational models of choice in Economics.

Overweighting of rare outcomes under DFD is a robust empirical phenomenon. It is usually associated with separate attention induced by an explicit reminder of a good or a bad rare outcome, which triggers emotions of hope and fear. Nevertheless, the impact of rare outcomes under DFE has not been settled yet. Previous studies suggested that the neglect of rare outcomes while making DFE is mainly due to ignorance about them caused by lack of experience. However, whether a rare outcome is still neglected (or underweighted) when its existence is known by observation, or when it is anticipated, under DFE has been unclear.

This dissertation explores the DFD-DFE gap, specifically the weighing of uncertainty under DFD and under DFE, in more detail. The main premise of this thesis is that the DFD-DFE gap has two components: (1) probability weighting, and (2) subjective beliefs. The first component is essential for understanding the differences between DFD and DFE with respect to attitudes towards outcome probabilities. For example, a decision maker can assign higher or lower decision weights to favorable outcome probabilities (optimism vs. pessimism), or she can be more or less sensitive to changes in outcome probabilities under DFE than under DFD. In this thesis, such attitudes towards outcome probabilities are modelled by source dependent probability weighting functions. This means that the impact of learning experience and ambiguity under DFE are observable through a comparison of probability weighting functions that are specific to the sources of described risk and experienced uncertainty.

The second component, understanding subjective beliefs, is also essential because DFE entail probabilistic inference while DFD does not. Specifically, the decision maker under DFE first has to estimate the probabilities of outcomes based on her own sampling observations. On the contrary, no probability estimation is required under DFD because the objective probabilities are readily available to the decision maker. Importantly, the probability estimations may deviate from objective probabilities, due to sampling error (e.g. under-observation of rare outcomes), over-or under-estimation of probabilities or belief updating. This discrepancy is another source of the DFD-DFE gap.

In what follows, chapters 2 and 3 aim to shed light on the DFD-DFE gap by disentangling beliefs from probability weighting under DFE. Chapter 2 provides an empirical investigation of the impact of learning experience on probability weighting while controlling for the impact of probability estimations and ambiguity. Different from previous DFE studies, the experiment in this chapter examined DFE under risk rather than under ambiguity. Thus, DFD and DFE treatments were two different cases of risk that differed only with respect to information acquisition: through description or through sampling experience. Our experimental design also resolved a number of methodological difficulties in previous studies such as sampling bias, an aggregation problem, and the impact of utilities. This chapter starts with a critical review of the literature on the DFD-DFE gap.

Chapter 3 examines DFE under ambiguity. It introduces a two-stage decision model for DFE that emphasizes the role of prior beliefs. The two-stage model assumes that (1) the subjective probabilities are estimated in a Bayesian manner, combining prior beliefs with the sampling observations, and (2) the estimated probabilities are transformed by a source dependent probability weighting. The first stage of the model provides a Bayesian explanation for overestimation of infrequent outcomes that is commonly found in previous studies on probability judgments, as well as a natural way to estimate the probabilities of always-observed outcomes – i.e. outcomes with observed relative frequency of one – whose certainty is not known for sure. A source dependent probability weighting in the second stage captures deviations from SEU under experienced uncertainty. The two-stage model was tested by reanalyzing data sets available from previous studies by Glöckner et al. (2016) and Erev et al. (2010).

Chapter 3 used a Bayesian method of updating as a working hypothesis, and estimated probability weighting functions parametrically. The Bayesian updating method was proved to be useful in the analysis due to its tractability, and the model was successful in disentangling beliefs from probability weighting. Nevertheless, the descriptive validity of Bayes' rule is often questioned in empirical studies. Moreover, different parametric specifications of probability weighting functions are supported on empirical, axiomatic and meta-theoretical grounds. Following up on chapter 3, the subsequent chapters present two independent studies on the aforementioned components of the DFD-DFE gap that can inform future studies on DFE and on DFD.

Chapter 4 focuses on subjective beliefs. This chapter introduces a tractable model of non-Bayesian belief updating in a signal setup where the decision maker receives binary signals from a source of uncertainty. Assuming that a decision maker is born in a hypothetical state of ignorance, the model interprets her beliefs as a posterior probability conditional on all the *perceived* signals that she has received from the source of uncertainty. The model accommodates common updating biases such as *conservatism* referring to the reluctance of extracting enough information from sampled observations, and *confirmatory bias* referring to the tendency to misread evidence contradicting prior beliefs. Accordingly, the model quantifies *conservatism* by a likelihood of missing some signals regardless of its support for the prior beliefs, and *confirmatory bias* by a likelihood of misperceiving signals contradicting prior beliefs. The model was tested in a laboratory experiment.

Chapter 5 focuses on probability weighting. It performs an experimental test of a theory of probability weighting developed by Prelec (1998). Prelec's compound-invariant family provides an appealing way to model probability weighting and is widely used in empirical studies. Prelec (1998) gives an axiomatic foundation for this function. Luce (2001) points out that Prelec's behavioral condition, *compound invariance,* is hard to test empirically, and he proposes a simpler condition, *reduction invariance*, to characterize Prelec's weighting function that is easier to test empirically. Following up on Luce's suggestion, this chapter investigates the empirical validity of this condition in a laboratory experiment.

This thesis mainly contributes to the emerging field of DFE by exploring the role of probability weighting and subjective beliefs. First, it clarifies the controversy about the DFD-DFE gap by a carefully designed laboratory experiment (chapter 2). Second, it develops a parsimonious decision model for DFE that can successfully account for previous findings on the DFD-DFE gap in the literature (chapter 3). Third, it introduces a tractable non-Bayesian updating model whose insights are expected to inform future studies of DFE (chapter 4). Lastly, it provides an empirical test of the foundations a family of probability weighting function that can be used in parametric estimations under DFE as well as under DFD (Chapter 5).

Chapter 2

Are Black Swans Really Ignored? Re-examining Decisions from Experience

with Yu Gao

2.1 Introduction

Studies of decisions from experience (henceforth, DFE) investigate decision situations in which people rely on personal experiences when facing uncertainty. Decision makers often have no access to possible choice outcomes, let alone to the corresponding probabilities. Instead, they make decisions based on the past observations in their memory. DFE better captures real life decisions than traditional 'Decisions from Description' (henceforth, DFD) where payoffs and probabilities are fully specified, which rarely happens in real life. In the usual sampling paradigm of DFE (Hertwig et al. 2004), subjects learn about unknown payoff distributions by drawing samples with replacement. With merely these cases in memory, they make their final decisions.

Since Barron & Erev (2003) and Hertwig et al. (2004), an intriguing discrepancy between the two decision paradigms, which is called the DFD-DFE gap, has received plenty of attention. The common view in the DFE literature is that people make decisions from experience as if they are underweighting rare and extreme events, so called "black swans", which are often overweighted under the DFD paradigm (for a review, see Hertwig & Erev, 2009). This pattern implies a reversal of the inverse S-shaped probability weighting that has been documented by many empirical studies under DFD (Abdellaoui, 2000; Bleichrodt & Pinto, 2000; Bruhin, Fehr-Duda, & Epper, 2010; Booij, van Praag, & van de Kuilen, 2010; Fehr-Duda, De Gennaro, & Schubert, 2006; Gonzalez & Wu, 1999; Tversky & Kahneman, 1992)[1].

The DFE literature has suggested that the DFD-DFE gap is a robust empirical phenomenon. Although the under-sampling of rare events due to reliance on small samples (sampling error) partly explains the early findings of the gap (Fox & Hadar, 2006; Hadar & Fox, 2009; Hertwig et al., 2004), later studies have shown that it does not provide a

[1] See Wakker (2010), section 7.1 for a survey.

complete account (Barron & Ursino, 2013; Camilleri & Newell, 2009; Hau, Pleskac, Kiefer, & Hertwig, 2008; Hau, Pleskac, & Hertwig, 2010; Ungemach, Chater, & Stewart, 2009). Importantly, unlike risk with known probabilities in DFD, the ambiguity in DFE stemming from unknown outcome probabilities – and even from unknown set of possible outcomes – is another cause of the gap (Abdellaoui, L'Haridon, & Paraschiv, 2011; Glöckner, Hilbig, Henninger, & Fiedler, 2016; Kemel & Travers, 2016).

Despite the robustness of the DFD-DFE gap, whether it actually amounts to a reversal of the inverse S-shaped probability weighting is still unclear in the literature. In addition to the sampling error and ambiguity, there are two extra confounds that render the inferences about probability weighting problematic in DFE studies. The first confound concerns an aggregation problem when there is a lack of control over the sampling experience of subjects. Because of the random nature of the sampling process – where the sampling is made with replacement and subjects decide when to stop sampling – each subject relies on her own distinct subjective experiences. Importantly, this heterogeneity in experience at the individual level causes potential distortions at the aggregate level due to averaging artifacts (Estes, 1956; Estes, 2002; Sidman, 1952). The problem of aggregation is explained in section 2.3.3.

The second confound concerns the role of utilities. Early studies in the DFE literature argue about the underweighting of rare outcomes in an "as-if" sense. Specifically, the underweighting is typically inferred from a preference for sure gains over expected-value-equivalent lotteries involving unlikely gains (for example, a preference for a sure $1 over a lottery with 10% chance of winning $10 and $0 otherwise). However, the absolute weighting of probabilities stays unclear as the aversion to unlikely gains may as well be due to concave utility (possibly coupled with an unbiased probability weighting) as it may be due to an underweighting of unlikely events. Later studies controlled for utilities by estimating them together with probability weighting functions using a parametric approach. Nevertheless, one concern about simultaneous parametric estimations is the potential interactions between the parameters of utility and probability weighting functions (Gonzalez & Wu, 1999, p. 152; Scheibehenne & Pachur, 2015, pp. 403-404; Stott, 2006, p. 112; Zeisberger, Vrecko, & Langer, 2012).

This paper provides a measurement of probability weighting under DFE by resolving the aforementioned problems, and thus improving validity. First, we used Barron & Ursino's (2013) adjustment of the sampling paradigm to obtain a control over the sampling experience of each individual subject. Specifically, all of our subjects were required to carry out complete sampling from finite outcome distributions without replacement. Hence, they acquired the sampling information that matched with the objective probabilities without any sampling error. Second, this way we also avoid the confounding effects of unknown probability attitudes, well documented in the literature (Ellsberg, 1961; Trautmann & Van de Kuilen, 2015). Third, we avoided the aggregation problem as explained in more detail later.

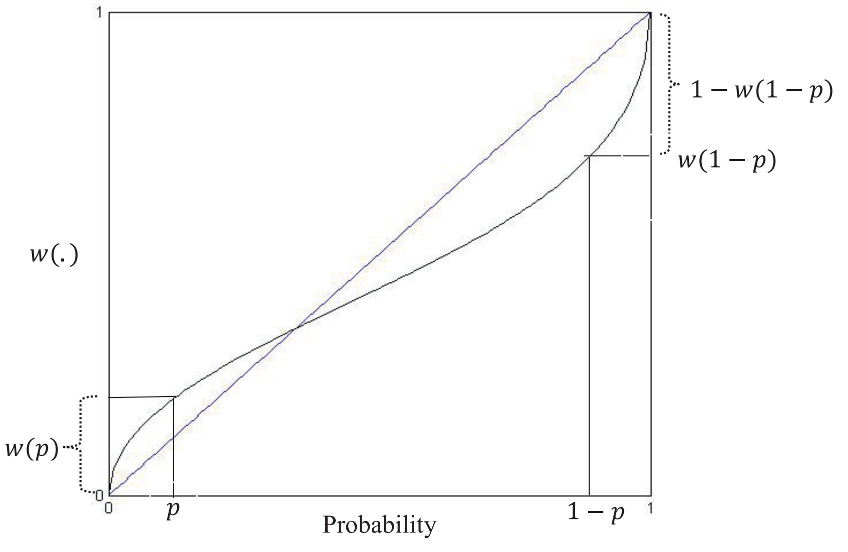
Fourth, we measured probability weighting by a rigorous two-stage methodology (Abdellaoui, 2000; Bleichrodt & Pinto, 2000; Etchart-Vincent N. , 2004; Etchart-Vincent N. , 2009; Qiu & Steiger, 2011). In particular, this controlled for the utility curvature in the first stage. Thus, each choice in the second stage directly indicated overweighting or underweighting of probabilities. The experimental setup enabled us to identify the direction and the magnitude of the deviations from expected utility (henceforth, EU), without relying on any parametric assumptions about probability weighting. Parametric estimations were implemented as a supplement of our nonparametric measures to test robustness and smooth out response errors.

2.2 Deviations from EU due to probability weighting

We restrict our attention to probability-contingent binary prospects in the gain domain. A binary prospect of winning α with probability p and β otherwise is denoted $\alpha_p\beta$. Under rank dependent utility (henceforth RDU), for $\alpha \geqslant \beta \geqslant 0, \alpha_p\beta$ is evaluated by $w(p)U(\alpha) + \big(1 - w(p)\big)U(\beta)$ where U is the utility function and w the probability weighting function. Throughout, we assume binary RDU. Most other non-EU theories, in particular both versions of Prospect Theory for gains (henceforth PT, Kahneman & Tversky, 1979; Tversky & Kahneman, 1992), and Gul's (1991) Disappointment Aversion Theory, agree with the binary RDU in the evaluation of binary prospects (Observation 7.11.1 in Wakker, 2010, p. 231). Hence, our analysis applies to all these theories.

RDU deviates from EU when $w(.)$ is not the identity. Thus, the risk attitude of a decision maker depends not only on the utility curvature as in EU but also on probability weighting. The common finding with the DFD paradigm is an inverse S-shaped (first concave and overweighting, then convex and underweighting) probability weighting function (Figure 2.1).[2] The steepness of the probability weighting function at both end points implies that the rare and extreme outcomes in general receive too much decision weight. When a rare outcome with probability p is desirable, its impact given by $w(p)$ is overweighted because of the overweighting of small probabilities ($w(p) > p$). This increases the attractiveness of the prospect, and leads to risk seeking. Similarly, when a rare outcome with probability p is unfavorable, its impact given by $1 - w(1-p)$ is overweighted because of the underweighting of large probabilities ($w(1-p) < 1 - p$). This decreases the attractiveness of the prospect, and leads to risk aversion.

Figure 2.1 Inverse S-shaped probability weighting function



The pattern of inverse S-shaped probability weighting is commonly interpreted as the reflection of both cognitive and motivational deviations from EU (Gonzalez & Wu, 1999). On the one hand, the simultaneous overweighting and underweighting of extreme probabilities implies insufficient sensitivity to intermediate probabilities. This effect is called likelihood insensitivity, and points to cognitive limitations in discriminating different levels of uncertainty. On the other hand, underweighting of moderate probabilities

[2] For evidence against inverse S, see Qiu & Steiger (2011), van de Kuilen & Wakker (2011) and Krawczyk (2015).

(such as, $w(0.5) < 0.5$) suggests a pessimistic attitude towards risk in the major part of the probability domain. This effect points to motivational deviations from EU.

2.3 The DFD-DFE gap

Hertwig & Erev (2009) considers three DFE paradigms: partial feedback, full feedback, and sampling paradigms. The essential feature shared by all three DFE paradigms is that subjects learn about unknown payoff structures by solely relying on their experiences. In the partial feedback paradigm, subjects make repeated choices and receive feedback about the realized outcomes (Barron & Erev, 2003). In the full feedback paradigm, subjects also learn about the forgone outcomes from the unchosen options (Yechiam & Busemeyer, 2006). Differently, the sampling paradigm involves a single – rather than repeated – choice preceded by a purely exploratory and inconsequential sampling period in which subjects draw outcomes from unknown payoff distributions with replacement, usually as many times as they wish (Hertwig et al, 2004; Weber , Shafir, & Blais, 2004).

All three paradigms lead to similar behavioral patterns with an apparent underweighting of rare and extreme outcomes, which contradicts the common empirical findings from DFD. Although the empirical findings with all three paradigms are alike, the two feedback paradigms are inherently different from the sampling paradigm (for an empirical comparison of three DFE paradigms, see Camilleri & Newell, 2011, also see the theoretical discussion of Gonzalez & Dutt, 2011). In particular, repeated choices in the two feedback paradigms, as opposed to single decisions in the sampling paradigm, induce long-run payoff considerations due to accumulating income (Wullf, Hills, & Hertwig, 2015). This predicts more expected value maximization in repeated choices by the law of large numbers (Keren & Wagenaar, 1987; Lopes, 1982; Tversky & Bar-Hillel, 1983). Furthermore, distinct psychological factors, such as reinforcement learning, and the hot stove effect[3], also play a role in repeated decisions with feedback (March, 1996; Denrell & March, 2001). Erev & Barron (2005) reviews the effects that lead to deviations from expected value maximization in repeated choice paradigms. The sampling paradigm, on the other hand, is more comparable with the DFD paradigm as both involve single decisions. Therefore, the intriguing gap between the sampling paradigm and DFD has received most

[3] The hot stove effect was first introduced by Mark Twain based on his observation that if a cat jumped on a hot stove, then she would never jump on a hot stove again. However, the cat would never jump even on a cold stove.

attention in the DFE literature. The current paper also focuses on the sampling paradigm of DFE.

2.3.1 The information asymmetry account and the sampling error

The main premise of the DFD-DFE gap is that the way in which the information about uncertain prospects is acquired matters. In other words, experience matters (Hau et al., 2008). Fox & Hadar (2006) and Hadar & Fox (2009) argue that there is an important caveat associated with this premise. DFE and DFD differ from each other not only in terms of the way that the information is acquired but also in terms of the information available to subjects. Indeed, whereas the objective probabilities and outcomes are known in DFD, they remain partially unknown in DFE. This means that subjects in DFE have to rely on their own subjective probability judgments based on the sampling information they acquire. Importantly, subjective probabilities are prone to diverge from objective probabilities due to potential distortions either in the sampling process or in subjective probability judgments. This generates an information asymmetry between DFE and DFD. Fox & Hadar (2006) indicates that the underweighting of rare outcomes observed by Hertwig et al. (2004) is almost entirely caused by the sampling error as subjects often under-observe, or even never observe the rare outcomes due to reliance on small samples. On the other hand, judgment error and underestimation of rare outcomes are not found to be significant sources of the gap.

Later studies test this information asymmetry account of the DFD-DFE gap by reducing or completely eliminating the sampling error. Several papers demonstrated that the gap is actually persistent when the subjects are obliged to draw large or even perfectly accurate samples from underlying probability distributions (Barron & Ursino, 2013; Camilleri & Newell, 2009; Hau et al., 2008; Hau et al., 2010; Ungemach et al., 2009). Moreover, subjective probability judgments are usually found well calibrated although their correlation with observed relative frequencies is imperfect (but see also Barron & Yechiam, 2009). These findings suggest that the DFD-DFE gap is not just information asymmetry, but indeed a robust psychological phenomenon.

2.3.2 DFE and DFD: Two different sources of uncertainty

Although drawing large or representative samples solve the problem of systematic sampling error, the uncertainty about the outcome probabilities as well as the set of possible outcomes remains. This residual uncertainty makes DFE a case of ambiguity whereas DFD is a case of risk. Several studies show that the gap is reduced or even reversed by manipulating the degree of ambiguity in DFE. In addition to information provision regarding the certainty or possibility of outcomes, Glöckner et al. (2016) also points out the impact of the type of problems used in the experiments, which may lead to context dependent subjective beliefs.

In a design that is intermediate between DFE and DFD, Abdellaoui et al. (2011) and Kemel & Travers (2016) find inverse-S pattern in DFE with more pronounced pessimism than in DFD. This result reflects ambiguity aversion. Kellen, Pachur, & Hertwig (2016) and Glöckner et al. (2016) find even more pronounced likelihood insensitivity in DFE. These findings are consistent with the previous ambiguity literature (Abdellaoui, Baillon, Placido, & Wakker, 2011; Fox & Tversky, 1998; Tversky & Fox, 1995; Wakker, 2004).

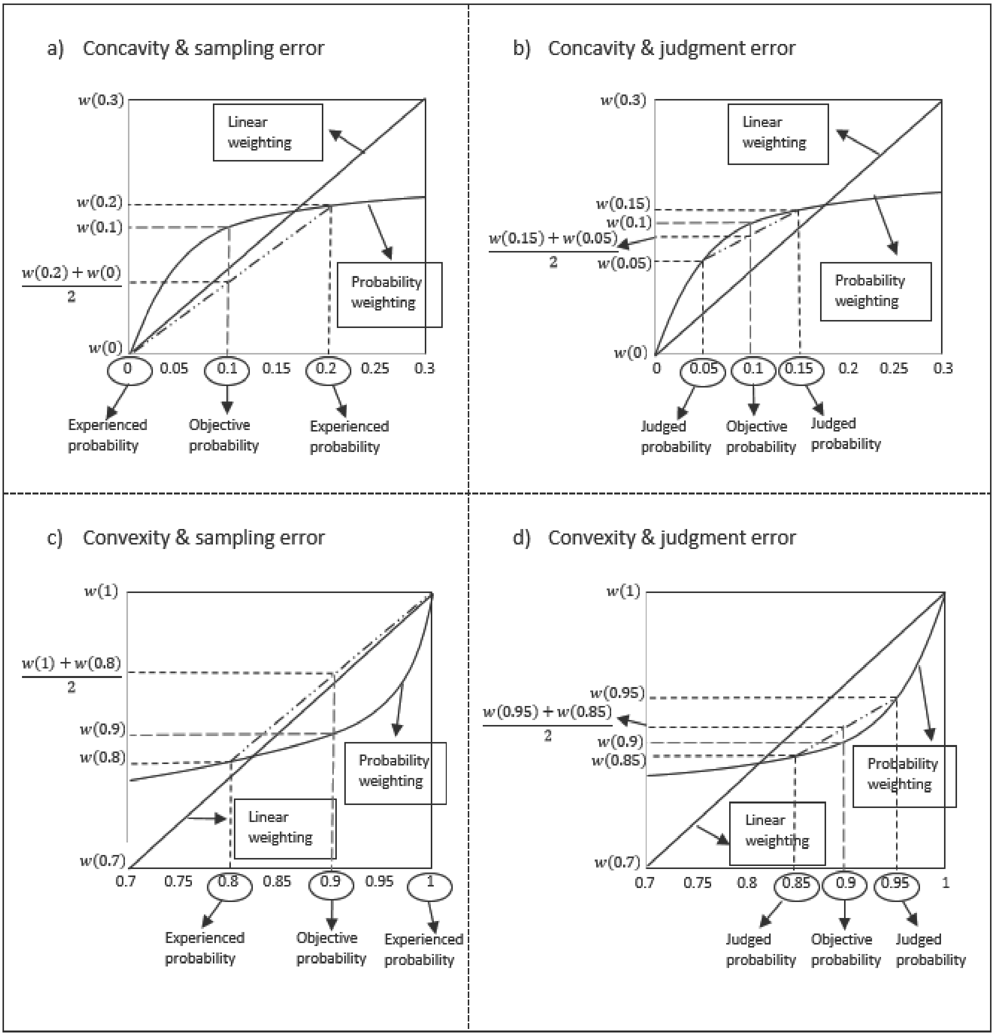
2.3.3 Problem of aggregation in the sampling paradigm

As explained before, experienced probabilities differ from objective probabilities either due to sampling error or due to judgment errors. As a result, each subject makes choices based on her own subjectively experienced probabilities. Notably, as the aggregation of such individual choices amounts to taking the average of the weightings – rather than the weighting of the average – of experienced probabilities, the concave-convex curvature of the inverse S-shaped probability weighting function may lead to an erroneous DFD-DFE gap.

To illustrate, assume that all subjects in DFE and DFD have the same probability weighting function depicted in figure 2.2a, which is concave and overweight 10% probability of a rare and favorable outcome. For the sake of the example, also assume that each subject in DFE draws only 5 times, in which half of the subjects never observe the rare outcome, and the other half observe it once. Therefore, assuming that the subjects do not commit a judgment error, the experienced probabilities will be either 0% or 20%. In this case, aggregating choices over all subjects' amounts to averaging the weightings of 0% and 20% rather than weighting the average 10%. This makes the aggregate choice appear as if 10% is underweighted due to concavity whereas in reality it is overweighted (see figure 2.2a).

The same effect, although probably smaller in size, also applies when there is no sampling error but only judgment error. Figure 2.2b illustrates the case where the subjects in DFE accurately observe 10% probability, however, half of them underestimate it as 5% whereas the other half overestimate it as 15%. As a result, the aggregate choice appears as if 10% is weighted less in DFE than in DFD (see figure 2.2b).

Figure 2.2 Distortions due to aggregation



By the dual effect, convex probability weighting for large probabilities moves aggregate choices in the direction of overweighting (see figures 2.2c and 2.2d). Together with the concavity for small probabilities, this implies a reversed or attenuated inverse S at the aggregate level, which is what the DFD-DFE gap also suggests. This theoretical conjecture is indeed indirectly supported by the findings of Rakow, Demes, & Newell (2008). In their yoked design, each subject in the DFE treatment is matched with a subject in the DFD treatment who receives the same sampling information in description format. Thus, equating the heterogeneity of the sampling information across the two treatments, they observe that the DFD-DFE gap is almost completely eliminated (also see the discussion of Hau et al. 2010 on the amplification effect in yoked design).

2.3.4 Underweighting or not?

Along with the aforementioned issues, the controversy about the DFD-DFE gap concerns whether it can actually give rise to underweighting of rare outcomes. Early studies of DFE infer underweighting from aggregate patterns of risk seeking and/or risk aversion. Rakow & Newell (2010, p. 6) points out that the gap often amounts only to a discrepancy in risk attitudes (e.g. different degrees of risk seeking for small probability gains), suggesting a less pronounced overweighting in DFE compared to DFD, rather than an absolute underweighting. Moreover, even a reversal in risk attitudes (e.g. risk aversion for small probability gains in DFE as opposed to risk seeking in DFD) may not be sufficient to conclude about the absolute underweighting of rare outcomes under DFE as a concave utility along with an unbiased weighting might also lead to risk aversion.

Later studies report quantitative estimations of probability weighting under DFE, also by controlling for the role of utilities. However, the present evidence on the shape of probability weighting functions is mixed. Hau et al. (2008) and Ungemach et al. (2009) document linear weighting and underweighting respectively, based on the same set of problems used by Hertwig et al. (2004). Among those studies that used larger problem sets, Abdellaoui et al. (2011b), Kemel & Travers (2016), and Cubitt, Kopsacheilis, & Starmer (2016) report less pronounced overweighting whereas Barron & Ursino (2013) and Frey, Mata, & Hertwig (2015) report underweighting. Other recent studies by Glöckner et al. (2016) and Kellen et al. (2016) reports even more pronounced overweighting under DFE. Differences in methodologies and the use of different choice tasks are possible sources of the discrepancy (Glöckner et al. 2016). For a further discussion of these discrepant results, see a recent meta-analysis by Wulff, Canseco, & Hertwig (2016).

Our experiment aims to clarify the controversy by resolving the aforementioned four confounds. Different from previous studies, our adjustment of the sampling paradigm turns the DFD-DFE comparison into a pure comparison of two cases of risk that differ only in terms of information acquisition, being experience or description.

2.4 Method

Our experimental procedure consists of two stages. In the first stage, the utility function of each subject is elicited using the trade-off (TO) method of Wakker and Deneffe (1996). The TO method is a well-established method that has been commonly used in studies investigating probability weighting (Abdellaoui, 2000; Abdellaoui, Vossmann, & Weber, 2005; Bleichrodt & Pinto, 2000; Etchart-Vincent N. , 2009; Etchart-Vincent N. , 2004; Qiu & Steiger, 2011). The method entails the elicitation of a standard sequence of outcomes that are equally spaced in utility units. The elicitation procedure consists of a series of adaptive indifference relations. For two fixed gauge outcomes G and g, and a selected starting outcome x_o with $x_0 > G > g$, $x_1 > x_0$ is elicited such that the subject is indifferent between prospects ${x_1}_p g$ and ${x_0}_p G$. Then, x_1 is used as an input to elicit $x_2 > x_1$ such that the subject is indifferent between ${x_2}_p g$ and ${x_1}_p G$. This procedure is repeated n times in order to obtain the standard sequence $(x_0, \ldots, x_n)$ with indifferences ${x_{i+1}}_p g \sim {x_i}_p G$ for $0 \leq i \leq n-1$. Under RDU, these indifferences result in $U(x_1) - U(x_0) = U(x_2) - U(x_1) = \cdots = U(x_{n-1}) - U(x_n)$ (for the derivation, see Appendix 2.1). A remarkable feature of the TO method is that it elicits these equalities irrespective of what the probability weighting is. Therefore, it is robust against most distortions due to non-expected utility maximization.

Once the standard sequence of outcomes has been obtained, we obtain the utility function of each individual by parametrically estimating the power specification $U(x) = x^{\alpha}$ with $\alpha > 0$ after scaling of $x_i s$ as $x_i = \frac{x_i - x_0}{x_n - x_0}$. We use parametric estimation in order to smooth out errors, and better capture the utility curvature. The parameter α is calculated using an ordinary least squares regression without intercept, $\log\big(U(x)\big) = \alpha \log(x) + \varepsilon$ where $\varepsilon \sim N(0, \sigma^2)$.

In the second stage of our procedure, we measure probability weighting using several binary choice questions. The questions are constructed based on the subject-specific outcome sequences obtained from the first stage. Subjects choose between a risky prospect ${x_k}_q x_j$ and a sure outcome s_q, where x_k and x_j are two distinct elements of the elicited outcome sequence with $x_k > x_j$, and s_q is equal to the certainty equivalent of ${x_k}_q x_j$ under EU based on the power utility function estimated in the first stage:

$$s_q = U^{-1}\left[qU(x_k) + (1-q)U\left(x_j\right)\right]. \tag{2.1}$$

That is, s_q would be equivalent to ${x_k}_q x_j$ if the subject with the given utility did not weigh probabilities. Hence by construction, the following logical equivalences hold for given preference relations under RDU.

$${x_k}_q x_j \prec s_q \;\Leftrightarrow w(q) < q \; (underweighting\;) \tag{2.2}$$

$${x_k}_q x_j \sim s_q \;\Leftrightarrow w(q) = q \; (EU) \tag{2.3}$$

$${x_k}_q x_j \succ s_q \;\Leftrightarrow w(q) > q \; (overweighting) \tag{2.4}$$

Because we do not allow indifference in our experiment, each individual choice will reveal either overweighting or underweighting of probability q. Our method makes the deviations from EU observable at the aggregate level. For instance, an overweighting of q can be detected when the majority of subjects choose the risky ${x_k}_q x_j$ as in (4).

Barron & Ursino (2013) also investigates the DFD-DFE gap under risk (their experiment 1) similar to our study by using a different two-stage experimental procedure. Their procedure replicates the well-known DFD-DFE gap. However, it does not make inferences about the actual over- or under- weighting of rare outcomes under DFE and DFD[4].

[4] Their first stage obtains an indifference relation under DFD which implies $w(1-q) * U(X) = w(q) * U(\$40)$, where the probability q is either 0.1 or 0.2, depending on the treatment, and X is elicited. Their second stage looks at deviations from this indifference under DFE and DFD. Their findings indicate deviations only under DFE, suggesting less weighting of q and/or more weighting of $(1-q)$ under DFE, i.e. $w(1-q) * U(X) > w(q) * U(\$40)$, consistent with the DFD-DFE gap.

2.5 Experimental design

Subjects and Incentives

The experiment was performed at the ESE-EconLab at Erasmus University in 5 group sessions. Subjects were 89 Erasmus University students from various academic disciplines (average age 23 years, 40 female). All subjects were recruited from the pool of subjects who had never participated in any economic experiment in our lab before, to avoid experienced subjects in TO method. We paid each subject a €5 participation fee. In addition, at the end of each session, we randomly selected two subjects who could play out one of their randomly drawn choices for real. The ten subjects who played for real received €60.70 on average. Over the whole experiment, the average payment per subject was €12.37.

Procedure

The experiment was run on computers. Subjects were separated by wooden panels to minimize interaction. To prevent the impact of variations in memory limitations, all subjects were provided with paper and pen in case they wished to take notes. Before they started with the main parts of the experiment, they read the general instructions with detailed information about the payment procedure, the user interface, and the type of questions they would face. The subjects could ask questions at any time during the experiment. The experiment consisted of two successive stages without a break in between. Each stage started with its corresponding instructions, and several training questions to familiarize subjects with the stimuli. Each session took 45 minutes on average, including the payment phase after the experiment.

Stimuli

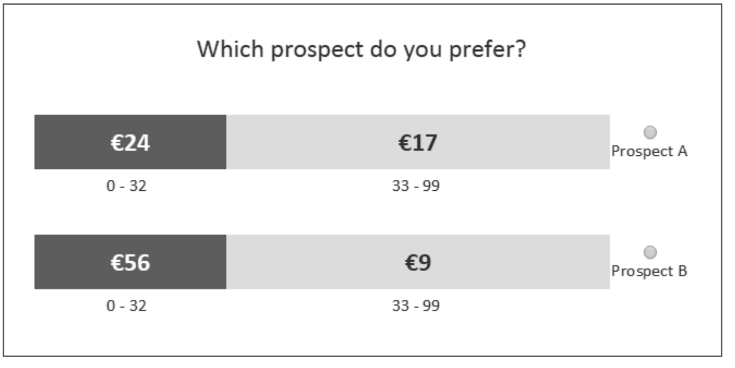
Stage 1: measuring utility. In the first stage of the experiment, a standard sequence of outcomes was elicited using the TO method. We measured $x_1, x_2, x_3, x_4, and\ x_5$ from the following five indifferences, with $p = 0.33, G = 17, g = 9$, and $x_0 = 24$:

$$\boldsymbol{24_p G \sim x_{1_p} g},\ x_{1_p} G \sim x_{2_p} g,\ \boldsymbol{x_{2_p} G \sim x_{3_p} g},\ x_{3_p} G \sim x_{4_p} g,\ \boldsymbol{x_{4_p} G \sim x_{5_p} g}.$$

Indifferences were obtained by a bisection method requiring 7 iterations for each x_i. In addition, the last iteration of one randomly chosen x_i was repeated at the end of stage 1, in order to test the reliability of the indifferences. Hence, subjects answered a total of 36 questions in this part. The bisection iteration procedure is described in Appendix 2.2. The prospects were presented on screen as in Figure 2.3.

In this part, risk was generated by two ten-faced dice each generating one digit of a random number from 00 to 99. The outcome of prospects depended on the result of two dice physically rolled by subjects in case the question was played for real at the end of the experiment.

Figure 2.3 Choice situation in the TO part



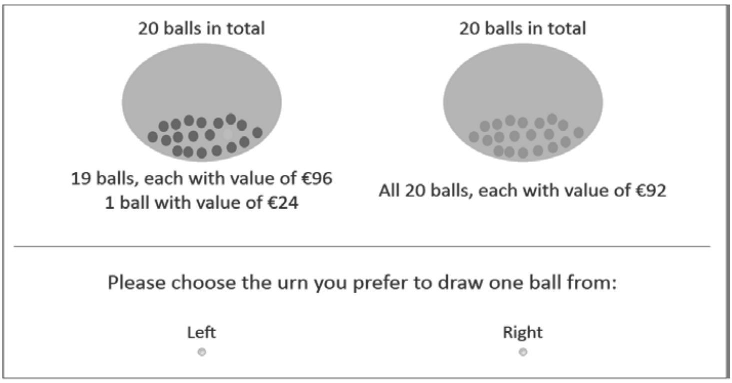
Stage 2: DFD and DFE. Before the start of the second part, each subject was randomly assigned to one of the two treatments: DFE or DFD. Subjects in both treatments answered 7 subject-specific binary choice questions. Each question entailed a choice between a risky prospect ${x_5}_q x_1$ and the safe prospect s_q as described in the method section. Note that both x_1 and x_5 were endogenously determined, and varied between subjects.[5] Values of s_q were always rounded to the nearest integer. The seven probabilities used for the investigation of probability weighting were $0.05, 0.10, 0{,}20, 0.50, 0.80, 0.90$ and 0.95. Within each treatment, the orders of the seven questions were counterbalanced. The position of the risky prospect and the safe prospect were also randomized in each question.

Prospects were represented by Ellsberg-type urns containing 20 balls with different monetary values attached to them. This means that all the aforementioned probabilities

[5] We used the elicited x_1 as the minimum outcome of the risky prospects to avoid problems related to the extreme behavior of power utility near its origin (Wakker 2008), i.e. x_0 in our design. In particular, for $\alpha < 1$, the slope of the power utility converges to infinity as x tends to the origin. This implies extreme risk aversion near the origin. Similarly, $\alpha > 1$ implies extreme risk seeking near the origin.

were fractions of 20; i.e. 5% is 1 out of 20, 10% is 2 out of 20, etc. The two treatments differed from each other in terms of how the contents of the urns were learnt. In the DFD treatment, the contents of the urns were explicitly described to the subject. Figure 2.4 shows a screen shot of a choice situation for DFD.

Figure 2.4 Choice situation in DFD

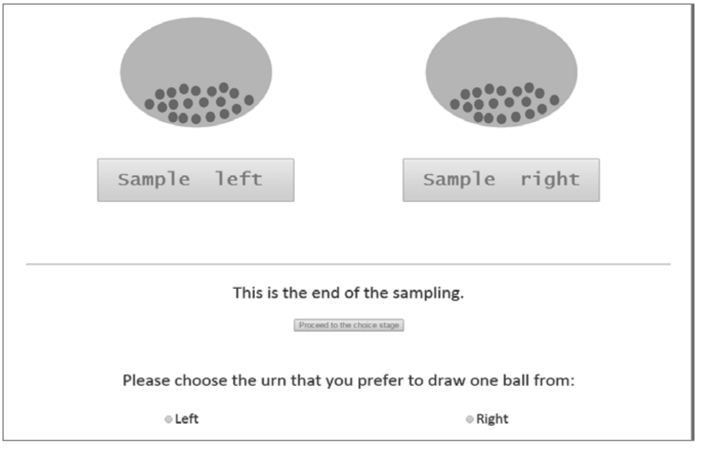


Subjects in the DFE treatment were initially given no information about the contents of the urns except for the total number of balls. They could only learn about the outcome compositions of the urns by sampling each and every ball one-by-one without replacement, and observing the monetary values attached. Figure 2.5 shows a screen shot of the sampling phase in the DFE treatment. Subjects sampled balls from urns by clicking "Sample left" or "Sample right" on the screen. Each time, the monetary outcome attached to the ball sampled was shown to the subject for 1.5 seconds, and then disappeared. Subject could sample at their own speed, in whichever order they preferred, and switch as many times as they wanted, but they could only proceed to the choice stage after sampling all the balls in both urns.

Figure 2.6 shows the screen shot of the choice stage in DFE. In case a question in this part was drawn for the payment at the end of the experiment, the experimenters physically created the relevant urn seen on the screen by filling an opaque urn with 20 ping-pong balls painted to dark blue or light blue, each associated with the payoffs in question (see Figure 2.4). Then, the subject drew a ball from the urn, which determined her payoffs.

Figure 2.5 Sampling stage in DFE

Figure 2.6 Choice stage in DFE



Subjects in the DFD treatment faced 21 extra questions following the main set of 7 questions to equalize the length of the two treatments. These extra questions were for another research project.

2.6 Results

Reliability and Consistency of Utility Elicitation

In the TO part, each subject repeated one choice faced in one of the five elicitations. The repeated choice was randomly selected among the last steps of the iterations. Because the subjects were very close to indifference at the last step, this was the strongest test of consistency. Subjects made the same choice in 70.8% of the cases. Reversal rates up to one third are common in the literature (Stott, 2006; Wakker, Erev, & Weber, 1994). Especially, if the closeness to indifference is taken into account, our reversal rates are satisfactory. Among the reversed cases, repeated indifferences were higher than the original indifference values in 42.3% of the times, which did not indicate any systematic pattern (p=0.56, two-sided binomial). Overall, repeated indifference values did not differ from original elicitations (p=0.44, Wilcoxon sign-rank).

In our data, one subject reached the possible lower bound of x_i's in all 5 cases. Consequently, her standard sequence was not well spaced enough for the estimations of s_q with Equation 2.1.[6] We excluded this subject from the following analysis. The analysis with this subject included does not alter our conclusions.

Utility Functions

Table 2.1 gives the descriptive statistics for the elicited outcome sequence. The parameter α of the power utility $\mathrm{u(x)} = \mathrm{x}^{\alpha}$ was estimated at the individual level by ordinary least squares regression. The average R^2 over all individual utility estimations was 0.985 which indicated that our estimations fit the data well.

The summary statistics for the mean and median α are reported in the last row of Table 2.1. The aggregate data did not deviate from linearity (p=0.92, Wilcoxon sign-rank). Although the mean alpha suggested slight convexity, this was due to the outliers in our data. Three subjects exhibited extreme convexity with $\alpha > 2$, and the Skewness/Kurtosis test rejected the normality of the distribution of α's (p=0.00). Utilities did not differ across the two treatments (p=0.84, Wilcoxon rank-sum) as the first stage of the experiment was the same for all subjects.

Our data suggested slightly more evidence for concavity at the individual level. Based on the α parameters that were significantly different than 1 at 5% significance level, 30 subjects exhibited concavity ($\alpha < 1$), and 23 subjects exhibited convexity ($\alpha > 1$). The proportions of concave and convex utilities did not differ from each other (p=0.41, two-sided binomial).

---

[6] She got ($x_5 - x_1 = 8$). Therefore, the resulted estimations, $s_{0.05} = x_1$ and $s_{0.95} = x_5$, made the preference for $x_{5_{0.05}}x_1$ over $s_{0.05}$ and the preference for $s_{0.95}$ over $x_{5_{0.95}}x_1$ trivial because of the domination of the safe or the risky prospect.

Table 2.1 Descriptive statistics of the elicited outcome sequence (N=88)

	Mean	S. Dev	Min	Median	Max
x_0	24.00	0.00	24.00	24.00	24.00
x_1	60.36	23.48	30.00	58.00	118.00
x_2	90.36	42.58	36.00	80.00	212.00
x_3	125.23	65.89	46.00	102.00	306.00
x_4	164.18	91.13	52.00	134.00	400.00
x_5	204.14	116.25	58.00	160.00	494.00
α	*1.05*	*0.36*	*0.41*	*0.99*	*2.65*

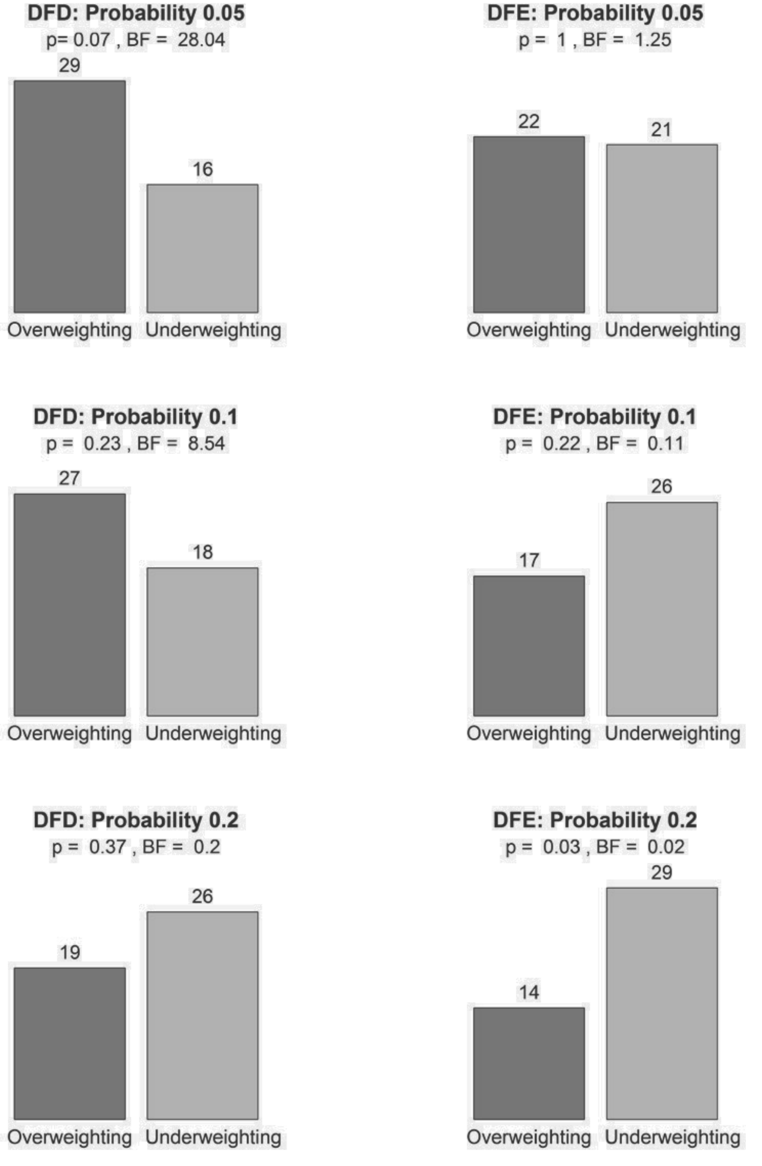
Probability Weighting: DFE vs. DFD

Aggregate data. In this section, we report the aggregate choices in the direction of overweighting and underweighting according to (2) and (4) in the Method section. The proportions of overweighting and underweighting of small and large probabilities are given in Figures 2.7 and 2.8 respectively.

The aggregate choices replicated the common DFD-DFE gap at the extreme probabilities. Overall, the DFD-DFE gap indicated significantly less overweighting of rare outcomes based on a repeated measures logistic regression ($z = -2.15$, p=0.031).[7] Based on individual hypothesis tests, the gap was significant at 0.95 (p=0.02, χ^2); and marginally significant at 0.10, and 0.90 (p=0.06, and p=0.07 respectively, χ^2). The gap at probability 0.05 was not significant (p=0.20, χ^2), although the trend suggested reduced overweighting in DFE. There was also no apparent DFD-DFE gap in the middle range, $0.20 \leq q \leq 0.80$ (p=0.35, p=0.92, and p=0.37 for $q = 0.20, 0.50$, and 0.80 respectively, χ^2).

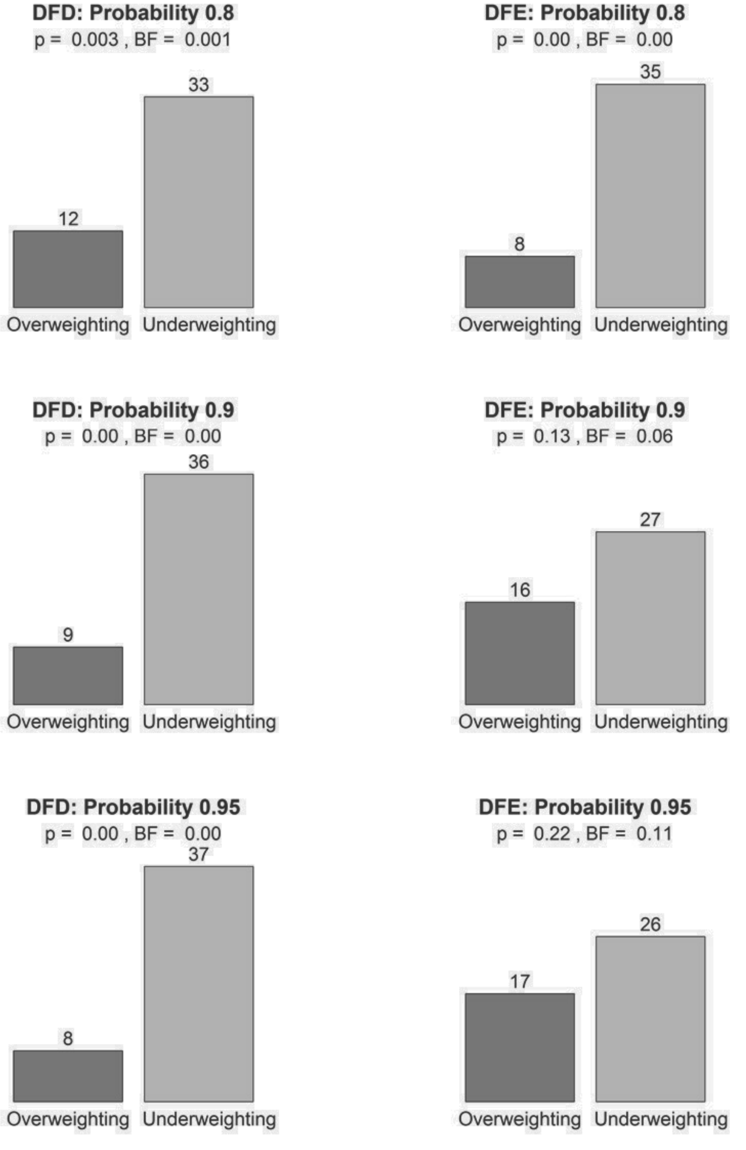
[7] Note that the overweighting of good rare outcomes amounts to overweighting of 0.05, 0.10, and 0.20 whereas the overweighting of bad rare outcomes amounts to underweighting of 0.80, 0.90, and 0.95.

Figure 2.7 Weighting of Small Probabilities



Notes: p-values are for the two-sided binomial tests. Bayes factors (BF) indicate evidence for the null hypothesis that the probability is overweighted. Higher BF indicates higher support for overweighting of the given probability. The numbers above bars are the number of subjects who revealed the correspondent probability weighting patterns in choices.

Figure 2.8 Weighting of Large Probabilities



Notes: p-values are for the two-sided binomial tests. Bayes factors (BF) indicate evidence for the null hypothesis that the probability is overweighted. Higher BF indicates higher support for overweighting of the given probability. The numbers above bars are the number of subjects who revealed the correspondent probability weighting patterns in choices.

In what follows, we focus on absolute overweighting and underweighting of probabilities under DFD and DFE. We first test the deviations from unbiased weighting in either directions with the classical two-sided binomial tests for proportions. In addition, to interpret the relative evidence for overweighting and underweighting, we report Bayes factors for the null hypothesis of overweighting against the alternative hypothesis of underweighting. Bayes factors state the relative evidence for the null hypothesis. For instance, a Bayes factor of 10 indicates that overweighting is 10 times more likely than underweighting for the given probability. Following Jeffreys (1961), a Bayes factor between 3 and 10 is interpreted as "some evidence", a Bayes factor between 10 and 30 is interpreted as "strong evidence", and a Bayes factor larger than 30 is interpreted as "very strong evidence" for the null of overweighting. Similarly, Bayes factors between 0.1 and 0.33, between 0.03 and 0.1, and less than 0.03 are interpreted respectively as "some evidence", "strong evidence", and "very strong evidence" for the alternative hypothesis of underweighting.[8]

As shown in Figure 2.7, for small probabilities, we found a marginally significant deviation from unbiased weighting at 0.05 (p=0.07) under DFD. Interpreting from Bayes factors, there was strong evidence of overweighting 0.05 (BF=28.04), some evidence of overweighting 0.1 (BF=8.54) and some evidence of underweighting 0.2 (BF=0.2). Under DFE, we only found a significantly biased weighting at 0.2 (p=0.03). Interpreting from Bayes factors, there was very strong evidence of underweighting 0.2 (BF=0.02) and some evidence of underweighting 0.1 (BF=0.11). There was no evidence for the underweighting or the overweighting of 0.05 (BF=1.25).

For large probabilities as shown in Figure 2.8, under DFD, we found significant biases in weighting of probabilities 0.8, 0.9 and 0.95 (p=0.00 for all). The Bayes factors indicated very strong evidence for underweighting of 0.8, 0.9 and 0.95 (BF=0.00 for all). Under DFE, we found significant bias only at 0.8 (p=0.00). The Bayes factors suggested very strong evidence of underweighting of 0.8 (BF=0.00), strong evidence of underweighting 0.9 (BF=0.06) and some evidence of underweighting 0.95 (BF=0.11).

Lastly, we examined the weighting of the moderate 0.5 probability. 38 out of 45 subjects in the DFD treatment and 36 out of 43 subjects in the DFE treatment underweighted 0.5. Hence, the deviations from unbiased weighting was highly significant

[8] Bayes factors are computed with the package BayesFactor in R (Morey, Rouder, Jamil, & R Core Team, 2015)

at 0.5 in both treatments (p=0.00 for both treatments, two-sided binomial tests). The Bayes factors also indicated very strong evidence in favor of underweighting at 0.5 (BF<0.03 for both treatments).

To summarize, while replicating the common inverse-S pattern under DFD, our aggregate data did not provide evidence for a reversal of inverse-S pattern under DFE. In particular, we did not observe significant deviations from unbiased weighting at extreme probabilities 0.05, 0.1, 0.9 and 0.95 under DFE. Notably, there was no convincing evidence for the underweighting of small probabilities 0.05 and 0.1, and there was more evidence for underweighting than overweighting at large probabilities.

Individual data. Next, we examine the shape of probability weighting functions at the individual level. We classify each subject's probability weighting function as inverse S-shaped, S-shaped, pessimistic or optimistic based on the number of over – and under – weightings at three small and three large probabilities examined in Figures 2.7 and 2.8. Specifically, a probability weighting function is inverse S-shaped if it overweights at least two out of three small probabilities and underweights at least two out of three large probabilities, at the same time. An S-shaped probability weighting function is implied by the opposite pattern. Similarly, a pessimistic probability weighting function underweights at least two small and two large probabilities at the same time, and the opposite pattern implies an optimistic probability weighting function.

Table 2.2 Type of Probability Weighting Functions

	Inverse S-shaped	**S-Shaped**	**Pessimistic**	**Optimistic**
DFD	51% (23)	9% (4)	36% (16)	4% (2)
DFE	42% (18)	23% (10)	33% (14)	2% (1)
Gap	9% (p=0.40)	-14% (p=0.08)	3% (p=0.82)	2% (p=1)

Notes: The number of probability weighting functions is given in the parenthesis. p-values are for the (two-sided) Fisher's exact test.

The classification results are in Table 2.2. The probability weighting functions were mainly classified as inverse S-shaped, S-shaped or pessimistic while the proportion of optimistic weighting functions was negligible in both treatments. Among the three main types, the majority of the probability weighting functions was inverse S-shaped in the DFD treatment (p=0.00, one-sided binomial, H0: Proportion of inverse S is $\frac{1}{3}$ among inverse S, S and pessimistic types). The inverse S-shape was also the most frequent type in the DFE treatment but it was not the majority (p=0.13, one-sided binomial, H0: Proportion of inverse S is $\frac{1}{3}$ among inverse S, S and pessimistic types).

Overall, our individual level analysis suggested reduced, but persistent, inverse S pattern in the DFE treatment. The preceding results are valid without requiring any parametric assumptions or specification of the stochastic nature of errors. The parametric analysis in the next section supplements our nonparametric results.

Parametric estimations. We made the parametric analysis of probability weighting functions by implementing Bayesian hierarchical estimation procedure. This procedure enables reliable aggregate and individual level estimations with limited data available per subject. It was recommended by Nilsson, Rieskamp, & Wagenmakers (2011) and Scheibehenne & Pachur (2015), and employed by several other studies for estimating RDU and cumulative PT components (Balcombe & Fraser, 2015; Kellen, Pachur, & Hertwig, 2016; Lejarraga, Pachur, Frey, & Hertwig, 2016).

We estimated Goldstein & Einhorn's (1987) weighting function given by $w(q) = \frac{\delta q^{\gamma}}{\delta q^{\gamma}+(1-q)^{\gamma}}$. The parameter γ determines the curvature and captures the sensitivity towards changes in probabilities. Here, $\gamma < 1$ indicates inverse S-shape and likelihood insensitivity, and $\gamma > 1$ indicates S-shape and extreme likelihood sensitivity. The parameter δ determines the elevation, and captures the degree of pessimism. For $\delta = 1$, we have $w(0.5) = 0.5$. Lower (higher) values of δ indicates less (more) elevation and more (less) pessimism. Following Kruschke (2011), we evaluate the credibility of likelihood insensitivity and pessimism based on the ranges of 95% intervals from posterior distribution of parameters. The details on estimation procedures are in Appendix 2.3.

Table 2.3 Group level mean parameters

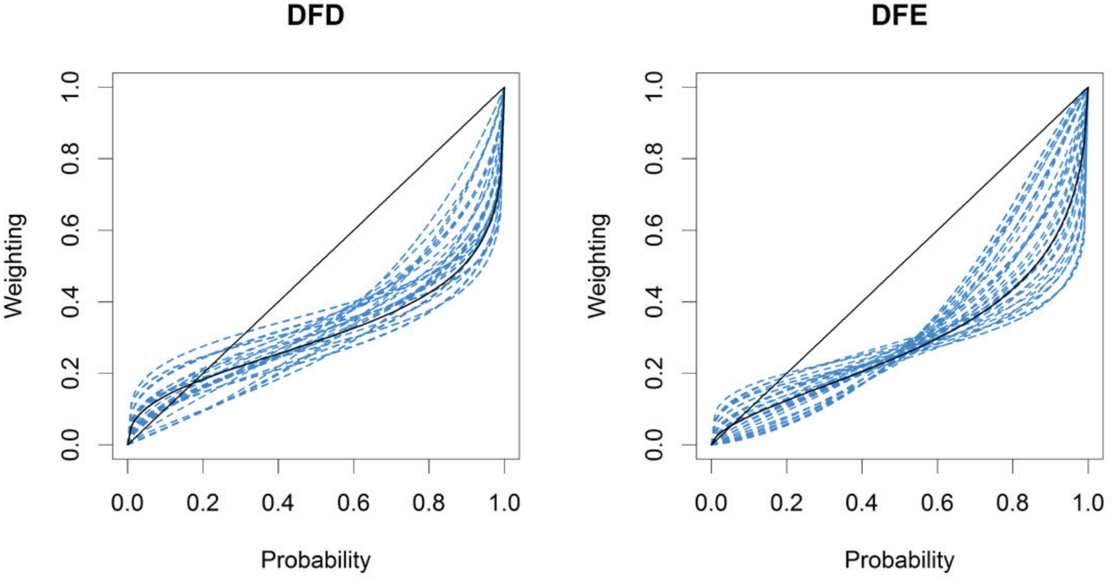
	γ	δ
DFD	0.430	0.407
	[0.234, 0.675]	[0.259, 0.590]
DFE	0.611	0.331
	[0.372, 0.868]	[0.198, 0.508]
Gap	-0.181	0.076
	[-0.517, 0.160]	[-0.152, 0.304]

Notes: Estimated parameters are the means of the posterior distributions of the group level means. 95% credibility intervals are given in square brackets.

We report the estimated group level mean parameters and corresponding 95% credibility intervals in Table 2.3. Figure 2.9 shows the estimated probability weighting functions. The estimated parameters indicated credible likelihood insensitivity and pessimism in both treatments as $\gamma = 1$ and $\delta = 1$ fell on the right side of 95% credibility intervals. The DFD-DFE gap in terms of likelihood insensitivity and pessimism was not credible, although the difference in likelihood insensitivity was suggestive. Hence, we observed a less pronounced inverse S-shape in the DFE weighting function, while the elevation was comparable across the two treatments (black curves in Figure 9).

At the individual level, pessimism ($\delta < 1$) was credible for all the subjects in both treatments. Likelihood insensitivity was credible for 51% (23 out of 45) of the subjects in the DFD treatment and for 29% (13 out of 43) of the subjects in the DFE treatment. While there was no subject with likelihood sensitivity ($\gamma > 1$) in the DFD treatment, 23% (10 out of 43) subjects in the DFE exhibited likelihood sensitivity, although it was never credible.

Figure 2.9 Probability weighting functions



Notes: Blue/dashed curves are individual level probability weighting functions based on the means of individual level posterior distributions. Black curve is the group level probability weighting function based on the mean of the posterior distribution of the group level mean.

2.7 Discussion

Our adjustment of the sampling paradigm with complete sampling of outcomes allowed us to observe the pure impact of sampling experience on risk attitudes. Both nonparametric and parametric analysis indicated that the sampling experience attenuates but does not reverse biases at extreme probabilities. Our results suggested that sampling experience mainly attenuates likelihood insensitivity but it does not have much impact on pessimism towards risk.

The de-biasing effect of sampling experience can be explained by two possible factors. First, the two informationally-identical treatments may suggest distinct cognitive processes for different information formats as argued by Gigerenzer & Hoffrage (1995). In particular, insensitivity to probabilities diminishes, similar to deviations from Bayesian updating, when the probabilistic information is acquired through sequential sampling in terms of natural frequencies. Other studies by Hogarth & Soyer (2011) and Hogarth, Lejarraga, & Soyer (2015) also emphasize the importance of the structure of the learning environment for reduction of biases in judgment and decision making. In particular, a kind learning environment, where the samples collected by the decision maker provide an accurate representation of the target population, is a necessary condition for unbiased judgments and choices. Our experimental design provides a kind learning environment in the absence of sampling biases and ambiguity.

As regards the second factor, the DFD-DFE gap can signify other internal biases due to memory limitations and/or inattention (Camilleri & Newell, 2011). To avoid these potential confounds in our experiment, we provided our subjects with paper and pen and reminded them that they can keep track of the outcomes during the sampling stage in DFE. We observed that more than half of the subjects in the DFE treatment took notes. Hence, our results were less likely to be driven by misremembering the past observations.

Contrary to our findings under risk, more pronounced pessimism and likelihood insensitivity were reported by some previous studies of DFE concerning ambiguity (Abdealloui et al., 2001; Glöckner et al., 2016; and Kellen et al., 2016). Such impacts of ambiguity are prevalent in the literature, and will be the topic of Chapter 3. Here, our conclusions on the impact of sampling experience on probability weighting are consistent with some other previous findings. Gottlieb, Weiss, & Chapman (2007), Hilbig & Glöckner (2011), and Humphrey (2006) also report reduced probability weighting with different variants of the sampling paradigm under risk. Erev, Ert, Plonsky, Cohen, & Cohen (2015, pp. 7-11), Jessup, Bishara, & Busemeyer (2008), van de Kuilen & Wakker (2006), and van de Kuilen (2009) report significant convergence to EU maximization under risk in repeated choice settings, when immediate feedback after each choice is available but not when it is unavailable. These results also suggest the distinct impact of experience in repeated choice settings (also see Lejarraga & Gonzalez, 2011 on strong impact of experience).

Two-Stage Design, Non-parametric and Parametric Analysis

Our two stage experimental design avoided potential interdependencies between utility and probability weighting components of RDU, which was reported by previous studies. Moreover, it enabled a reliable non-parametric analysis of probability weighting functions without relying on specific functional forms, which can be subject to distortions. Our parametric estimations were consistent with our nonparametric analysis. To further test the descriptive adequacy of the parametric Bayesian estimations, we compared posterior predictions of the estimated model with the actual data observed (see Appendix 2.3, Figure A2.2). The model was accurate in predicting choices.

Despite the aforementioned advantages of the nonparametric approach, one might still have concerns about our two-stage design. One concern is the error propagation in the chained procedure. In particular, the stimulus for the measurement of probability weighting in the second stage is determined based on the utilities elicited in the first stage. Thus, any error in the calculation of s_q from the first stage may result in a bias in probability weighting measurements. However, studies investigating this point have shown that this problem is indeed negligible (Abdellaoui et al. 2005; Bleichrodt and Pinto 2000). Moreover, high goodness of fit in estimations of utility functions, and the replication of common qualitative patterns of probability weighting under DFD confirm the validity of our procedure.

Another concern is incentive compatibility of the TO method due to its adaptive nature (later stimuli being determined by previous choices). However, no previous studies have found this to be a problem in experiments (Abdellaoui, 2000; Bleichrodt, Cillo, & Diecidue, 2010; Qiu & Steiger, 2011; Schunk & Betsch, 2006; Van de Kuilen & Wakker, 2011). Hence, in the terminology of Bardsley et al. (2010), there is only a concern for theoretical incentive compatibility but not for behavioral incentive compatibility (p. 265). Still, as a precautionary measure, our bisection procedure also included filler questions in the iteration process, aiming to make the detection of our adaptive design even more difficult. Our data did not show any evidence of strategic choices (Appendix 2.2).

Lastly, our experimental design makes an implicit assumption that the sampling experience has an impact on the probability domain but not on utilities. This assumption enabled us to measure utilities under the more efficient DFD paradigm in the first stage. The assumption was empirically supported in previous studies by Abdellaoui, L'Haridon, & Paraschiv (2011) and by Cubitt et al. (2016), where the utilities were estimated under DFE separately. It was also supported in estimations with the data sets of Glöckner et al. (2016) and of Erev et al. (2010). These two studies are investigated in Chapter 3.

2.8 Conclusion

This paper clarifies the controversy about the DFD-DFE gap. Our strictly controlled sampling paradigm isolates the impact of the sampling experience from other confounds, and the two stage design reveals the exact weighting of probabilities under DFE. The experimental findings support the DFD-DFE gap. However, the gap does not amount to a reversal of the inverse S-shaped probability weighting, and there is no actual underweighting of rare and extreme outcomes in DFE. Our findings illustrate the importance of the learning experience in reducing irrationalities. Decisions from experience do not reverse an irrationality into another irrationality but rather reduce the cognitive impairment of likelihood insensitivity. Black swans are not ignored under DFE.

Appendix 2.1

Derivation of the Standard Sequence of Outcomes in TO Method

Under RDU, indifferences $x_{i+1_p}g \sim x_{i_p}G$ imply $w(p)U(x_{i+1}) + (1 - w(p))U(g) = w(p)U(x_i) + \big(1 - w(p)\big)U(G)$. A rearrangement of this equation shows $U(x_{i+1}) - U(x_i) = \frac{(1-w(p))}{w(p)}[U(G) - U(g)]$ for all $0 \leq i \leq n-1$. Because the right hand side of the equation is fixed by the design, the indifferences result in $U(x_1) - U(x_0) = U(x_2) - U(x_1) = \cdots = U(x_{n-1}) - U(x_n)$.

Appendix 2.2

Bisection Procedure

The iteration process serves to measure $x_1, x_2, x_3, x_4,$ and x_5 from the following indifferences, with $p = 0.33, G = 17, g = 9, x_0 = 24$:

$$x_{0_p}G \sim x_{1_p}g,\ x_{1_p}G \sim x_{2_p}g,\ x_{2_p}G \sim x_{3_p}g,\ x_{3_p}G \sim x_{4_p}g,\ x_{4_p}G \sim x_{5_p}g$$

For each x_i, it took five choice questions to reach the indifference point. Subjects always chose between two prospects: $x_{i_p}g$ and $x_{i-1_p}G$ for $i = 1, \ldots, 5$. The procedure was as follows.

1. The initial value of x_i was determined as $x_{i-1} + 4(G - g) = x_{i-1} + 32$.
2. x_i was increased by a given step size when $x_{i-1_p}G$ was chosen over $x_{i_p}g$, and decreased when $x_{i_p}g$ was chosen over $x_{i-1_p}G$ as long as $x_i > x_{i-1}$. In case $x_i \leq x_{i-1}$, x_i was increased in order to ensure outcome monotonicity.
3. The initial step was $4(G - g)$ =32. The step sizes were halved after each choice.
4. The indifference point was reached after five choices.
5. The largest possible value of x_i was $x_{i-1} + 32 + 32 + 16 + 8 + 4 + 2 = x_{i-1} + 94$.
6. The smallest possible value of x_i was $x_{i-1} + 32 - 32 + 16 - 8 - 4 - 2 = x_{i-1} + 2$. The fourth term on the left hand side (+16) ensured the outcome monotonicity (see point 2).

One concern for the TO method and the bisection iteration process is the incentive compatibility due to the adaptive design. A subject who is fully aware of the adaptive design can strategically drive the value x_i upwards by pretending to be extremely risk averse in the bisection questions. In this way, he or she can increase the expected values of prospects in the subsequent questions for the elicitation of x_{i+1}. To make it more difficult for our subjects to fully grasp the process, we included two filler questions in the iteration process of each x_i. The two filler choices were after the first and the third choice questions for every x_i. In these questions, x_i was changed in the direction that is opposite to the changes described in point 2 above. These questions had no further impact on the flow of the procedure.

Our data did not suggest any strategic behavior. While an awareness of the adaptive design from the outset is fairly unlikely, learning during the experiment would lead to increasing distances between x_is. This means that a systematic learning of the strategic choice during the experiment would give us larger distances between x_5 and x_4 than between x_1 and x_0. On the contrary, the median distances in our data were 26 and 34 respectively, and did not differ significantly (Wilcoxon sign-rank, p-value=0.54).

Appendix 2.3

Bayesian Hierarchical Estimation Procedure

We implemented the Bayesian hierarchical estimation procedure as follows. Goldstein & Einhorn's (1987) probability weighting function is $w(q) = \frac{\delta q^{\gamma}}{\delta q^{\gamma}+(1-q)^{\gamma}}$. The probability of choosing the risky prospect was calculated using Luce's (1959) stochastic choice function, which gave a better fit to our data than the logit function. It is $\Pr(choosing\ risky\ option) = \frac{RDU_{risky}^{\varphi}}{RDU_{risky}^{\varphi}+RDU_{safe}^{\varphi}}$, where φ is the noise parameter. After normalizing $U(x_1) = 0$, and $U(x_5) = 1$; $RDU_{risky} = w(q) * U(x_5) + (1 - w(q)) * U(x_1) = w(q)$, and $RDU_{safe} = U(x_q) = q$ by construction. Thus, the choice function implies random choice when $w(q) = q$, consistent with (3) in the Method section.

In the estimations, individual level parameters γ_i and δ_i were constrained by using plausible ranges based on the previous findings in the literature and on the findings from our nonparametric analysis. Given the limitations of the dataset, and especially the small number of observations at individual level, to ensure the identifiability, mildly small ranges were used for constraining individual level parameters. The ranges of the prior distributions were from 0.1 to 2 for γ_i and from 0.1 to 1.5 for δ_i. The range chosen for γ_i

allows a wide array of curvature ranging from strong inverse S-shape to strong S-shape. The range chosen for δ_i implies that $w(0.5)$ is between $\frac{1}{11}$ and $\frac{3}{5}$, which is considered as a reasonable range given the previous findings in the literature and our nonparametric results suggesting strong underweighting at 0.5.

To facilitate hierarchical modelling, following Rouder and Lu (2005), Nilsson et al. (2011) and Scheibehenne & Pachur (2015), we used probit transformations of individual level parameters γ_i and δ_i with linear linkages, i.e. $\gamma_i = 1.9 * \Theta(\gamma_i') + 0.1$ and $\delta_i = 1.4 * \Theta(\delta_i') + 0.1$ where Θ is the cumulative distribution function of the standard normal distribution. The probitized parameters γ_i' and δ_i' are assumed to come from normal distributions with $N(\mu_\gamma, \sigma_\gamma)$ and $N(\mu_\delta, \sigma_\delta)$ respectively. The priors of the group level means, μ_γ and μ_δ, were assumed to follow standard normal distributions, which result in uniform distributions with the aforementioned ranges when they are transformed back to rate scale. The priors of the group level standard deviations, σ_γ and σ_δ, were uniformly distributed ranging from 0 to 10.

The individual level noise parameters φ_i were assumed to come from a lognormal distribution. Similarly, to facilitate the hierarchical modelling, we used logarithmic transformations of φ_i, i.e. $\varphi_i = exp(\varphi_i')$, where the prior of φ_i' was assumed to follow $N(\mu_\varphi, \sigma_\varphi)$. The group level mean, μ_φ, was assumed to be uniformly distributed ranging from -2.3 to 2.3, which results in a uniform distribution ranging from 0.1 to 10 in the exponential scale. The group level standard deviation σ_φ was uniformly distributed ranging from 0 to 1.33. The upper bound of 1.33 was determined as the standard deviation of the prior distribution of the group level mean, $U(-2.3, 2.3)$, following Nilsson et al. (2011, p. 88).

The MCMC algorithm was implemented in WinBUGS run through R software. Three chains, each with 60000 iterations were run, after a burn-in of 10000 iterations. To reduce the autocorrelation, only every 10th sample was recorded. Convergence was verified by Gelman-Rubin statistics, and by visual inspection of trace plots.

Figure A2.1 shows the posterior histograms for the group level mean parameters. Figure A2.2 shows the predictive performance of the estimations by comparing the median numbers of overweighting predicted by the posterior distributions of group level parameters with the actual numbers of overweighting observed in our data. The model predictions match with the observed data for 0.2 and 0.9 in the DFD treatment, and for

0.05 and 0.1 in the DFE treatment. The predictions for the other probabilities were close to the actual data in the DFE treatment. The predictions for 0.05 and 0.8 in the DFD treatment indicated some misalignment with the actual data, although they performed well in the rest of the probabilities.

Figure A2.1 Posterior histograms for group level means

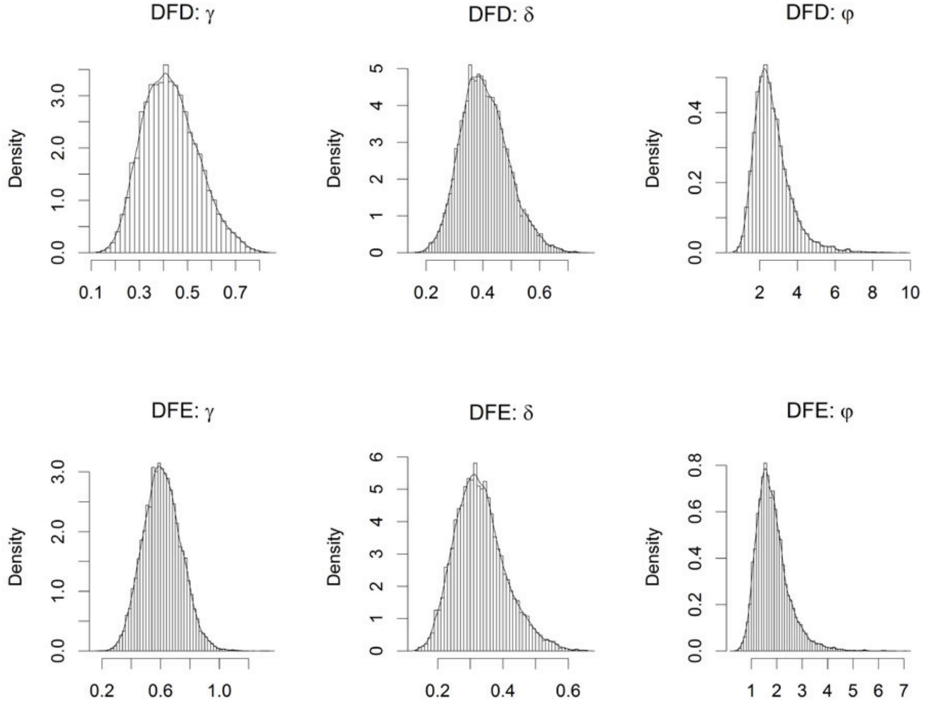
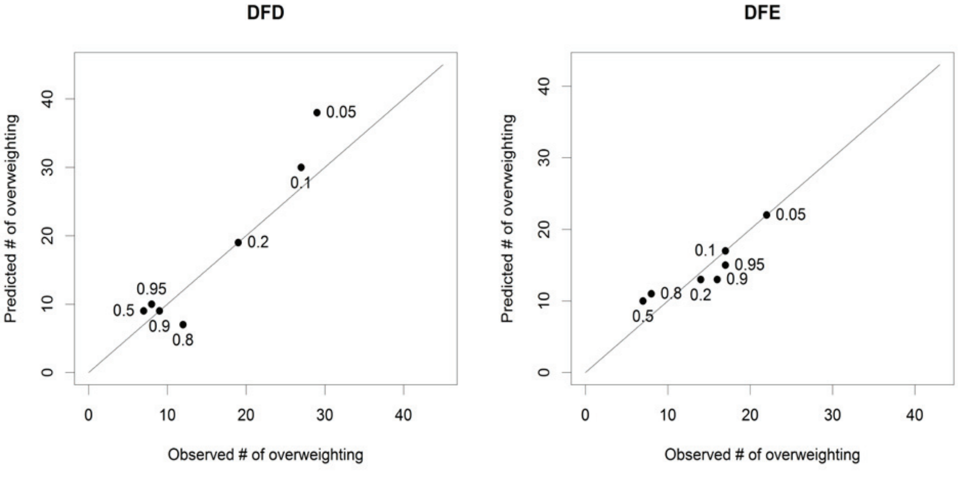


Figure A2.2 Posterior predictions based on group level parameters



Chapter 3

The Role of Prior Beliefs in Decisions from Experience

3.1 Introduction

Early studies of decisions from experience (henceforth, DFE) suggested that people make choices as if they underweight the impact of rare outcomes. This empirical observation is inconsistent with the findings from traditional decisions from description (henceforth, DFD) and with the predictions of prospect theory (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992), the most prominent theory for risk and uncertainty. Since the influential studies by Barron & Erev (2003) and Hertwig et al. (2004) introducing the intriguing DFD-DFE gap, an ever-growing DFE literature has clarified two main factors underlying the gap.

First, under-observation of the rare outcomes in small samples, also known as the sampling error, is a major factor underlying the underweighting. This implies that observed relative frequencies of outcomes rather than the objective probabilities, which are unknown to the decision maker, should count in DFE (Fox & Hadar, 2006). Controlling for sampling error, the DFD-DFE gap still amounts to less overweighting, but not to underweighting, under DFE as it was claimed originally (Ungemach et al., 2009; Hau et al., 2009; Camilleri et al., 2009).

The second factor concerns the information asymmetry between DFD and DFE (Hadar & Fox, 2009). Whereas DFD involves risk (known probabilities), DFE involves ambiguity due to incomplete information about the set of possible outcomes and probabilities. Importantly, the decision maker also lacks a priori knowledge about certainty or possibility of outcomes, which is relevant in the presence of always – or never – sampled outcomes. Indeed, several studies have shown that the gap is reduced or even reversed if the information asymmetry is reduced by providing information about the possible outcomes in prospects or in the absence of sure outcomes in choice problems (Abdellaoui, L'Haridon, & Paraschiv, 2011; Glöckner, Hilbig, Henninger, & Fiedler, 2016; Hadar & Fox, 2009; Kemel & Travers, 2016; Kellen, Pachur, & Hertwig, 2016). The recently found reversed DFD-DFE gap, implying even more pronounced overweighting of rare outcomes under DFE, is consistent with the previous literature on ambiguity (Abdellaoui, Vossmann, & Weber, 2005; Abdellaoui, Baillon, Placido, & Wakker, 2011; Fox & Tversky, 1998; Fox, Rogers, & Tversky, 1996; Tversky & Fox, 1995; Tversky & Wakker, 1995). The decreased likelihood sensitivity[9] under ambiguity is commonly attributed to the overestimation of infrequent outcomes due to sub-additive subjective beliefs or regression to the mean effects in probability estimations (Erev, Wallsten, & Budescu, 1994; Fiedler, Unkelbach, & Freytag, 2009; Fiedler & Unkelbach, 2014; Rottenstreich & Tversky, 1997; Tversky & Koehler, 1994).

This paper points to the role of prior beliefs as another important factor in the DFD-DFE gap. Except for a few studies eliciting introspective judged probabilities (Hau et al. 2008; Camilleri & Newell 2009; Ungemach et al. 2009), previous studies usually approximate subjective probabilities with observed relative frequencies neglecting the role of prior beliefs. However, the importance of the subjective prior beliefs is particularly evident in the face of the ambiguous nature of DFE because every subject brings his own prior expectations about the experimental setting into the laboratory. For example, even though not specified explicitly, a subject can reasonably anticipate the range of possible outcomes, and predict that extreme losses or gains do not occur very frequently for ethical reasons or the budgetary constraints of the experimenter. The ecological rationality account of Pleskac & Hertwig (2014) illustrates the importance of such intuitions under ambiguous situations. If prior beliefs are not incorporated into the analysis of subjective probabilities, then the estimations of probability weighting may be confounded because the impact of prior beliefs will incorrectly be modeled through probability weighting.

This study puts forward a more complete account of subjective probabilities under DFE, which involves a combination of prior beliefs with observed relative frequencies. As a working hypothesis, the present account proposes a Bayesian updating method for the estimation of subjective probabilities. Notably, the Bayesian updating of an ignorance prior will estimate the probability of an infrequent outcome higher than its observed relative frequency. This gives a rational basis for the regressions to the mean effects in probability estimations, also documented in the previous DFE studies eliciting judged probabilities.

Hence, I introduce a two-stage decision model for DFE according to which (1) subjective probabilities are estimated using a Bayesian updating method developed by Rudolf Carnap (1952); (2) and the estimated probabilities are transformed using prospect

[9] The concept of likelihood insensitivity was explained in section 2.2.

theory's rank- and sign-dependent probability weighting (Tversky & Kahneman, 1992). Besides being a normative method for belief updating, Carnap's method, introduced in the next section, is also psychologically natural. Following the source method of Abdellaoui et al. (2011), the probability weighting in the second stage captures deviations from Bayesian rationality, and it is assumed to be source dependent. This means that having revealed the subjective probabilities, the model allows for different attitudes towards described and experienced probabilities observable through different probability weightings. It should be noted that the current model differs from the two-stage model of Tversky & Fox (1995) and Fox & Tversky (1998), which attributes ambiguity attitudes to sub-additive beliefs under uncertainty. Another distinguishing feature of the current account is that it adheres to the revealed preference approach of (behavioral) economics by relying on choice-based probabilities rather than introspective probability judgments.

The two-stage model is empirically tested by reanalyzing the data sets of Glöckner et al. (2016), as well as the Technion Prediction Competition data set of Erev et al. (2010). As will be illustrated later, the model successfully disentangles the role of beliefs from preferences in DFE. Accordingly, the reversed DFD-DFE gap in probability weighting is estimated to be considerably smaller when prior beliefs are controlled for. Moreover, the classic DFD-DFE gap is also reduced, or even reversed, under some plausible assumptions on subject's prior expectations about the set of possible outcomes. Overall, the robust likelihood insensitivity under DFE suggests further deviations from Bayesian rationality due to ambiguity. Lastly, model comparisons based on Bayesian Information Criteria (BIC) scores also indicate that the two-stage model performs better than the single stage approach using observed relative frequency approximation of subjective probabilities. Thus, the two-stage model provides a parsimonious way to analyze DFE by adding only one extra parameter to the preceding models.

3.2 Carnap's updating method and the two-stage model

The current paper makes use of the inference method that Rudolf Carnap[10] put forward to quantify the degree of confirmation of a hypothesis stating that the next observation from a

[10] Rudolf Carnap is a well-known philosopher of science who also contributed to the theory of probability by providing a logical definition of probability (Carnap, 1945; 1950; 1952). Briefly, his theory views probability as a logical relation between two statements, namely the degree of confirmation of a hypothesis *h* on the

population will be the outcome x_i based on the evidence that a previous sample of N observations contains n_i observations from the outcome x_i:

$$p_i = \frac{cp_i^0 + N\frac{n_i}{N}}{c + N} .$$

Thus, as in every Bayesian approach, the method combines a prior probability p_i^0 of the outcome x_i with the observed relative frequency $\frac{n_i}{N}$. The respective weights are proportional to a constant $c > 0$ and the total number of observations N. The prior probability p_i^0 together with the constant c represents the complete prior knowledge of the decision maker. A common intuitive interpretation is that the prior knowledge of the decision maker can be thought to be roughly equivalent to a hypothetical sample consisting of c observations with relative frequency p_i^0. In Bayesian inference, the method is also known from the updating of the conjugate beta family and the conjugate Dirichlet family for its multinomial extension (Winkler, 1972; Wilks, 1962; Zabell, 1982).

Carnap's method is empirically appealing. First, the posterior probability of an outcome always lies between the prior and the observed relative frequency. The estimation converges to the relative frequency as more and more observations are accumulated, reflecting increasing confidence in empirical probabilities. In the case where there are two possible outcomes, a "flat" prior representing ignorance is captured by $p_i^0 = \frac{1}{2}$ and $c = 2$, which turns the formula into the posterior mean of a uniform beta prior (Winkler, 1972). This case is illustrated in Figure 3.1. The posterior estimations tend to the 50/50 prior especially when the number of observations is small. This tendency reduces significantly as the number of observations increases from 5 to 40.

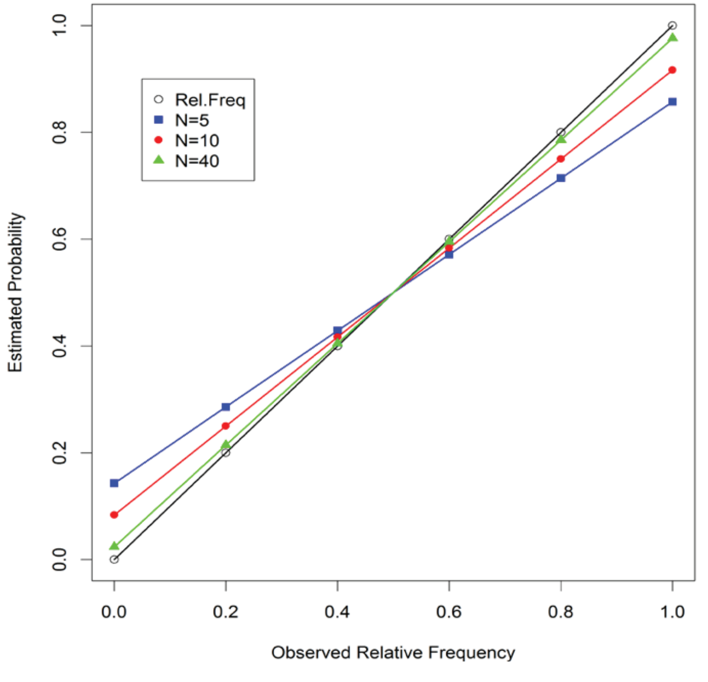
Second, the method reduces to relative frequency when c converges to 0. Carnap (1945, p. 86) points out the major problem of using relative frequencies in estimations of probabilities concerning always – or never – observed outcomes. This problem is also commonly encountered in DFE experiments. In particular, assigning 1 or 0 probability to these outcomes may be implausible. A famous historical example of this issue is Laplace's (1825) sunrise problem, asking the likelihood of the sun rising tomorrow. Laplace's rule of succession for dealing with the problem, $p_i = \frac{1+N}{2+N}$, is simply the restricted version of

evidence e (Carnap, 1945, p. 72), which is analogous to the concept of the degree of belief in the theory of subjective probability.

Carnap's method under ignorance illustrated in Figure 3.1.[11] For example, the method results in $p_i = \frac{1+10}{2+10} = \frac{11}{12} \cong 92\%$ when the observed relative frequency is $10/10$. The posterior estimation converges to certainty when N increases.

Carnap (1952) justifies the appropriateness of the method by providing logical axioms for it. Wakker (2002) presents the axioms in a decision theoretic context, and highlights the normative status of the method. The first property is positive relatedness of the observations. It means that an extra observation from an outcome only increases its likelihood. The second property is exchangeability. It means that only the number of observations from the outcomes matters, regardless of the order of observations. The third property is disjoint causality. It means that there is no causal relationship between different outcomes. Therefore, the probability of an outcome x_i depends only on the number of observations of itself (n_i) and of not-itself $(N - n_i)$, regardless of which other outcomes were observed among the $(N - n_i)$ other outcomes.

Figure 3.1 Posterior estimations with Carnap's method when $\mathrm{p}_\mathrm{i}^0 = \frac{1}{2}$ and $\mathrm{c} = 2$



[11] A historical review of the rule of succession is in Zabell (1989).

In principle, Carnap's properties are applicable to the DFE experiments, where the sample information is obtained from a fixed outcome distribution with replacement, i.e. from a stationary and independent process. It is worth noting that the properties can be violated due to subjects' unjustified beliefs about the random processes and the cognitive illusions such as the hot hand and gambler's fallacies (Tversky & Kahneman, 1971; Kahneman & Tversky, 1972; Ayton & Fischer, 2004; Sundali & Croson, 2006). However, these effects have been mainly documented in repeated settings such as in feedback paradigms of DFE (Barron & Yechiam, 2009) and in probability matching tasks (Sundali & Croson, 2006) but not in the sampling paradigm, where the observations are made only for the purpose of information acquisition.

Having constructed beliefs using Carnap's method, the two-stage model assumes that prospects are evaluated by prospect theory (Tversky & Kahneman, 1992) in the second stage. In what follows, I denote a prospect with outcomes $x_1, \dots, x_n$ with respective probabilities $p_1, \dots, p_n$ by $(p_1: x_1, \dots, p_n: y_n)$. The prospect theory value of an experienced prospect with $x_1 > \dots > x_k > 0 > x_{k+1} > \dots > x_n$ is

$$PT(p_1: x_1, \dots, p_n: x_n)$$
$$= \sum_{i=1}^{k} u(x_i)\,[w_e^+(p_i + \dots + p_1) - w_e^+(p_{i-1} + \dots + p_1)]$$
$$+ \sum_{j=k+1}^{n} u(x_j)\,[w_e^-(p_j + \dots + p_n) - w_e^-(p_{j+1} + \dots + p_n)].$$ [12]

The utility $u(.)$ is strictly increasing and continuous with $u(0) = 0$. The probability weighting functions $w_e^s(.)$ for gains ($s = +$) and losses ($s = -$) are strictly increasing with $w_e^s(0) = 0$ and $w_e^s(1) = 1$. Here, the subscript e designates the experienced source of ambiguity. Specifically, $w_e^s(.)$ measures the weighting of subjective probabilities under DFE. Prospects under DFD are similarly evaluated by prospect theory, where $w_e^s(.)$ is replaced by $w^s(.)$ measuring the weighting of objective probabilities. Hence, different attitudes towards experienced ambiguity and described risk can be captured by differences between $w_e^s(.)$ and $w^s(.)$.

[12] Here $p_{i-1} + \dots + p_1 = 0$ when $i = 1$, and $p_{j+1} + \dots + p_n = 0$ when $j = n$.

3.3 Testing the two-stage model

The following sections provide an empirical test of the two-stage model by parametric estimations of the prospect theory components under the two-stage model. I use Goldstein & Einhorn's (1987) two-parameter family for probability weighting, and the commonly used power family for utility. The choice probabilities are calculated using the stochastic logit rule.

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

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

$$u(x) = \begin{cases} x^{\alpha} & if\ x \geq 0 \\ -\lambda(-x)^{\beta} & if\ x < 0 \end{cases}$$

$$p(A,B) = \frac{1}{1 + e^{-\sigma(PT(A)-PT(B))}}$$

The parameter δ^s determines the elevation of probability weighting, and measures the pessimism/optimism of the decision maker. Higher δ^s leads to more elevation, and thus more optimism in the gain domain and more pessimism in the loss domain. The parameter γ^s determines the curvature of the probability weighting function and it captures sensitivity towards probabilities. For, $\gamma^s < 1$ the probability weighting is inverse S-shaped reflecting likelihood insensitivity. The parameter λ determines the degree of loss aversion. To avoid extra complexity, utility curvature in the gain and in the loss domain is assumed to be the same by constraining $\alpha = \beta$. This assumption avoids an identification problem in the estimation of loss aversion (Wakker, 2010, section 9.6), and it is empirically supported by previous findings (Tversky & Kahneman, 1992). For $\alpha, \beta < 1$, the utility curve is concave in the gain domain and it is convex in the loss domain. Lastly, the parameter σ in the logit formula determines the sensitivity to differences in prospect theory values of prospects.

Three different cases of subjective priors are considered in the estimations. The first case assumes symmetric prior probabilities, p_i^0, equally distributed over all the outcomes that are believed to be possible in a prospect, and the constant c in Carnap's formula is treated as a free parameter to be estimated together with the other parameters. Hereafter, this will be called the *Carnap prior* case. The second case concerns the *ignorance prior* that was already mentioned in the previous section. The *ignorance prior* is a special case of the *Carnap prior* case. It assumes that the prior knowledge of the subject is equivalent to a hypothetical sample that contains one and only one observation from each of the possible outcomes. For instance, for a prospect with k possible outcomes, the prior probability of every outcome is $\frac{1}{k}$ and $c = k$. The third case suppresses prior beliefs altogether and simply approximates subjective probabilities with observed relative frequencies. In other words, this is Carnap's method with $c = 0$. Following the Bayesian terminology, this case will be called the *diffuse prior* (Winkler, 1972, p. 178). This case has been commonly used in the previous DFE studies.

3.3.1 Accounting for the reversed DFD-DFE gap: A reanalysis of Glöckner et al. (2016)

Contrary to previous studies of DFE, Glöckner et al. (2016) reports reversed DFD-DFE gap based on an analysis of four experimental data sets. They find more pronounced overweighting of small probabilities under DFE than under DFD. The authors attribute the discrepancy to information asymmetry between DFD and DFE conditions in early DFE studies, especially in the presence of sure outcomes whose certainty is not known by subjects in DFE conditions. They explain the reversal of the gap in the absence of such outcomes by regression to the mean effects in probability estimations due to noise and reduced evaluability under uncertainty. They also point out the possibility of an alternative explanation with updating of ignorance priors in a footnote (footnote 8, p. 490). This section tests this alternative explanation by re-examining the four data sets using the two-stage model. The data sets are made available by the authors at Open Science Framework: https://osf.io/d9f8q/.

Data

The first data set is obtained from a previous study by Glöckner, Fiedler, Hochman, Ayal, & Hilbig (2012). The second and the third data sets are based on the Experiments 1 and 2 in Glöckner et al. (2016). These experiments replicate the experiment by Glöckner et al. (2012) with slight procedural variations. The choices concern only the gain domain in these three data sets. The fourth data set is based on Experiment 3 in Glöckner et al. (2016). This data set contains choices in the gain, loss, and mixed domains. All the choice problems in the study of Glöckner et al. (2012) and in Experiment 2, and a large majority of the choice problems in Experiment 1 and in Experiment 3 involve a choice between two two-outcome prospects. The rest of the problems involves a choice between a two-outcome prospect and a sure outcome. Subjects in DFE conditions were informed about the number of possible outcomes in prospects, except for half of the subjects in Experiment 2. In what follows, I will refer to these two conditions within the DFE condition of Experiment 2 as *DFE-informed* and *DFE-uninformed*. Readers are referred to Glöckner et al. (2016) for more details on the experimental design.

Analysis

Parameters are estimated by the method of maximum likelihood using the estimation routine in STATA software described by Harrison (2008). Standard errors are cluster-corrected at the individual subject level. The DFD-DFE gap is tested by the significance of dummy variables for DFE treatment. Model comparisons are based on BIC scores.

To control for potential interactions between utility curvature and elevation of probability weighting commonly documented in previous studies (Gonzalez & Wu, 1999, p. 152; Scheibehenne & Pachur, 2015, pp. 403-404; Stott, 2006, p. 112; Zeisberger, Vrecko, & Langer, 2012), same utility functions under DFD and DFE conditions are assumed in the estimations. This avoids a problem of identification in ambiguity aversion that trends in utility and probability weighting elevation across DFD and DFE conditions may pose. For instance, in the gain domain, whereas less (more) utility curvature under DFE than under DFD implies ambiguity seeking (aversion), less (more) elevated probability weighting under DFE than under DFD implies ambiguity aversion (seeking).

Unconstrained estimations are reported in Appendix 3. Constrained estimations outperform unconstrained estimations in all data sets based on the BIC scores. Consistent with the previous studies by Abdellaoui, L'Haridon, & Paraschiv (2011) and by Cubitt et al. (2016), equality of utilities across DFD and DFE are supported in all data sets of Glöckner et al. (2016), only except for the *DFE-uninformed* condition of Experiment 2. Differences in unconstrained estimations are mentioned in the discussion.

Results

The estimation results based on *diffuse, ignorance* and *Carnap priors* are in Table 3.1a, Table 3.1b, and Table 3.1c. Last columns report the estimation results with the pooled data set.

Comparison of BIC scores across different prior assumptions under DFE are in Table 3.2. The resulting probability weighting functions are in Figure 3.2.

The estimations based on *diffuse priors* indicate significant DFD-DFE gap with respect to the likelihood sensitivity parameters $\gamma^{+/-}$ in all data sets (Table 3.1a). There is less likelihood sensitivity in the DFE condition than in the DFD condition. No significant DFD-DFE gap is observed in other model components, except for the error parameter in Experiment 1. Here, we observe significantly more error proneness under DFE than under DFD. Trends in Experiment 2 also suggest less elevated probability weighting under *DFE-informed* condition than under DFD condition, and more elevated probability weighting under *DFE-uninformed* condition than under DFD condition, although these effects are not significant (p=0.144, and p=0.259 respectively). The elevation of probability weighting differs across *DFE-informed* and *DFE-uninformed* conditions (p=0.031).

Table 3.1a. Estimations based on *diffuse priors*

	Glöckner et al. (2012) (DFD / DFE)	Experiment 1 (DFD / DFE)	Experiment 2 (DFD / $\mathrm{DFE_{informed}}$ / $\mathrm{DFE_{uninformed}}$)	Experiment 3 (DFD / DFE)	Pooled (DFD / DFE)
α	0.647	0.965	0.763	0.971	0.826
c	- / 0	- / 0	- / 0	- / 0	- / 0
δ^+	0.557/0.558	0.369/0.330	0.595/0.472/0.706	0.617/0.683	0.527/0.519
γ^+	0.732/0.560*	0.720/0.579*	0.983/0.512***/0.571***	0.560/0.424**	0.806/0.540***
δ^-				1.081/1.202	1.154/1.257
γ^-				0.868/0.458***	0.792/0.374***
λ				0.960	1.156
σ	2.157/2.012	0.844/0.617*	1.287/1.289/1.379	1.472/1.307	1.188/1.117
N	2581	3049	6092	5069	16791
LL	-1143.92	-1436.374	-2837.71	-2653.457	-8282.492
BIC	2342.832	2928.906	5762.568	5409.286	16681.73

Notes: Stars indicate DFD – DFE gap. $*p < 0.05$. $**p < 0.01$. $***p < 0.001$. *The first numbers in cells indicate the estimated parameters for DFD condition. In the column for Experiment 2, the second numbers are based on DFE-informed condition and the third numbers are based on DFE-uninformed condition.*

Table 3.1b. Estimations based on *ignorance priors*

	Glöckner et al. (2012) (DFD / DFE)	Experiment 1 (DFD / DFE)	Experiment 2 (DFD / $\text{DFE}_{\text{informed}}$ / $\text{DFE}_{\text{uninformed}}$)	Experiment 3 (DFD / DFE)	Pooled (DFD / DFE)
α	0.648	0.972	0.770	0.966	0.828
c	- / 2	- / 2	- / 2	- / 2	- / 2
δ^{+}	0.556/0.557	0.365/0.331	0.590/0.474/0.705	0.621/0.690	0.525/522
γ^{+}	0.732/0.670	0.721/0.715	0.984/0.623***/0.676**	0.559/0.502	0.806/0.651***
δ^{-}				1.085/1.225	1.149/1.272
γ^{-}				0.868/0.540**	0.791/0.445**
λ				0.961	1.162
σ	2.149/1.971	0.826/0.610*	1.258/1.280/1.342	1.486/1.332	1.181/1.114
N	2581	3049	6092	5069	16791
LL	-1147.33	-1433.739	-2833.117	-2648.986	-8274.346
BIC	2349.651	2923.63	5753.382	5400.349	16665.43

Notes: Stars indicate DFD – DFE gap. $^{*}p < 0.05$. $^{**}p < 0.01$. $^{***}p < 0.001$. *The first numbers in cells indicate the estimated parameters for DFD condition. In the column for Experiment 2, the second numbers are based on DFE-informed condition and the third numbers are based on DFE-uninformed condition.*

The gap in likelihood sensitivity is reduced under the two-stage model using the *ignorance priors* (Table 3.1b). In particular, the gap is less pronounced in all the data sets, and insignificant in Glöckner et al. (2012), in Experiment 1, and in the gain domain of Experiment 3. The DFD-DFE gap in error parameter in Experiment 1, and the trends in the elevation of probability weighting are replicated here. The elevation of probability weighting differs only across *DFE-informed* and *DFE-uninformed* conditions (p=0.035).

The estimations under the two-stage model demonstrate further reductions of the DFD-DFE gap in likelihood insensitivity when the *Carnap priors* are used (Table 3.1c). In this case, the gap is insignificant in all data sets except for Glöckner et al. (2012) and for the *DFE-uninformed* condition of Experiment 2. Surprisingly, Carnap's c is found significantly negative in the data set of Glöckner et al. (2012). This result resembles representativeness in probability updating where too much weight is assigned to the relative frequencies at the expense of prior probabilities (Grether, 1980; Griffin & Tversky, 1992). As the negative c induces underestimation of rare outcomes, the reversed gap is more pronounced here. In the rest of the estimations, the constant c is estimated positive although not significantly different from 0. The estimations with the pooled data set indicates $c > 0$ ($p = 0.08$), and the gap is accommodated in the gain domain but not in the loss domain. The gap in the error parameter in Experiment 1, and the trends in probability weighting elevation are also observed here. The elevation of probability weighting differs across *DFE-informed* and *DFE-uninformed* conditions (p=0.038).

Table 3.1c. Estimations based on *Carnap Priors*

	Glöckner et al. (2012) (DFD / DFE)	Experiment 1 (DFD / DFE)	Experiment 2 (DFD / $\mathrm{DFE_{informed}}$ / $\mathrm{DFE_{uninformed}}$)	Experiment 3 (DFD / DFE)	Pooled (DFD / DFE)
α	0.647	0.981	0.768	0.962	0.829
c	- / -1.603***	- / 7.618	- / 6.626 / 0.501	- / 10.192	- / 2.833
δ^+	0.556/0.538	0.360/0.328	0.592/0.477/0.704	0.625/0.697	0.525/0.522
γ^+	0.732/0.428**	0.723/1.013	0.984/0.815/0.600**	0.559/0.735	0.806/0.690
δ^-				1.089/1.263	1.147/1.277
γ^-				0.868/0.788	0.791/0.470*
λ				0.962	1.164
σ	2.156/2.033	0.804/0.591*	1.266/1.286/1.356	1.499/1.335	1.179/1.110
N	2581	3049	6092	5069	16791
LL	-1140.568	-1432.486	-2830.442	-2645.871	-8273.931
BIC	2343.984	2929.152	5765.460	5402.644	16674.33

Notes: Stars indicate DFD – DFE gap. $*p < 0.05$. $**p < 0.01$. $***p < 0.001$. *The first numbers in cells indicate the estimated parameters for DFD condition. In the column for Experiment 2, the second numbers are based on DFE-informed condition and the third numbers are based on DFE-uninformed condition.*

Table 3.2. Comparison of *Diffuse, Ignorance* and *Carnap Priors* under DFE

		Glöckner et al. (2012)	Experiment 1	Experiment 2 *DFE-informed*	Experiment 2 *DFE-uninformed*	Experiment 3	Pooled
Diffuse Prior	LL	-558.416	-775.766	-626.460	-660.924	-1340.574	-4056.785
	BIC	1138.302	1573.724	1274.696	1343.771	2720.436	8158.757
Ignorance Prior	LL	-561.826	-773.264	-621.364	-661.202	-1336.142	-4048.726
	BIC	1145.124	1568.721	1264.504	1344.328	2711.572	8142.64
Carnap Prior	LL	-555.065	-772.175	-619.279	-660.674	-1333.061	-4048.339
	BIC	1138.757	1573.939	1267.591	1350.580	2713.266	8150.903
N		1283	1632	1420	1492	2585	8412
$BF_{Diffuse-Ignorance}$		30.296	0.082	0.006	1.321	0.012	0.0003
$BF_{Diffuse-Carnap}$		1.255	1.113	0.029	30.099	0.028	0.020
$BF_{Ignorance-Carnap}$		0.041	13.585	4.681	22.783	2.333	62.271

Notes: $BF_{Null-Alternative}$ *indicates relative evidence for the null hypothesis against the alternative. BF values higher than 1 indicate evidence in favor of the null hypothesis. BF values smaller than 1 indicate evidence in favor of the alternative hypothesis.*

The comparisons of the BIC scores indicate that the two-stage model accounts for the data at least as good as the model based on the *diffuse prior*. The two-stage model with *ignorance priors* outperforms the model with *diffuse priors* in all data sets except for the data set of Glöckner et al. (2012) and for *DFE-uninformed* condition of Experiment 2. Interpreting from Bayes Factors[13] in Table 3.2, there is strong evidence in Experiment 1; and very strong evidence in *DFE-informed* condition of Experiment 2, in Experiment 3, and in the pooled data set in favor of *ignorance priors* over *diffuse priors* (for derivation of Bayes Factors from BIC scores, see Wagenmakers, 2007). There is almost no evidence in favor of *diffuse* or *ignorance priors* in *DFE-uninformed* condition of Experiment 2. There is also very strong evidence in favor of two stage model with *Carnap priors* over the model with *diffuse priors* in *DFE-uninformed* condition of Experiment 2, in Experiment 3, and in the pooled data set. There is almost no evidence in favor of *diffuse* or *Carnap priors* in Glöckner et al. (2012) and in Experiment 1. We find more evidence in favor of *ignorance priors* over *Carnap priors* in all data sets except in Glöckner et al. (2012). This suggests that using a free updating parameter c does not make a significant contribution in accounting for the data compared to using a constrained $c = 2$ (*ignorance prior*).

Discussion

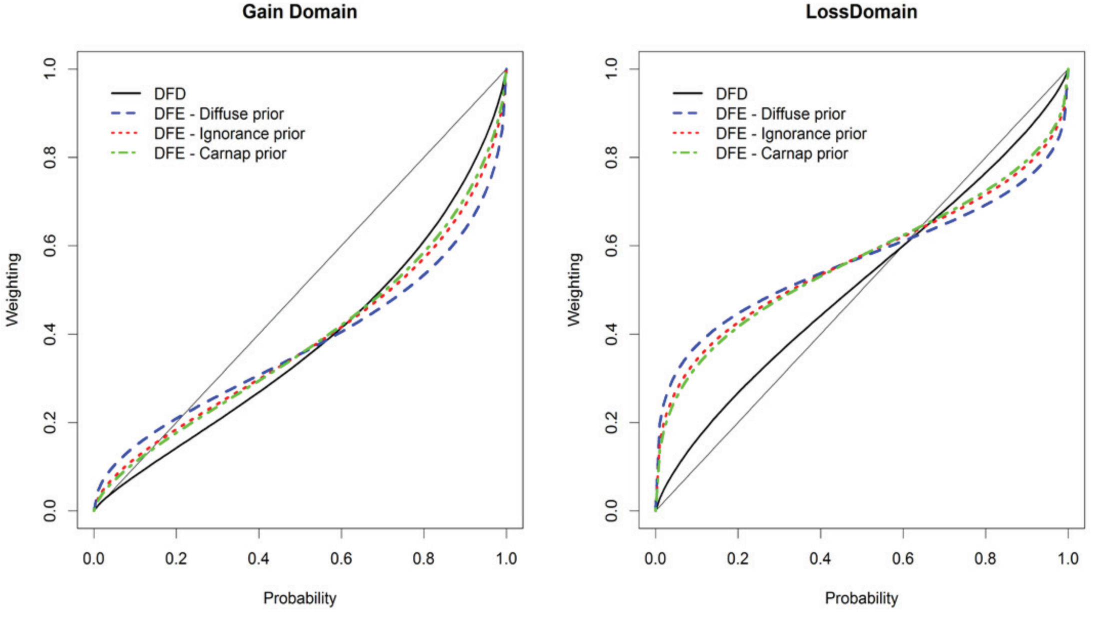
The aforementioned results support the updating account under DFE. The two-stage model with *ignorance prior* provides the best account of the data, and the gap in likelihood insensitivity is persistent as shown in Figure 3.2. This is consistent with ambiguity-generated likelihood insensitivity reported in previous studies (Abdellaoui, Baillon, Placido, & Wakker, 2011; Dimmock, Kouwenberg, & Wakker, 2015). No significant DFD-DFE gap is observed in any other prospect theory parameters.

While our main conclusions are replicated in unconstrained estimations without assuming equal utilities across DFD and DFE conditions, we only observe further differences in utility and elevation of probability weighting in Experiment 2 (Appendix 3, Table A3.1). Here, the trends suggest more utility curvature, and more elevation of probability weighting under DFD than under both *DFE-informed* and *DFE-uninformed* conditions. The trend in utility is found significant for the *DFE-uninformed* condition, and the trend in elevation is found significant for the *DFE-informed* condition. It should be noted that these trends in utility and probability weighting elevation imply opposite effects

[13] See section 2.6 in Chapter 2, for interpretation of Bayes Factors.

in terms of ambiguity aversion as mentioned previously. A reversed pattern, i.e. less utility curvature and less elevation under DFD than under DFE, is also observed in Experiment 1, although these are not significant.

Figure 3.2 Probability weighting functions based on the pooled data set of Glöckner et al.



Notes: Solid black lines show probability weighting under DFD. Dashed blue lines show probability weighting under DFE when the diffuse prior is used. Dotted red lines show probability weighting under DFE when the ignorance prior is used. Dot-dash green lines show probability weighting under DFE when the Carnap prior is used.

3.3.2 Accounting for the classic DFD-DFE gap: A reanalysis of Erev et al. (2010)

The study by Erev et al. (2010) reports the classic DFD-DFE gap. The ambiguity in the DFE (sampling) condition of this study is augmented by the lack of information about the certainty and possibility of outcomes. Whereas the ambiguity due to unknown probabilities can be easily studied in a tractable manner as illustrated in the previous section, the additional ambiguity about the set of possible outcomes poses a more complex problem to deal with. In particular, the subject's prior beliefs about the ambiguous outcome space are not easily observable. This section first introduces some plausible assumptions about prior beliefs under such ambiguous situations to explain the classic DFD-DFE gap. Then, the assumptions are empirically tested under the two-stage model by reanalyzing the Technion choice prediction competition data set of Erev et al. (2010).

Prior beliefs over ambiguous outcome space

Here, I describe two possible considerations in a subject's mind. The first is called *context-dependent expectations*, and it was put forward by Glöckner et al. (2016). To illustrate, consider the following choice problem under DFD involving two options: Option A is a sure outcome of 8.7 and Option B is a risky prospect $(0.91\colon 9.6,\ 0.09\colon -6.4)$ (taken from figure 3 in Glöckner et al.). Under DFE, the subject does not know the number of possible outcomes in the options, and therefore she is not aware of the certainty of the outcome 8.7 that she observes from Option A successively. Glöckner et al. (2016) argues that while forming beliefs about an option, the subject will not only use the information that she sampled from the very same option but also the information that she gathered from the other option. Accordingly, the rare outcome observed from Option B may be projected upon Option A. Specifically, the experience of the rare outcome -6.4, along with the common outcome 9.6 in Option B, can create an expectation that a similarly bad and rare outcome also exists in Option A. Hence, her belief about Option A involves an ambiguous prospect with two outcomes, 8.7 and ~ -6.4, rather than a sure outcome 8.7. Notably, this makes Option A less attractive, and therefore she may prefer Option B. However, if her prior beliefs are not taken into account, her preference for Option B gives the impression that she is underweighting the small probability of -6.4 in Option B.

More specifically, suppose that the subject updates her beliefs according to Carnap's method, and she uses an *ignorance prior*. Furthermore, she makes 10 observations from each option, always observes the outcome of 8.7 from Option A, and observes the outcome of 9.6 with relative frequency of $\frac{9}{10}$ and the outcome of -6.4 with relative frequency of $\frac{1}{10}$ from Option B. Thus, the relative frequency approximation of her subjective beliefs misleadingly implies that she is making a choice between the sure outcome $A\colon(\frac{10}{10}\colon 8.7)$ and a two-outcome prospect $B\colon(\frac{9}{10}\colon 9.6,\ \frac{1}{10}\colon -6.4)$, whereas she might indeed be making a choice between two two-outcome prospects, $A\colon(\frac{1+10}{2+10}\colon 8.7,$

$\frac{1+0}{2+10}:\sim -6.4)$ and B: $(\frac{1+9}{2+10}: 9.6,\ \frac{1+1}{2+10}: -6.4)$, as implied by *context dependent expectations* and Carnap's updating method.

The second scenario is a natural extension of the *context-dependent expectations*. It assumes that the subject will have more comprehensive beliefs by expecting that any outcome that she is aware of within a given problem can in principle exist in both options. Hence, all the outcomes observed from the two options are considered to be possible in both options. This scenario will be called *comprehensive expectations*. Continuing with the previous example, it leads to a choice between option A: $\left(\frac{1+0}{3+10}: 9.6,\ \frac{1+10}{3+10}: 8.7,\ \frac{1+0}{3+10}: -6.4\right)$ and option B: $(\frac{1+9}{3+10}: 9.6, \frac{1+0}{3+10}: 8.7\ , \frac{1+1}{3+10}: -6.4)$, where the probabilities are calculated based on the Carnap's method with $p_i^0 = \frac{1}{3}$ and $c = 3$. This scenario has a less clear prediction for the attractiveness of the sure prospect. In the present problem, adding good and bad outcomes, 9.6 and -6.4, to Option A will impact its attractiveness depending on the relative steepness of the lower parts of the probability weighting curves in the gain and in the loss domains. For instance, more overweighting of $\frac{1}{13}$ in the loss domain than in the gain domain decreases the attractiveness.

In the following analysis, the perceived set of possible outcomes is constructed based on the *context-dependent* or *comprehensive expectations*. Then, the parametric estimations will be done under the two stage model by using *Carnap, ignorance* and *diffuse priors*.

Data

I focus on the DFD and DFE sampling conditions in Erev et al. (2010). The study consists of an estimation data set and a competition data set. The two data sets are pooled in the current analysis. The pooled data set contains 40 subjects in the DFD condition and 80 subjects in the DFE condition. Each subject makes 60 choices in the DFD condition and 30 choices in the DFE condition. The problems always involve a choice between a sure outcome and a two-outcome risky prospect. Prospects were equally divided into gain, loss, and mixed domains. Subjects in the DFE condition were provided with minimal information about the content of the two prospects. Importantly, unlike in Glöckner et al. (2016), they do not know the number of possible outcomes in prospects, and therefore they lack the information about the certainty of an always-observed outcome. They make a single choice after an exploratory sampling stage. The readers are referred to Erev et al. for more details on the experimental design.

Analysis

The same estimation method is used as in the reanalysis of Glöckner et al. (2016). Utility is assumed to be the same across different conditions. Results from unconstrained estimations without assuming the same utility is reported in Appendix 3. Constrained estimations outperform unconstrained estimations based on BIC scores. Differences in unconstrained estimations are discussed.

Results

The estimation results are in Table 3.3. Comparison of BIC scores across different prior assumptions are in Table 3.4. The resulting probability weighting functions under the *Carnap prior* assumption are in Figure 3.3.

The parameter estimations under the DFE condition using the *diffuse prior* replicate the classic DFD-DFE gap. Specifically, the DFE condition indicates more likelihood sensitivity compared to the DFD condition. This means that the rare outcomes are less overweighted under DFE than under DFD. It should be noted that the underweighting of rare events claimed in the early DFE studies is not found here. This happens mainly because of the correction of the sampling error by using observed relative frequencies rather than unknown objective probabilities. There is no DFD-DFE gap with respect to other prospect theory parameters.

The estimation results based on *context-dependent expectations* imply different conclusions about the gap in likelihood sensitivity depending on the prior assumptions. Under the assumption of *ignorance prior,* the gap in likelihood sensitivity is persistent in the gain domain; and it is insignificant in the loss domain. Under the assumption of *Carnap prior,* the gap in likelihood sensitivity is insignificant in the gain domain; and it is significantly reversed in the loss domain. Carnap's constant c is marginally different from 0 (p=0.055). The estimated $c = 0.212$ means that for the median number of 5 draws, a never-observed rare outcome receives 2% probability. No significant gap is observed in other model parameters.

The estimations based on *comprehensive expectations* consistently indicate insignificant DFD-DFE gap in likelihood sensitivity both in the gain and in the loss domain. However, the trends indicate a reversal of the classic DFD-DFE gap as we observe less likelihood sensitivity under DFE than under DFD. The reversal is particularly noticeable in the case of *Carnap priors*. Under the assumption of *Carnap prior,* the parameter c is different from 0 ($p = 0.007$). The estimated $c = 0.994$ means that a never-observed rare outcome receives 5.5% probability for the median sample size. The trends in the elevation of probability weighting indicate less elevation under DFE than under DFD in the gain domain, and more elevation under DFE than under DFD in the loss domain. These imply ambiguity aversion in both domains.

Table 3.3 Estimation results with the data set of Erev et al. (2010)

		Context-Dependent Expectations		Comprehensive Expectations	
	Diffuse Prior (DFD/DFE)	Ignorance Prior (DFD/DFE)	Carnap Prior (DFD/DFE)	Ignorance Prior (DFD/DFE)	Carnap Prior (DFD/DFE)
α	0.926	0.851	0.904	0.857	0.892
λ	1.143	1.133	1.211	1.125	1.118
c	-/0	-/2	-/0.212	-/3	-/0.994**
δ^+	0.779/0.799	0.784/0.600	0.780/0.766	0.784/0.512	0.782/0.542
γ^+	0.594/0.891**	0.588/0.943**	0.581/0.778	0.590/0.532	0.595/0.387
δ^-	1.167/1.266	1.186/1.184	1.168/1.453	1.185/1.421	1.177/1.570
γ^-	0.592/0.921**	0.584/0.691	0.597/0.289**	0.584/0.453	0.587/0.345
σ	1.054/1.424	1.246/1.798	1.071/1.671	1.238/2.569	1.153/2.177
N	4800	4800	4800	4800	4800
LL	-2516.465	-2590.941	-2505.980	-2524.151	-2505.262
BIC	5134.647	5283.599	5122.152	5150.019	5120.717

Notes: Stars indicate DFD – DFE gap. $*p < 0.05$. $**p < 0.01$. $***p < 0.001$. *The first numbers in the cells indicate the estimated parameters for the DFD condition.*

The two stage model with *Carnap prior* account for the data better than the model with *diffuse prior*. Interpreting from Bayes Factors in Table 3.3, there is very strong evidence in favor of *Carnap priors* over *diffuse priors* under both accounts. The two-stage model with ignorance priors did considerable worse than the models with *Carnap* and *diffuse priors.* With *Carnap priors*, *comprehensive expectations* perform only slightly better than *context-dependent expectations* ($BayesFactor = 2.419$).

Table 3.4. Comparison of *Diffuse, Ignorance* and *Carnap Priors* under DFE

		Context-Dependent Expectations	Comprehensive Expectations
Diffuse Prior	LL	-1202.812	-1202.812
	BIC	2444.541	2444.541
Ignorance Prior	LL	-1274.809	-1208.356
	BIC	2588.534	2455.628
Carnap Prior	LL	-1191.92	-1191.037
	BIC	2430.54	2428.773
N		2400	2400
$BF_{Diffuse-Ignorance}$		$> 10^{31}$	255.571
$BF_{Diffuse-Carnap}$		$< 10^{-3}$	$< 10^{-3}$
$BF_{Ignorance-Carnap}$		$< 10^{-34}$	$< 10^{-5}$

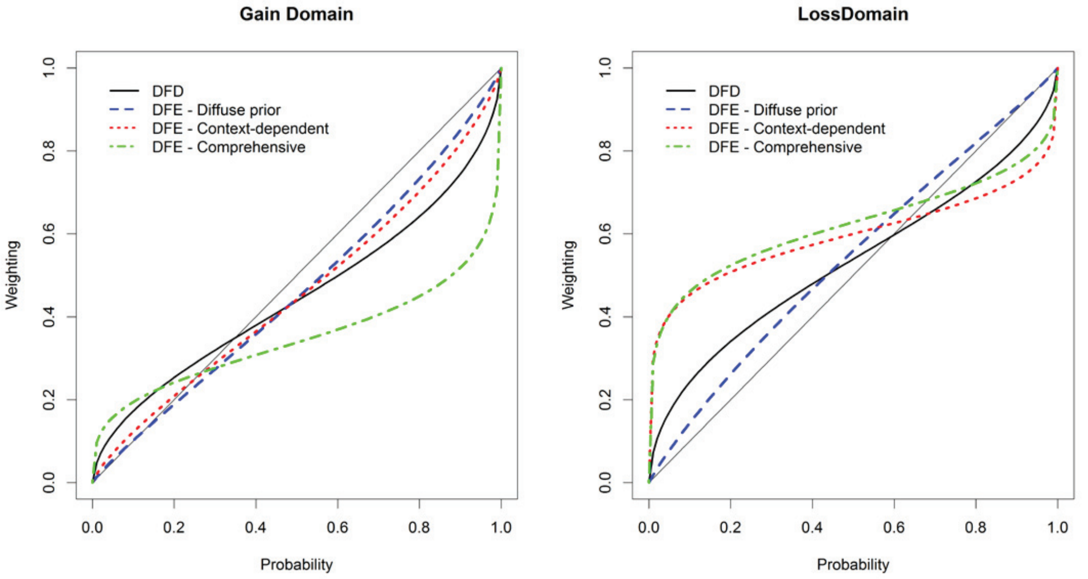
Notes: $BF_{Null-Alternative}$ *indicates relative evidence for the null hypothesis against the alternative. BF values higher than 1 indicate evidence in favor of the null hypothesis. BF values smaller than 1 indicate evidence in favor of the alternative hypothesis.*

Discussion

The reanalysis of Erev et al. (2010) indicates that the classic DFD-DFE gap can be reversed when the prior beliefs are taken into account. The reversal is consistent with the findings of Glöckner et al. (2016). Different from our estimations with the data sets of Glöckner et al., *Carnap prior* performed better than the *ignorance priors*. This discrepancy can be attributed to the differences in the experimental procedures. In particular, in the studies reported in Glöckner et al., subjects were informed about the number of possible outcomes in prospects. This might have reinforced the use of ignorance priors. No such information was available to the subjects in the experiments of Erev et al. (2011).

Our main conclusions are replicated in the unconstrained estimations (Appendix 3, Table A3.3). Furthermore, here, we also observe more utility curvature under DFE than under DFD under the assumption of *ignorance priors*. However, the assumption of *ignorance prior* accounts for the data considerably worse than *Carnap priors* and no significant gap in utility is observed under the assumption of *Carnap priors* (Table A3.4).

Figure 3.3 Probability weighting functions from the data set of Erev et al. (2010)



Notes: Solid black lines show probability weighting under DFD. Dashed blue lines show probability weighting under DFE when the diffuse prior is used. Dotted red lines show probability weighting under DFE based context dependent expectations. Dot-dash green lines show probability weighting under DFE based on comprehensive expectations.

3.4 General Discussion

DFD-DFE gap

The weighing of uncertainty under DFE concerns both probabilistic inference and probability weighting. The aforementioned two-stage model gives a refined analysis of probability weighting under DFE by modelling probabilistic inference with a Bayesian method of updating. The findings with the two-stage model showed that the rational updating of symmetric priors accommodates the commonly found regressive probability estimations, and it explains a considerable part of the reversed DFD-DFE gap. The remaining gap is explained by the source dependent probability weighting. The reanalysis of the classic DFD-DFE gap confirmed the validity of the two-stage model by revealing the persistence of the enhanced likelihood insensitivity under DFE, which is consistent with the reversed DFD-DFE gap.

There are two possible factors that may give rise to different probability weighting under DFD and DFE: sampling experience and ambiguity. Contrary to the reversed DFD-DFE gap, previous studies by van de Kuilen & Wakker (2006), van de Kuilen (2009) and Humphrey (2006) on experienced risk; and studies by Ert & Trautmann (2014) and Kemel & Travers (2016) on experienced ambiguity report that the sampling experience reduces, rather than enhances, likelihood insensitivity. While the experienced ambiguity in the present study can explain the discrepancy with the former studies on experienced risk, a possible reason for the discrepancy with the previous studies on experienced ambiguity can be different methodologies used in these studies. Kemel & Travers (2016) uses certainty equivalents of experienced prospects, rather than binary choice data, to elicit PT parameters. Therefore, their method requires comparisons of experienced prospects with explicitly described certain outcomes. Similarly, Ert & Trautmann (2014) focuses on choices between experienced ambiguous prospects and described risky prospects. Future research can clarify the impact of sampling experience when the choice is between two experienced ambiguous prospects as in Glöckner et al. (2016) and Erev et al. (2010).

The enhanced likelihood insensitivity is a common finding in the ambiguity literature (Wakker, 2010, p. 292). This residual deviation from Bayesian rationality can be explained by perceived ambiguity in estimated probabilities (Dimmock, Kouwenberg, Mitchell, & Peijnenburg, 2015). Specifically, acknowledging the uncertainty about his probability estimation, the decision maker can consider a range of possible probabilities around his estimate. However, the range is very likely to be asymmetric around small probabilities such as 5% because there is much more room between 5% and 100% than between 0% and 5%. As a result, a decision maker who is weighting the rare outcome with an average of minimum and maximum of the perceived range of probabilities – as in the $\alpha - maxmin$ model (Hurwicz 1951; Luce & Raiffa 1957) – is likely to assign a weight higher than the small probability estimate of 5%. Such multiple prior accounts of ambiguity are common in the behavioral economics literature (Baillon, Bleichrodt, Keskin, L'Haridon, & Li , in press; Chateauneuf, Eichberger, & Grant, 2007; Ghirardato, Maccheroni, & Marinacci, 2004; Gilboa & Schmeidler, 1989; Marinacci, 2015).

The Bayesian method of updating

This paper uses a tractable Bayesian updating method in analyzing subjective probabilities under DFE. Despite its promising performance in accounting for the previous empirical findings, the descriptive validity of the method can be questioned. First, as in any Bayesian approach, in Carnap's method, the strength and the weight of evidence, i.e. $\frac{n_i}{N}$ and N, receive equal emphasis in evaluation. Notwithstanding this normative property, an influential study by Griffin & Tversky (1992) indicates that people systematically focus on the strength of evidence while paying insufficient attention to the credibility. This tendency results in either representativeness (overconfidence) or conservatism (underconfidence) in probability judgments. A recent study by Kvam & Pleskac (2016) replicates these findings in an environment where the information is accumulated by observation similar to the sampling paradigm.

Although Carnap's method cannot differentiate the relative impact of the strength and the weight of evidence, biases similar to representativeness and conservatism can still be observed by negative values of c. In particular, $-N < c < 0$ implies too much updating in the direction of sample information resembling to representativeness, and $c < -N$ implies too much updating in the direction of prior beliefs resembling to an extreme case of conservatism.

Second, the Bayesian updating method assumes perfect memory, and hence, no recency effects in probability judgments. Although some early studies report recency effects in DFE, the evidence is still mixed (see the comprehensive meta-analysis by Wulff et al. 2016). Nonetheless, there are also ways to capture these effects within Carnap's formula. One way is to simply use the sampling information from the second half of the observed sequence of outcomes as if the first half has been forgotten. Such modelling of recency effects is discussed in Ashby & Rakow (2014) and in Wulff & Pachur (2016) (the sliding window model). Another way is to assign different weights to the relative frequencies observed in the first and in the second half of the sequence. For example, taking N_1 and N_2, the sample sizes of the first and second half the observations with $N_1 + N_2 = N$, one can use $N_1' = N_1$ and $N_2' = \varphi N_2$ where $\varphi > 1$ implies recency.

3.5 Conclusion

The preceding literature on DFE has extensively argued for the role of sampling error and ambiguity in the DFD-DFE gap. This study points to another important factor, being prior beliefs. The Bayesian approach, taken as a working hypothesis in this study, is shown to be useful in resolving the controversy about the gap by offering a tractable way to analyze prior beliefs. Importantly, the DFD-DFE gap is almost fully accommodated when prior beliefs are taken into account. The residual gap is explained by perceived ambiguity. Bayesian updating does remarkably well in explaining experimental data despite its normative nature. A promising topic for future research will concern more descriptive methods for analyzing beliefs under DFE.

Appendix 3

Table A3.1 Unconstrained estimations with the data sets in Glöckner et al. (2016)

		Glöckner et al. (2012)	Experiment 1	Experiment 2	Experiment 3	Pooled
α	DFD	0.652	1.058	0.617	0.954	0.850
	DFE – Diffuse prior	0.642	0.853	0.836/1.015**	0.991	0.796
	DFE – Ignorance prior	0.644	0.871	0.845/1.040**	0.981	0.800
	DFE – Carnap prior	0.642	0.890	0.851/1.028**	0.972	0.801
c	DFD	-	-	-	-	-
	DFE – Diffuse prior	0	0	0	0	0
	DFE – Ignorance prior	2	2	2	2	2
	DFE – Carnap prior	-1.602***	6.456	6.661/0.881	10.117	2.768
δ^+	DFD	0.553	0.320	0.695	0.637	0.511
	DFE – Diffuse prior	0.561	0.392	0.432*/0.527	0.661	0.541
	DFE – Ignorance prior	0.565	0.386	0.433*/0.517	0.673	0.542
	DFE – Carnap prior	0.541	0.376	0.431*/0.522	0.684	0.542
γ^+	DFD	0.732	0.736	0.961	0.559	0.810
	DFE – Diffuse prior	0.560*	0.552*	0.521***/0.595***	0.423*	0.536***
	DFE – Ignorance prior	0.670	0.684	0.634**/0.707**	0.502	0.645***
	DFE – Carnap prior	0.427**	0.919	0.833/0.648*	0.733	0.682
δ^-	DFD				1.074	1.091
	DFE – Diffuse prior				1.211	1.341
	DFE – Ignorance prior				1.243	1.357
	DFE – Carnap prior				1.288	1.361
γ^-	DFD				0.856	0.787
	DFE – Diffuse prior				0.462***	0.371***
	DFE – Ignorance prior				0.544**	0.443**
	DFE – Carnap prior				0.793	0.467*
λ	DFD				1.001	1.237
	DFE – Diffuse prior				0.916	1.068
	DFE – Ignorance prior				0.917	1.081
	DFE – Carnap prior				0.920	1.084
σ	DFD	2.117	0.639	2.123	1.512	1.099
	DFE – Diffuse prior	2.045	0.865	1.023/0.625*	1.268	1.232
	DFE – Ignorance prior	1.997	0.825	1.009/0.578*	1.308	1.218
	DFE – Carnap prior	2.066	0.778	0.991/0.601*	1.326	1.212
LL	Diffuse prior	-1143.916	-1434.251	-2831.996	-2652.874	-8281.178
	Ignorance prior	-1147.327	-1431.981	-2826.839	-2648.487	-8273.231
	Carnap prior	-1140.565	-1431.107	-2824.415	-2645.445	-8272.874
BIC	Diffuse prior	2350.680	2932.683	5768.568	5425.181	16698.56
	Ignorance prior	2357.502	2928.142	5758.254	5416.407	16682.66
	Carnap prior	2351.833	2934.417	5770.836	5418.853	16691.68

Notes: Stars indicate DFD – DFE gap. $*\,p < 0.05$. $**\,p < 0.01$. $***\,p < 0.001$. *In the column for Experiment 2, the first numbers are based on DFE-informed condition and the second numbers are based on DFE-uninformed condition. There is no significant difference across DFE-informed and DFE-uninformed conditions.*

Table A3.2 Comparison of *Diffuse, Ignorance* and *Carnap Priors* under DFE, based on unconstrained estimations with the data sets in Glöckner et al. (2016)

		Glöckner et al. (2012)	Experiment 1	Experiment 2 DFE-informed	Experiment 2 DFE-uninformed	Experiment 3	Pooled
Diffuse Prior	LL	-558.414	-774.531	-626.225	-657.932	-1340.263	-4056.058
	BIC	1145.456	1578.653	1281.483	1345.096	2735.529	8175.377
Ignorance Prior	LL	-561.825	-772.261	-621.113	-657.887	-1335.877	-4048.111
	BIC	1152.279	1574.111	1271.259	1345.006	2726.755	8159.483
Carnap Prior	LL	-555.063	-771.387	-618.974	-657.602	-1332.834	-4047.754
	BIC	1145.911	1579.762	1274.240	1351.743	2728.528	8167.807
N		1283	1632	1420	1492	2585	8412
$BF_{Diffuse-Ignorance}$		30.311	0.103	0.006	0.956	0.012	0.0003
$BF_{Diffuse-Carnap}$		1.255	1.741	0.027	27.757	0.030	0.023
$BF_{Ignorance-Carnap}$		0.041	16.869	4.439	29.035	2.427	64.199

Notes: $BF_{Null-Alternative}$ *indicates relative evidence for the null hypothesis against the alternative. BF values higher than 1 indicate evidence in favor of the null hypothesis. BF values smaller than 1 indicate evidence in favor of the alternative hypothesis.*

Table A3.3 Unconstrained estimations with the data sets of Erev et al. (2010)

		Context-Dependent Expectations	Comprehensive Expectations
α	DFD	0.932	0.932
	DFE – Diffuse prior	0.917	0.917
	DFE – Ignorance prior	0.720**	0.734**
	DFE – Carnap prior	0.859	0.815
c	DFD	-	-
	DFE – Diffuse prior	0	0
	DFE – Ignorance prior	2	3
	DFE – Carnap prior	0.202	1.038*
δ^+	DFD	0.779	0.779
	DFE – Diffuse prior	0.803	0.803
	DFE – Ignorance prior	0.670	0.501
	DFE – Carnap prior	0.790	0.512
γ^+	DFD	0.599	0.599
	DFE – Diffuse prior	0.891*	0.891*
	DFE – Ignorance prior	0.925*	0.461
	DFE – Carnap prior	0.795	0.338
δ^-	DFD	1.167	1.167
	DFE – Diffuse prior	1.268	1.268
	DFE – Ignorance prior	1.301	1.565
	DFE – Carnap prior	1.504	1.686
γ^-	DFD	0.589	0.589
	DFE – Diffuse prior	0.924**	0.924**
	DFE – Ignorance prior	0.626	0.403
	DFE – Carnap prior	0.265**	0.307
λ	DFD	1.111	1.111
	DFE – Diffuse prior	1.158	1.158
	DFE – Ignorance prior	1.179	1.145
	DFE – Carnap prior	1.283	1.123
σ	DFD	1.055	1.055
	DFE – Diffuse prior	1.450	1.450
	DFE – Ignorance prior	2.681**	3.853*
	DFE – Carnap prior	1.874	2.871
LL	Diffuse prior	-2516.406	-2516.406
	Ignorance prior	-2584.025	-2518.27
	Carnap prior	-2505.965	-2503.452
BIC	Diffuse prior	5151.481	5151.481
	Ignorance prior	5286.72	5155.209
	Carnap prior	5137.076	5134.05

Notes: Stars indicate DFD – DFE gap. $*p < 0.05$. $**p < 0.01$. $***p < 0.001$.

Table A3.4 Comparison of *Diffuse, Ignorance* and *Carnap Priors* under DFE, based on unconstrained estimations with the data set of Erev et al. (2010)

		Context-Dependent Expectations	Comprehensive Expectations
Diffuse Prior	LL	-1202.785	-1202.785
	BIC	2460.053	2460.053
Ignorance Prior	LL	-1270.405	-1204.659
	BIC	2595.292	2463.781
Carnap Prior	LL	-1191.344	-1189.832
	BIC	2444.955	2441.929
N		2400	2400
$BF_{Diffuse-Ignorance}$		$> 10^{29}$	6.449
$BF_{Diffuse-Carnap}$		$< 10^{-3}$	$< 10^{-3}$
$BF_{Ignorance-Carnap}$		$< 10^{-32}$	$< 10^{-4}$

Notes: $BF_{Null-Alternative}$ *indicates relative evidence for the null hypothesis against the alternative. BF values higher than 1 indicate evidence in favor of the null hypothesis. BF values smaller than 1 indicate evidence in favor of the alternative hypothesis.*

Chapter 4

Signal Perception and Belief Updating

with Aurelien Baillon, Emmanuel Kemel, and Chen Li

4.1 Introduction

In standard economic models, from game theory to macroeconomics, decision makers incorporate new information using the rational gold standard of belief updating: the Bayes rule. Yet, studies from the psychology literature highlighted regular deviations from Bayesian updating. Famous examples are the confirmatory bias (see Oswald & Grosjean, 2004, for a review), in which people tend to neglect or even misinterpret signals contradicting their prior beliefs, and the conservatism bias (Phillips & Edwards, 1966; Edwards, 1968), in which people fail to sufficiently incorporate new information, resulting in posteriors that are *too* close to their priors.

Since the end of the 1990s, economists have proposed models to incorporate deviations from Bayesian updating. For instance, Rabin and Schrag (1999) modelled confirmatory bias as decision makers misreading signals that contradict their priors, which may give rise to behavioral biases such as overconfidence. Epstein (2006) provided an axiomatic foundation for non-Bayesian updating through a retroactively changing prior. Wilson (2014) modelled a decision maker with bounded memory, which can lead to the emergence of confirmatory bias and conservatism in belief formation.

Alternatively, the literature on *motivated beliefs* (see Bénabou & Tirole, 2016, for a review) models deviations from Bayesian updating through the decision maker's tradeoff between the accuracy and desirability of their beliefs. Strategies to cope with this tradeoff includes reality denial and wishful thinking. The motivated-belief approach is appealing in situations when people are motivated to attach values to their beliefs, such as when they think of their own abilities or of important aspects of their life (Bénabou & Tirole, 2002; 2006). It is not obvious though whether it would predict deviations from Bayesian updating when beliefs concern external, 'neutral' factors.

In this paper, we propose a theory of signal perception to model belief updating. We introduce two indices that are derived from the difference between people's *perceived* *signals*, revealed from their choices, and the actual signals received. Rabin and Schrag (1999) models the confirmatory bias as a probability of *misreading* a contradicting signal as confirming. Our first index, q, captures the same confirmatory tendency in belief updating. In particular, when $0 < q < 1$, it has the same interpretation as in Rabin and Schrag's model. Empirical evidence shows that people can also exhibit the opposite pattern (Eil & Rao, 2011). Our index captures disconfirmatory bias as well. Our second index p captures people's tendency of *missing* a signal, regardless of its agreement with the prior. Our model, combining q and p, is a portable extension of Bayesian updating in the sense of Rabin (2013). It can be incorporated in any model from macroeconomics or game theory by re-coding actual signals into perceived signals using transformations based on q and p.

After presenting our model, we show how perceived signals can be revealed from choices. In an experiment, we elicited subjects' beliefs and obtained a structural estimation of the indices, demonstrating the tractability of the model. We found clear evidence for both conservatism and confirmatory bias, showing that deviations of Bayesian updating may occur even in the absence of clear motivation. On average, subjects missed 65% of the signals and misread 17% of the signals contradicting their prior beliefs. We further explored factors influencing the indices. Consistent with previous findings (Griffin & Tversky, 1992), moderately informative signals led to more conservatism.

4.2 Perceived signal theory

4.2.1 Setup and perceived signals

We model a simple signal setup, in which a decision maker faces a mechanism producing independent and identically distributed binary signals. It produces *successes* with an unknown probability s (and *failures* with probability $1 - s$). The decision maker is interested in learning about the success rate s. We consider an initial state of ignorance, represented by a uniform probability measure $Prob(s)$ defined over $S \subset (0, 1)$. We assume that the support $\mathcal{S}$ is symmetric around 0.5, i.e. $p \in \mathcal{S} \Rightarrow (1 - p) \in \mathcal{S}$.

Before receiving a specific set of signals, the decision maker has a prior sample with α_0 successes and β_0 failures in his memory. Hence, his *prior beliefs* are $\Lambda(s;\ \alpha_0, \beta_0) = Prob(s|\alpha_0, \beta_0)$, abbreviated as $\Lambda(\alpha_0, \beta_0)$. When $\alpha_0 = \beta_0$, the mean of $\Lambda(\alpha_0, \beta_0)$ is equal to 0.5. The initial state of ignorance is a hypothetical construct that allows us to interpret the decision maker's beliefs in terms of signals. Departures from uniformity in prior beliefs are modelled by (possibly hypothetical) signals in the decision maker's mind.

After receiving a sequence of signals, his *posterior beliefs* becomes $\Lambda(\alpha_1, \beta_1)$. Define $\alpha = \alpha_1 - \alpha_0, \beta = \beta_1 - \beta_0$, and $\eta = \alpha + \beta$. These parameters measure how much the decision maker has updated his beliefs and therefore, how many signals (successes, failures) he has perceived. We call η the *perceived number of signals*, α the *perceived number of successes*, and β the *perceived number of failures*.

Consider a Bayesian updater with a uniform prior over $(0, 1)$, which is equivalent to $Beta(1, 1)$. If he observes a success, his posterior will also be a beta distribution, given by $Beta(2, 1)$. It would be $Beta(1, 2)$, had he observed a failure. After each success (failure), the first (second) parameter of the beta distribution is incremented by one. If the prior belief is $Beta(\alpha_0, \beta_0)$, with α_0 and β_0 possibly different from 1, then the expected probability of success is given by $\frac{\alpha_0}{\eta_0}$ with $\eta_0 = \alpha_0 + \beta_0$. Hence, the decision maker will expect success and failure to be equally likely iff $\alpha_0 = \beta_0$. In our application, we will use such a setting with beta distributions but the theory below does not rely on it.[14]

For a Bayesian updater, all signals are perceived without distortion: receiving n signals consisting of a successes and b failures implies $\alpha = a, \beta = b$, and $\eta = n$. However, this is not true for non-Bayesian updaters. Deviations from Bayesian updating can therefore be captured by differences between people's perceived signals (α, β, and η) and the actual signals they observe (a, b, and n). We study two sources of deviations: confirmatory bias and conservatism bias. Confirmatory bias captures people's tendency to "misread evidence as additional support for initial hypotheses" (Rabin & Schrag, 1999), whereas conservatism captures people's tendency to miss evidence and to not update enough their beliefs, without discriminating different types of signals.

4.2.2 Confirmatory bias

Following Rabin and Schrag (1999), we model the *confirmatory bias* as the probability q_c to misread a contradicting signal as confirming prior expectations. By symmetry, the opposite bias, that we called *disconfirmatory bias*, can be modelled as the probability q_d to misread a signal as contradicting prior expectations. If, according to the decision maker's

[14] For instance support $\mathcal{S}$ can be discrete

prior belief, successes are more likely than failures (i.e. $\alpha_0 > \beta_0$), the confirmatory bias gives:

$$\begin{cases} \alpha = a + q_c b \\ \beta = (1 - q_c)b \end{cases}, \tag{4.1}$$

whereas the disconfirmatory bias gives

$$\begin{cases} \alpha = (1 - q_d)a \\ \beta = b + q_d a \end{cases}. \tag{4.2}$$

If, according to the decision maker's prior belief, successes are less likely than failures (i.e. $\alpha_0 < \beta_0$), the confirmatory bias gives:

$$\begin{cases} \alpha = (1 - q_c)a \\ \beta = b + q_c a \end{cases}, \tag{4.3}$$

whereas the disconfirmatory bias gives

$$\begin{cases} \alpha = a + q_d b \\ \beta = (1 - q_d)b \end{cases}. \tag{4.4}$$

If successes and failures are equally likely according to the decision maker's prior belief, the perceived number of success and failure is not affected by confirmatory bias.

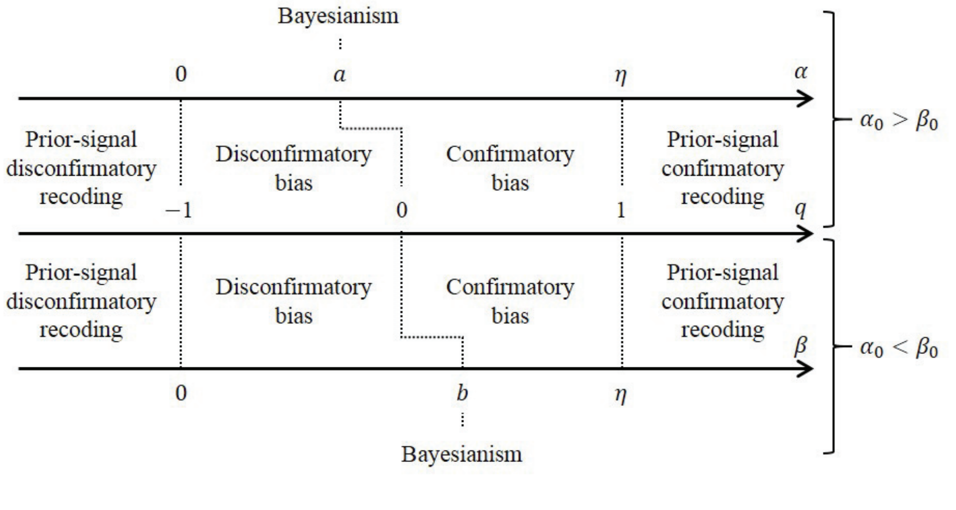
From observing perceived signals, either q_c or q_d can be determined whenever $\alpha_0 \neq \beta_0$. Consider the case $\alpha_0 > \beta_0$. If $a \leq \alpha \leq \eta$, there is evidence for confirmatory bias and q_c can be computed. In practice, we may even observe $q_c > 1$ when $\eta < \alpha$ (and therefore $\beta < 0$). In such a case, q_c is not a probability anymore but can still be used as an index of confirmatory bias. The case $q_c > 1$ indicates that the decision maker exhibits an extreme form of confirmatory bias, in which he even recodes the signals from his prior. We call such a case *prior-signal confirmatory recoding*. Figure 4.1 depicts all possible cases. The interpretation of the decision maker's perceived signals depends on his prior beliefs (α_0 and β_0) and his perceived number of successes α.

Moreover, we can combine q_c and q_d into a unique index of confirmatory bias q defined as:

$$q = \begin{cases} q_c & if\ (\alpha_0 > \beta_0\ and\ a \leq \alpha)\ or\ (\alpha_0 < \beta_0\ and\ b \leq \beta) \\ -q_d & if\ (\alpha_0 > \beta_0\ and\ a \geq \alpha)\ or\ (\alpha_0 < \beta_0\ and\ b \geq \beta) \end{cases} \quad . \tag{4.5}$$

Values of q in $[0,1]$ can be directly interpreted as probabilities to misread signal in a confirmatory way and values in $[-1,0]$ as minus probabilities to misread signal in a disconfirmatory way. The global index q is useful for empirical purposes. For instance, its distribution for the population can be estimated at once, without separating confirmatory biases from disconfirmatory biases (as is done for other attitude measures such as risk aversion).

Figure 4.1 Interpretation of q and relationship with α



4.2.3 Conservatism bias

We expand the confirmatory bias model by also considering a conservative decision maker's tendency to ignore signals. In this subsection, we introduce a measure of conservatism bias, and in the next subsection, we present a model in which both biases are combined in one model.

A conservative decision maker places too little weight on the sample information while updating and thereby tends to ignore some of the relevant information. We model the *conservatism bias* as a probability p to miss a signal. Hence, $\eta = (1-p)n$. The conservatism bias can affect both types of signals indistinguishably, leading to $\alpha = (1-p)a$ and $\beta = (1-p)b$. Bayesian updating implies $p = 0$. If $p = 1$, there is no updating at all.

Interestingly, p can also be interpreted if it lies outside the unit interval, but obviously not as a probability. The case $p > 1$ captures situations where the perceived number of signal is negative, suggesting that the decision maker received information that undermined his prior. For instance, a decision maker whose prior was too extreme, expecting successes almost exclusively, might be less confident in his beliefs after observing a few failures. In our perceived signal theory, such behavior corresponds to *prior signal destruction*.

By contrast, $p < 0$ means that the decision maker perceived too many signals. It can be further illustrated in the case of a Beta distribution. The posterior mean $\frac{\alpha_0+\alpha}{\eta_0+\eta}$ can be decomposed in terms of prior mean and sample mean:

$$\begin{aligned}\frac{\alpha_0+\alpha}{\eta_0+\eta} &= \frac{\alpha_0+(1-p)a}{\eta_0+(1-p)n} \\ &= \frac{\eta_0}{\eta_0+(1-p)n}\cdot\frac{\alpha_0}{\eta_0}+\frac{(1-p)n}{\eta_0+(1-p)n}.\frac{a}{n} \\ &= \frac{\eta_0}{\eta_0+(1-p)n}\cdot prior\ mean+\frac{(1-p)n}{\eta_0+(1-p)n}\cdot sample\ mean\end{aligned} \tag{4.6}$$

Bayes rule requires $p = 0$, i.e. the actual and the perceived number of signals match. A positive $p\,(< 1)$ decreases the impact of the sample mean, implying conservatism. The decision maker underweights the sample information and overweights the prior information. Negative p corresponds to *base rate neglect*, the decision maker assigning too much weight to the sample and neglecting his prior beliefs. Such behavior can be explained by the *representativeness heuristic* (Tversky & Kahneman, 1974), when decision makers assume that a sample must resemble the process it originates from and therefore tend to equate the process mean too much with the sample mean.

Figure 4.2 Interpretation of p and relationship q

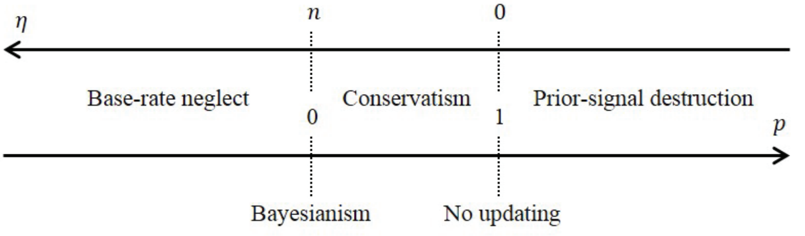


Figure 4.2 depicts the relationship between the perceived number of signal η and the conservatism index p. It shows that p is a simple rescaling of η such that p is independent of the actual sample size n.

4.2.4 Combining biases

In the combined model, the decision maker may miss signals (conservatism bias) and then misread those he did not miss (confirmatory bias). If, according to the decision maker's prior belief, successes are more likely than failures ($\alpha_0 > \beta_0$), the confirmatory bias in presence of conservatism bias gives (replacing q_c by q):

$$\begin{cases} \alpha = (1-p)a + q(1-p)b \\ \beta = (1-q)(1-p)b \end{cases}, \tag{4.7}$$

whereas the disconfirmatory bias gives (replacing q_d by $-q$)

$$\begin{cases} \alpha = (1+q)(1-p)a \\ \beta = (1-p)b - q(1-p)a \end{cases}. \tag{4.8}$$

The case $\alpha_0 < \beta_0$ is symmetric. If successes and failures were equally likely according to the decision maker's prior belief:

$$\begin{cases} \alpha = (1-p)a \\ \beta = (1-p)b \end{cases}. \tag{4.9}$$

In terms of observability, p can always be obtained by comparing η with n. Further, if $\alpha_0 \neq \beta_0$ and $p \neq 1$, then q can be obtained by first correcting a and b for conservatism (multiplying them by $1 - p$) and then applying the adequate equation of (dis)confirmatory bias. If after a first round of signals, the decision maker receives a second round of signals, p and q can be determined again using the posterior of the first round as the prior of the second round.

4.2.5 Measuring informativeness

Literature shows that deviations from Bayesian updating depend on various situational factors such as the strength of evidence. Griffin and Tversky (1992) find that moderate signals lead to insufficient updating while extreme signals lead to overreaction. The same set of signals may be deemed extremely informative by a decision maker but less so by another. The informativeness of signals thus depends both on the signals themselves and the prior of the decision maker. Hence a measure of informativeness should depend on $\alpha_0, \beta_0, a,$ and b.

We define our measure of informativeness as the *information gain (IG)*, also known as relative entropy or Kullback-Leibler divergence, between the prior $\Lambda(\alpha_0, \beta_0)$ and the posterior the decision maker would have if he were Bayesian $\Lambda(\alpha_0 + a, \beta_0 + b)$. Let $g(s)$ be the density function of prior, and $h(s)$ be the density of the Bayesian posterior. The IG is calculated as:

$$IG(\alpha_0, \beta_0, a, b) = \int_{[0,1]} h(s) log \frac{h(s)}{g(s)} ds. \tag{4.10}$$

The IG measure captures how much the signals should influence the decision maker's beliefs. It allows us to examine the impact of signal informativeness on belief updating biases.

4.3 Revealing perception through choices

To reveal people's perception of signals, it is necessary to make their beliefs observable. Belief elicitation methods in the literature, such as proper scoring rules (see Schotter & Trevino, 2014, for a survey in economics), often rely on the descriptive validity of expected value or expected utility to reveal people's true beliefs. In this paper, we consider two methods that do not rely on expected utility.

We are interested in the decision maker's belief about the unknown success rate s. Let $\mathcal{P}$ denote the σ-algebra on $(0, 1)$, which is the domain of s. *Events*, $E \in \mathcal{P}$, of interest to the decision maker are subsets of $(0, 1)$. The decision maker faces *(binary) acts*, denoted by $\gamma_E\delta$, which pays a positive money amount γ if event E happens and δ otherwise. The decision maker also faces *(binary) lotteries* $\gamma_\lambda\delta$, yielding γ with probability λ and δ otherwise.

Assume that the decision maker whose behavior towards lotteries can be represented by a function V satisfying first order stochastic dominance. The function V need not be expected utility and it therefore allows for deviations from expected utility such as in the paradoxes suggested by Allais (1953). The decision maker is *probabilistically sophisticated* (Machina & Schmeidler, 1992) if his behavior towards acts can be entirely explained by V and a probability measure Λ over $\mathcal{P}$. In other words, the assumption of a probabilistically sophisticated decision maker guarantees that choices are consistent with a probability measure and therefore is a sufficient condition to observe beliefs from choices.

We present two methods to elicit Λ irrespective of V. The first method to observe belief involves measuring *matching probabilities*, i.e. λ such that $\gamma_E\delta \sim \gamma_\lambda\delta$. Under probabilistic sophistication, this indifference implies $V(\gamma_{\Lambda(E)}\delta) = V(\gamma_\lambda\delta)$ and thus, $\Lambda(E) = \lambda$, thereby revealing beliefs. Many studies used *matching probabilities* to elicit people's beliefs (Raiffa, 1968; Spetzler & Stael von Holstein, 1975; Holt, 2007; Karni, 2009). The second method we consider involves elicitation of *exchangeable events*, events E and F, such that $\gamma_E\delta \sim \gamma_F\delta$. If probabilistic sophistication holds, the elicited indifference implies, $V(\gamma_{\Lambda(E)}\delta) = V(\gamma_{\Lambda(F)}\delta)$, and thus, $\Lambda(E) = \Lambda(F)$, providing constraints on the belief function. For instance, if they are complementary, then $\Lambda(E) = \Lambda(F) = \frac{1}{2}$. This method is based on the original idea of Ramsey (1931) (called ethically neutral events) and of De Finetti (1937) and has been long-known in decision analysis (Raiffa, 1968; Spetzler & Stael von Holstein, 1975). Recent experimental implementations can be found in Baillon (2008) and Abdellaoui, Baillon, Placido, & Wakker (2011).

Both methods have advantages and are therefore implemented in our experiment. Matching probabilities directly reveals the probability of an event whereas exchangeable events only reveal that two events are equally likely. Yet, matching probabilities require that the function V is the same for lotteries and for acts. If the decision maker tend to prefer lotteries to acts, exhibiting ambiguity aversion (Ellsberg, 1961), matching probabilities may be biased. Eliciting exchangeable events, which do not require the use of lotteries, is robust to this problem.[15] Implementing both methods will allow us to assess the possible impact of ambiguity attitude.

For empirical tractability, we assume that decision makers' beliefs follow a beta distribution. The beta family is both natural to model beliefs over a success rate and very tractable. Beta distributions are flexible and can take a wide array of shapes with different locations and dispersion for different parameters. Before and after they receive a set of signals, we elicit their priors and posteriors using the methods described above. We then estimate their perceived signals and are able to construct measures of their conservatism and confirmatory biases.

4.4 Experimental design

Subjects

Seven experimental sessions were conducted at the Erasmus School of Economics Rotterdam. The number of participants in each session varied between 20 and 27, summing up to 157 in total. Subjects were bachelor and master students at Erasmus University Rotterdam, with an average age of 21.3. Each session lasted one hour and fifteen minutes including instructions and payment.

Stimuli

During the experiment, subjects faced choice situations about acts whose payoffs depended on the actual color composition of a spinning wheel. The spinning wheel was covered by two (and only two) colors: yellow and brown. The color composition was randomly drawn from an opaque bag at the beginning of the experiment in front of all subjects by an *implementer* – one randomly selected subject.

The experiment consisted of alternating periods of choice and sampling (see Figure 4.3 for the flow). It started with a choice period in which subjects made choices without any knowledge about the color composition of the wheel. Then, the implementer spinned

[15] Ambiguity is sometimes assumed to be equivalent to the absence of probabilistic beliefs. As demonstrated theoretically by Chew and Sagi (2008) and empirically by Abdellaoui et al. (2011), one can preserve the existence of a belief function expressed in probabilistic terms and allow for the Ellsberg paradox. The decision maker is *within-source probabilistically sophisticated* if there exists a probability measure Λ defined over $\mathcal{P}$ and a function W satisfying first order stochastic dominance such that $\gamma_E\delta$ is evaluated $W(\gamma_{\Lambda(E)}\delta)$. Under this model, exchangeable events E and F still satisfy $\Lambda(E) = \Lambda(F)$.

the wheel three times and reported the resulting colors. Having acquired this new information, subjects made choices in the same choice situations (but potentially in different orders) again. The same procedure was repeated two more times.

Figure 4.3 Experimental flow

The color composition of the wheel stayed the same and unknown throughout the experiment, which means that in later choice periods, subjects made choices based on accumulated knowledge about the same wheel. For example, the last questionnaire was filled relying on the information of nine spins in total.

Matching probabilities

Figure 4.4 presents a choice list to elicit a matching probability. In each choice question, subjects had to choose between option W(heel) whose payoff depended on the actual color composition of the same spinning wheel, and option C(ard) whose payoff depended on a random draw from a deck of four cards of different suits: aces with heart, diamond, club, and spade, each with 25% probability.

The choice in the first line was pre-ticked for the subjects by the experimenters, as in this case, option C dominates option W since the proportion of brown cannot be 0% (otherwise there is only one color on the wheel). Similarly, the last line was also pre-ticked. Subjects were informed that as they move down the list, option W became better while option C stayed the same. Therefore, at one point, they may switch from preferring option C to option W.

The subjects' switching pattern in Figure 4.4 gave an interval $[y_{0.25}^{-}, y_{0.25}^{+}]$ for $y_{0.25}$ such that $20_{[0,y_{0.25}^{-}]}0 \prec 20_{0.25}0$ and $20_{[0,y_{0.25}^{+}]}0 \succ 20_{0.25}0$, implying that 0.25 was the matching probability of event $[0, y_{0.25}]$. We also elicited the corresponding intervals for $y_{0.5}$ and $y_{0.75}$. The choice lists were similar, except that the card options had more winning suits – two winning suits for 50% and three for 75%.

Figure 4.4 Choice list to elicit matching probabilities

Line		Option_W	Option_C	
1	brown = 0 %	○	✓	Card is ◆ (25% chance)
2	0% ≤ brown≤ 4%	○	○	Card is ◆ (25% chance)
3	0% ≤ brown≤ 8%	○	○	Card is ◆ (25% chance)
4	0% ≤ brown≤ 12%	○	○	Card is ◆ (25% chance)
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
22	0% ≤ brown≤ 84%	○	○	Card is ◆ (25% chance)
23	0% ≤ brown≤ 88%	○	○	Card is ◆ (25% chance)
24	0% ≤ brown≤ 92%	○	○	Card is ◆ (25% chance)
25	0% ≤ brown≤ 96%	○	○	Card is ◆ (25% chance)
26	0 % ≤ brown < 100 %	✓	○	Card is ◆ (25% chance)

Exchangeable events

Figure 4.5 presents a choice list used to elicit exchangeable events. In each choice question, subjects had to choose between two lotteries. Payoffs of both lotteries depended on the actual color composition of the spinning wheel. Take line 4 of the list as an example, Option L(eft) pays €20 if the actual brown proportion is no more than 12%, whereas Option R(ight) pays €20 if it is more than 12%. Subjects had to choose between the two lotteries in each line of the list, depending on their subjective judgment of the actual color composition of the wheel.

The first and the last lines were pre-ticked by similar dominance arguments as for matching probabilities, and subjects were told that as they move down from the list, option L became better and option R became worse. At some point, they may switch from preferring option R to option L.

Where subject switched in Figure 4.5 provided an interval $[y^-_{median}, y^+_{median}]$ for y_{median} such that $20_{[0,y^-_{median}]}0 \prec 20_{(y^-_{median},100]}0$ and $20_{[0,y^+_{median}]}0 \succ 20_{(y^+_{median},100]}0$. Therefore, for some $y_{median} \in [y^-_{median}, y^+_{median}]$, we have $20_{[0,y_{median}]}0 \sim 20_{(y_{median},100]}0$. The events $[0, y_{median}]$ and $(y_{median}, 1]$ were both exchangeable and complementary, meaning that the subjects assigned them probability $\frac{1}{2}$. Similarly, we elicited intervals for y_{low} and y_{high} such that $20_{[0,y_{low}]}0 \sim 20_{(y_{low},0.5]}0$, and $20_{[0.5,y_{high}]}0 \sim 20_{(y_{high},1]}0$, following the method of Abdellaoui, Bleichrodt, Kemel, & L'Haridon (2014). Choice lists to elicit y_{low} and y_{high} were similar, but with different start and end points of proportion intervals (from 0% to 50% for the former, and 50% to 100% for the latter).

Figure 4.5 Choice list to elicit exchangeable events

Line		Option_L	Option_R	
1	brown = 0%	○	✓	0% < brown ≤ 100%
2	0% ≤ brown≤ 4%	○	○	4% < brown≤ 100%
3	0% ≤ brown≤ 8%	○	○	8% < brown≤ 100%
4	0% ≤ brown≤ 12%	○	○	12% < brown≤ 100%
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
22	0% ≤ brown≤ 84%	○	○	84% < brown≤ 100%
23	0% ≤ brown≤ 88%	○	○	88% < brown≤ 100%
24	0% ≤ brown≤ 92%	○	○	92% < brown≤ 100%
25	0% ≤ brown≤ 96%	○	○	96% < brown≤ 100%
26	0% ≤ brown < 100%	✓	○	brown = 100%

Incentives

Each subject received a €5 show-up fee and a variable amount of €20 depending on one of his choices in one choice period (the implementer received a flat payment of €15). A prior incentive system (Johnson, et al., 2014) was implemented to determine for each subject which choice would matter for his final payment. Before the experiment started, each subject randomly drew a sealed envelope from a pile of 156 sealed envelopes each containing one choice question (subjects faced in total 6 choice situations, each with 26 choice questions). Subjects were informed that the question that would matter for their payment was in their envelope, and were told not to open their envelopes until the end of the experiment. To determine which choice period would matter, the implementer randomly drew a number from one to four. Further implementation detailed are reported in the appendix.[16]

4.5 Raw data

Table 4.1 summarizes the number of subjects and the color of spins in sampling periods in each session. For results reported in this section, we take the midpoint of the elicited intervals as the indifference values. For instance, we take $y_{median} = \frac{(y^{-}_{median}+y^{+}_{median})}{2}$.

Take the belief of a Bayesian updater with a uniform prior as the Bayesian benchmark. Figure 4.6 plots the difference between subjects' median belief (y_{median} in the exchangeability method and y_{50} in the matching method) of the yellow proportion and the Bayesian benchmark. A positive (negative) difference corresponds to an overestimation (underestimation) of the yellow proportion. In sessions with balanced signals, subjects' median beliefs did not deviate much from the Bayesian benchmark, however, in sessions (e.g. session 1 and 7) where they received extreme signals, deviations were high. For instance, in session 1, subjects only received Brown signals. Their median deviations were positive, suggesting an overestimation of the yellow proportion on the wheel. The overestimation can be caused by conservatism: subjects did not incorporate the signals sufficiently. A similar pattern was observed in session 6 where subjects only received yellow signals and underestimated the yellow proportion on the wheel.

[16] In particular, we also controlled for possible suspicion effects by letting the subjects choose on which color they would be betting.

Table 4.1 Description of sessions

Session	# Subjects	Received signals between rounds		
		1&2	2&3	3&4
1	24	BBB	BBB	BBB
2	27	BYY	YYY	BYB
3	20	BBB	BYB	YYB
4	20	BYB	BYB	BYY
5	23	BBY	YYY	BBY
6	20	YYY	YYY	YYY
7	23	YYY	YBB	YYY

Similarly, Figure 4.7 shows how the dispersion in subjects' beliefs ($s_{high} - s_{low}$ for the exchangeability method and $s_{75} - s_{25}$ for the matching method) differs from the Bayesian benchmark. A positive (negative) difference shows that subjects are under-precise (over-precise) as compared to the Bayesian benchmark. For both median and dispersion deviations, we observed persistent individual heterogeneity. In our structural model, we estimate the confirmation and conservatism indices while taking individual differences into account.

Figure 4.6 Median deviation from the Bayesian benchmark

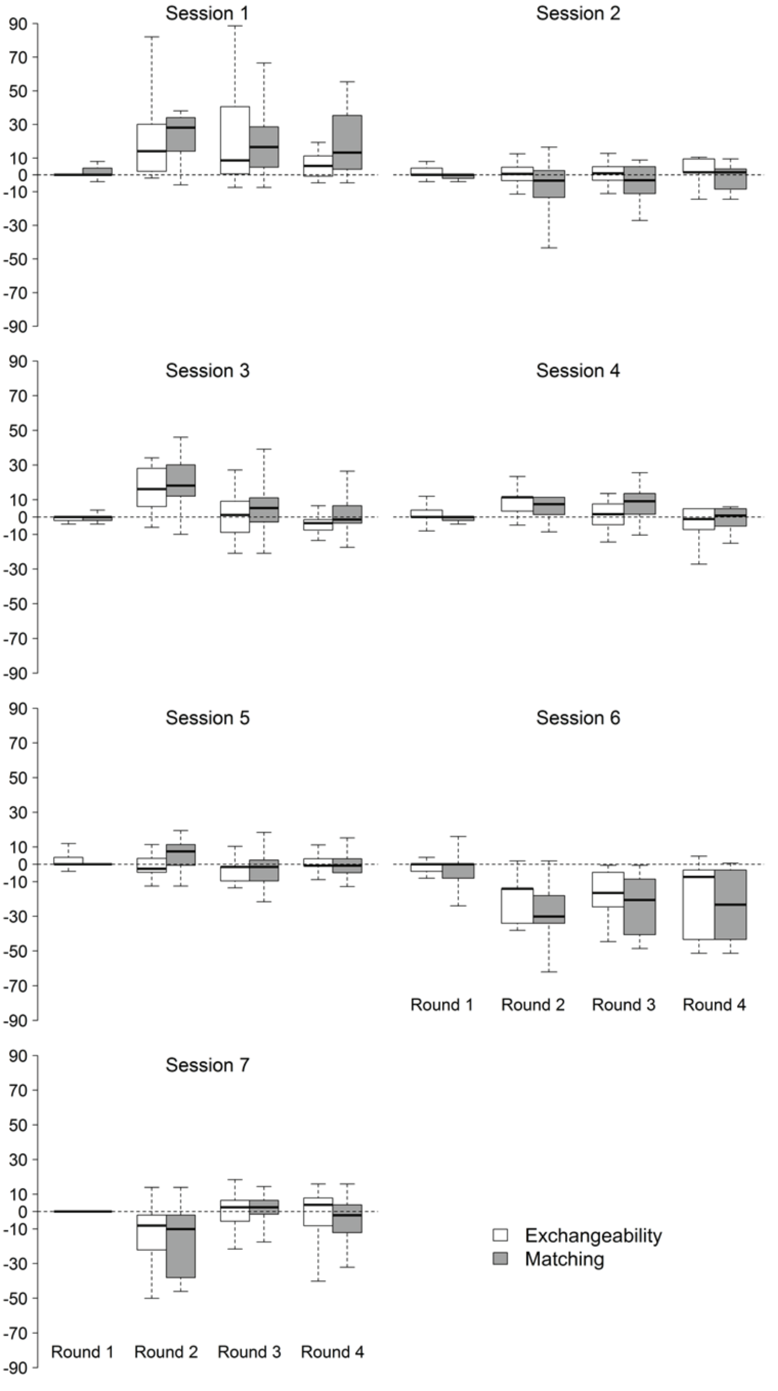
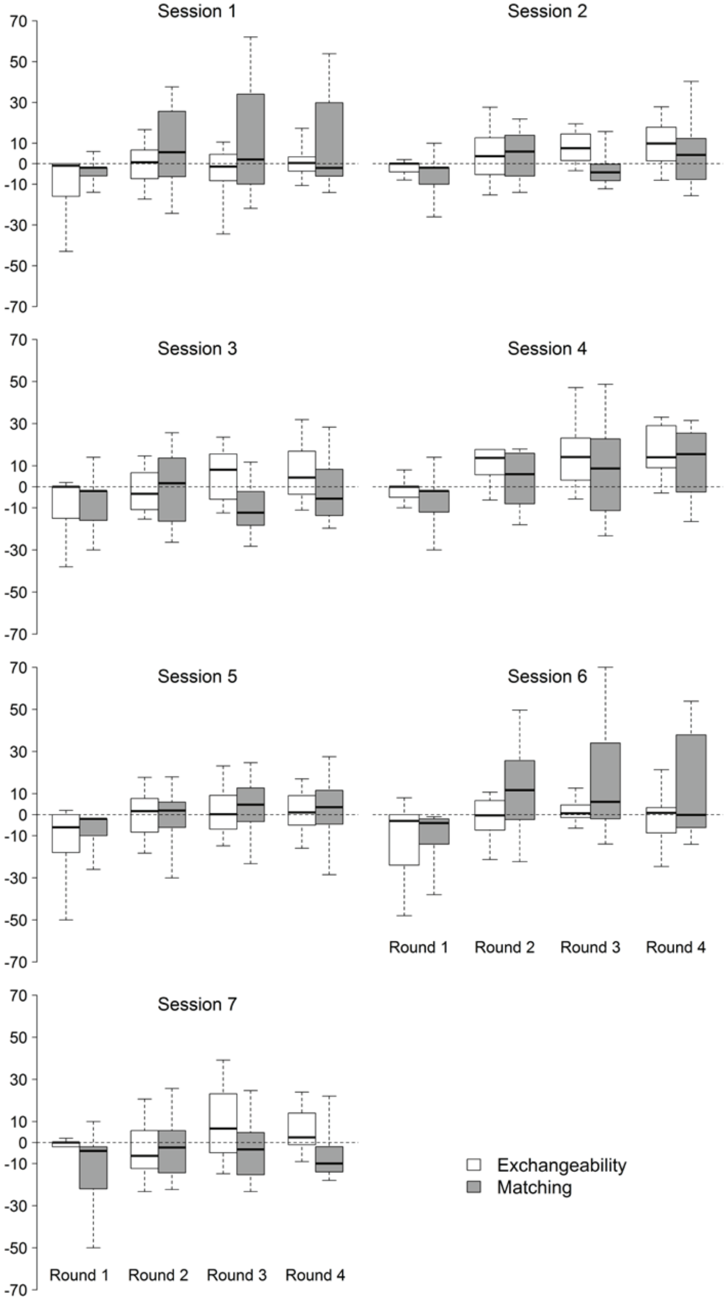


Figure 4.7 Dispersion deviation from Bayesian updating



4.6 Econometric Analysis

4.6.1 Econometric model

Measuring beliefs and deviations from Bayesian updating

The beliefs of a subject i at round j are assumed to follow a Beta distribution $Beta(.|\alpha_{i,j}, \beta_{i,j})$. The prior of subject i at round 1, determined by $\alpha_{i,1}$ and $\beta_{i,1}$, is assumed to be exogenous and will be estimated. Then, for rounds $j > 1$:

$$\alpha_{i,j} = \alpha_{i,j-1} + s(a_{i,j}, b_{i,j}, \alpha_{i,j-1}, \beta_{i,j-1}, p_{i,j}, q_{i,j})$$
$$\beta_{i,j} = \beta_{i,j-1} + f(a_{i,j}, b_{i,j}, \alpha_{i,j-1}, \beta_{i,j-1}, p_{i,j}, q_{i,j})$$

where s and f are the functions that determine respectively the perceived successes and failures, as modelled by equations 4.7 to 4.9. These functions depend on the current beliefs parameters $\alpha_{i,j-1}$ and $\beta_{i,j-1}$, the received signals $a_{i,j}$ and $b_{i,j}$ and the indices of deviations from Bayesian updating, $p_{i,j}$ and $q_{i,j}$. For a Bayesian, we have $s(a_{i,j}, b_{i,j}, \alpha_{i,j-1}, \beta_{i,j-1}, p_{i,j}, q_{i,j}) = a_{i,j}$ and $f(a_{i,j}, b_{i,j}, \alpha_{i,j-1}, \beta_{i,j-1}, p_{i,j}, q_{i,j}) = b_{i,j}$.

According to Figure 4.6, there is little or no heterogeneity in prior beliefs as measured in round 1. We therefore assume that α_1 and β_1 are constant across subjects. Much more heterogeneity in beliefs is observed for later rounds, both between and within sessions. Heterogeneity between sessions can be due to session-specific received signals that can be more or less surprising. Heterogeneity within sessions can be due to subjects' characteristics. Eventually, biases may also vary from one round to another, due to learning or fatigue. We attempt to account for these three possible sources of heterogeneity in our econometric analysis. To do so, we built a structural model that includes several explanatory variables of the deviation indices. Specifically, we assume that

$$p_{i,j} = p + \Gamma_p X_{i,j}$$
$$q_{i,j} = q + \Gamma_q X_{i,j}$$

where p and q are the intercepts of the deviation indices. When $\Gamma = 0$, these intercepts measure the aggregated indices over sessions, individuals and rounds. They should be equal to 0 if subjects perceived signals according to Bayes rule. Γ_p (respectively Γ_q) is the vector of coefficients that measure the impact of variables $X_{i,j}$ on index p (respectively q). The following explanatory variables are considered:

- a categorical variable denoting the round of the experiment,[17]
- the information gain of the received signals based on the individual specific prior,
- the squared information gain,
- subjects characteristics: including gender, and field of studies (economics or econometrics versus other).

We denote θ, the vector of coefficients to be estimated, with $\theta = (\alpha^1, \beta^1, p, q, \Gamma_p, \Gamma_q)$.

4.6.2 Estimating the model

Under our specification, the beliefs of a subject i at round j take the form of a probability distribution $\Lambda(.|\theta, a, b)$ where θ is a vector of coefficients and a and b are the received signals. This probability distribution is revealed by a series of choices, grouped within choice lists. Two types of choices lists are used. The first type, eliciting matching probabilities, considers a series of quantiles q_k and measures their corresponding values y_k^* such that $\Lambda(y_k^*) = q_k$. More precisely, these choice lists determine two values y_k^- and y_k^+ such that $20_{[0,y_k^-]}0 \prec 20_{q_k}0$ and $20_{[0,y_k^+]}0 \succ 20_{q_k}0$, i.e. $y_k^* \in [y_k^- \,;\, y_k^+]$.

The other type of choice lists, eliciting exchangeable events, considers intervals $[m_k, n_k]$ and measures the corresponding values y_k^* such that $\Lambda(m_k) - \Lambda(y_k^*) = \Lambda(y_k^*) - \Lambda(n_k)$ i.e. $\Lambda(y_k^*) = \frac{\Lambda(m_k) - \Lambda(n_k)}{2}$. Here again, the choice lists determine two values y_k^- and y_k^+ such that $20_{[m_k,y_k^-]}0 \prec 20_{[y_k^-,n_k]}0$ and $20_{[m_k,y_k^+]}0 \succ 20_{[y_k^+,n_k]}0$ i.e. $y_k^* \in [y_k^- \,;\, y_k^+]$.

For each individual i, round j and choice list k, the structural equation model provides a theoretical value $y_k^{th}(\theta, X_k)$ where θ is the vector of coefficients of our decision model, and X_k is the set of variables containing choice lists characteristics and other explanatory variables. In order to account for subject and/or specification errors, we assume that $y_k^* = y_k^{th} + \varepsilon_k$ with $\varepsilon_k \sim N(0, \sigma^2)$. Using this error specification, the likelihood of the observations provided by a given choice list is

[17] For p we consider round 2 as the reference and introduce dummy variables for round 3 and round 4. For q, preliminary analysis revealed that $\alpha_1 = \beta_1$, meaning that q is not defined for round 2. We considered round 3 as the reference and introduced a dummy variable for round 4.

$$
\begin{aligned}
p(y_k^* \in [y_k^- \,;\, y_k^+]) &= p(\varepsilon_k \in [y_k^- - y_k^{th}(\theta, X_k);\, y_k^+ - y_k^{th}(\theta, X_k)] \\
&= \Phi\left(\frac{y_k^+ - y_k^{th}(\theta, X_k)}{\sigma}\right) - \Phi\left(\frac{y_k^- - y_k^{th}(\theta, X_k)}{\sigma}\right) \\
&= l(\theta | y_k^+, y_k^-, X_k).
\end{aligned}
$$

This equation defines the likelihood of the vector of coefficients to be estimated, given the observations provided by choice lists and exogenous variables.

For a given individual i, the likelihood of a series of responses to choice lists (indexed by k), for each rounds (indexed by j), writes

$$
l_i(\theta) = \prod_j \prod_k l(\theta | y_{i,j,k}^+, y_{i,j,k}^-, X_{i,j,k})
$$

We estimate the vector of coefficients θ by maximizing the sum of log-likelihoods over individuals: $LL(\theta) = \sum_i l_i(\theta)$. This log-likelihood function is maximized by the BFGS algorithm.[18] In order to account for heterogeneity in individual error terms across rounds, specific error variances are estimated for each round. Inference is based on the (subjects) clustered standard-errors, computed from the variance-covariance matrix of individual scores.

4.6.3 Results

This section presents the estimated indices of deviation from Bayesian updating, and their explanatory variables. The results of the estimations are presented in Table 4.2.

Whatever the set of explanatory variables, the parameters α and β characterizing priors at round 1 and before receiving any signal, had very similar estimates: 1.5 and 1.4. The similarity of these two values suggests that the belief distribution of our representative subject was symmetrical. Consistently with the provided instructions, subjects did not expect one color to be more likely than the other, before receiving signals. It is nevertheless worth to note that priors were not perfectly uniform either, they exhibited a smaller variance and give slightly more probability weight to central than to extreme values of the $[0,1]$ interval.

[18] In order to avoid local maxima, for each estimations, suitable starting values were computed using grid search over 1000 possible vectors. After convergence, 10 additional estimations were run around estimated coefficients.

The first model introduces overall measures of conservatism bias (p) and confirmatory bias (q) for our representative subject. Both indices differed from 0. According to the estimated values, subjects exhibited a pronounced tendency to conservatism: they behaved as if they neglected 65% of the sample size of actual signals. Evidence for confirmatory bias was also observed: subjects behaved as if they misinterpreted 17% of signals contradicting their beliefs.

The other models (columns 2 to 4) enrich the analysis by introducing explanatory variables for the bias indices. When allowing biases to vary across rounds, we observed that conservatism bias was smaller for round 3 than for round 2, but the dummy variable for round 4 was not significant. For confirmatory bias, no significant differences were observed between rounds 3 and 4.[19]

Table 4.2 Results of Econometric Estimations

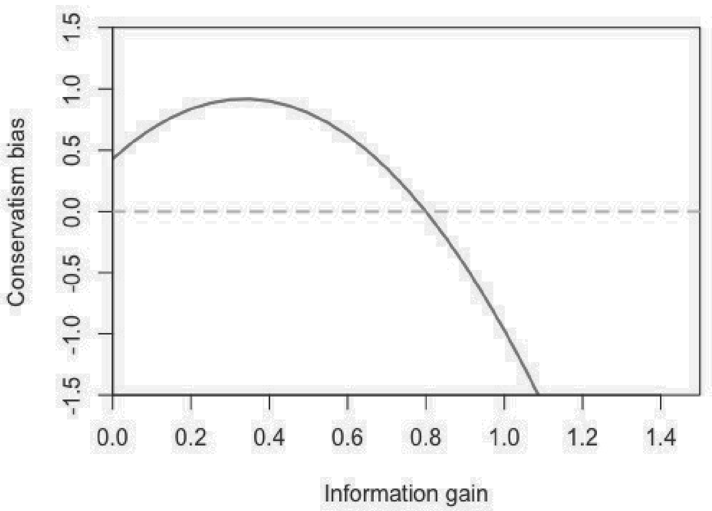
Coefficients	No Explanatory	Rounds	Rounds and Information	Round, information and subject's characteristics
σ_1	8.221 (0.086)***	8.133 (0.085)***	8.192 (0.095)***	8.184 (0.094)***
σ_2	12.143 (0.204)***	12.121 (0.204)***	12.237 (0.213)***	12.081 (0.211)***
σ_3	13.782 (0.27)***	13.696 (0.266)***	13.331 (0.263)***	13.496 (0.283)***
σ_4	13.389 (0.255)***	13.386 (0.265)***	13.229 (0.259)***	13.285 (0.28)***
α_1	1.449 (0.03)***	1.451 (0.031)***	1.485 (0.035)***	1.46 (0.034)***
β_1	1.368 (0.029)***	1.369 (0.03)***	1.378 (0.033)***	1.363 (0.034)***
$p_{Intercept}$	0.648 (0.009)***	0.673 (0.012)***	0.43 (0.053)***	0.374 (0.075)***
$q_{Intercept}$	0.166 (0.019)***	0.18 (0.028)***	0.414 (0.107)***	0.272 (0.17)ns
$p_{round=3}$		-0.09 (0.024)***	-0.256 (0.03)***	-0.119 (0.024)***
$p_{round=4}$		-0.01 (0.043)ns	-0.103 (0.069)ns	0.05 (0.063)ns
$q_{round=4}$		0.057 (0.075)ns	-0.339 (0.212)ns	-0.502 (0.108)***
p_{ig}			2.895 (0.263)***	1.435 (0.362)***
p_{ig^2}			-4.294 (0.358)***	-1.846 (0.433)***
q_{ig}			0.177 (0.568)ns	0.735 (3.198)ns
q_{ig^2}			-0.867 (0.711)ns	-2.984 (9.529)ns
$p_{gender=female}$				0.069 (0.014)***
$p_{major=Econ}$				0.085 (0.021)***
$q_{gender=female}$				0.047 (0.06)ns
$q_{major=Econ}$				0.156 (0.061)*

Notes: Clustered standard errors are reported between brackets. Stars report significance levels: ns for $p \geq 0.05$, * *for* $p < 0.05$, ** *for* $p < 0.01$, *** *for* $p < 0.001$

[19] The q index was not estimated for round 2 because people hold (approximately) symmetric beliefs in round 1.

The model of column 3 accounts for heterogeneity across rounds and across sessions by including information gain as an explanatory variable. Exploratory analysis suggested that the impact of information gain might be non-linear, and therefore a polynomial effect was considered. Information gain was not found to impact confirmatory bias, but impacted conservatism bias significantly. The coefficients associated to the two degrees of the polynomial were significant and suggested that the relationship was not monotonic, but inverse-U shaped. The shape of the estimated effect is represented in Figure 4.8. Moderately informative signals increased the biases, whereas very poorly or highly informative ones reduced them. It is noteworthy that very surprising signals were able to reverse the sign of the conservatism index, possibly leading to prior signals destruction.

Figure 4.8 Polynomial effect of information gain on p



4.6.4 Stability of results across measurement methods

Two different methods were used to measure beliefs. Estimations presented in the previous section pool observations from the two methods, assuming that they give similar patterns. We also tested this assumption.

The first and last models from Table 4.2 were re-estimated with all explanatory variables interacting with a dummy variable coding for the method used for estimations. For the simple model containing only intercept values of p and q, the dummy variable was found to have a significant impact on p. When measured with the exchangeability method (that is robust to ambiguity attitudes), the index of conservatism bias was lower by 0.133 (clustered standard error: 0.026). Regarding the confirmatory bias, the difference between measurement methods was estimated as 0.07 (clustered standard error: 0.10) but was not statistically significant.

The aforementioned result suggest that ambiguity can amplify conservatism. Their posterior beliefs, measured with matching probabilities, remain too diffuse. Controlling for ambiguity by the use of exchangeable events leads to less deviations from Bayesian updating.

The dummy variable denoting the measurement method was also included in the model with explanatory variables for biases (column 4), and possible interactions were allowed. The estimated model contains 32 coefficients. A likelihood test was run to check whether adding interactions between dummy variable for method and other explanatory variables increased the likelihood significantly. The p-value of the test is 0.07. This suggests that allowing for coefficients to interact with the method dummy does not increase the goodness of fit significantly. Therefore, the coefficients of the explanatory variables do not vary significantly with the measurement method.

4.7 Discussion

This paper models belief updating when a combination of conservatism and confirmatory bias may distort people's perception of signals received, thus incorporating new information insufficiently or asymmetrically. Our model provides an intuitive interpretation of the biases and makes them observable from revealed preferences. It extends Rabin and Schrag's (1999) model by accounting for more patterns of deviations from Bayesian updating.

The experiment illustrated how the indices could be estimated in a tractable manner. Thus, it provided the first structural estimation of the two well-known biases. The results showed evidence for both confirmatory bias and conservatism at aggregate level. On average, the confirmatory bias index was estimated as 0.17 suggesting that an opposite signal may be misread with 17% chance. The conservatism index was 0.65 suggesting a strong stickiness to priors. Furthermore, the conservatism bias also depended on the informativeness of signals as measured by how surprising the signals were given the prior beliefs. In particular, subjects were more conservative when the new signals were neither extremely surprising nor extremely non-surprising. This pattern is consistent with the previous findings indicating the tradeoff between the strength and the credibility of evidence (Griffin & Tversky, 1992; Massey & Wu, 2005; Kvam & Pleskac, 2016).

Our experiment contributes to the empirical investigation of confirmatory bias. Despite the abundance of theoretical models on confirmatory bias in economics literature, the main empirical findings for confirmatory bias mainly come from the psychology literature (for reviews, see Klayman, 1995; Nickerson, 1998; Oswald & Grosjean, 2004). However, the subjective nature of the psychological experiments do not allow a formal investigation of confirmatory bias due to the lack a normative benchmark for comparison of revised beliefs. Although there are a few field studies documenting evidence on confirmatory bias (Andrews, Logan, & Sinkey, 2015; Sinkey, 2015; Christandl, Fetchenhauer, & Hoelzl, 2011), there is still lack of evidence in standard Bayesian updating experiments. Several recent studies document evidence on asymmetric processing of information in Bayesian updating as in confirmatory bias, when the information has a valence or it is self-relevant (Coutts, 2016; Eil & Rao, 2011; Ertac, 2011). Different from our ego-neutral setting, these studies employ ego-related settings where subjects make inferences about their scores on intelligence tests or their physical attractiveness rated by other subjects in the same experimental session. Eil and Rao (2011) argue that confirmation of prior beliefs happens only when the confirming evidence supports a positive ego image. Specifically, people are more responsive to positive feedback compared to negative feedback about themselves regardless of their prior beliefs. Our results show that confirmatory bias can also arise in an ego-neutral setting. The direction of the bias, however, then depends on the informativeness of signals.

4.8 Conclusion

This paper studies biases in people's belief updating from a descriptive perspective. We modelled deviations from Bayesian updating by allowing perceived signals to differ from the signals people actually receive. It provides a natural interpretation of well-known biases and makes them observable from choices. Our model thus adheres to the revealed-preference approach of economics.

In our experiment, confirmatory bias and conservatism were dominant at the aggregate level, while individual heterogeneity persisted. The opposite of conservatism arose in situations where the signals were extremely surprising. This finding illustrates the relevance of allowing for different deviation patterns. Overall, our results replicated previous findings on Bayesian updating, suggesting that our model and the method are empirically valid.

Appendix 4

Detailed Experimental Procedure

Every subject received a subject ID upon arrival. In each session the subject whose ID started with M was invited to the front and introduced to all subjects as the implementer of that session. The implementer was then guided to a desk at the rear end of the room isolated by a wooden panel. The implementer would implement the randomization tasks to make sure that they were conducted in a fair and transparent manner.

Each session started with oral instructions by one of the experimenters – the instructor – using slides. Throughout the experiment, subjects could ask questions when anything was unclear. A training wheel was used during the instructions for illustration purpose. The training wheel was covered by blue and red, instead of brown and yellow to avoid potential misunderstandings and biases. The implementer first confirmed that the training wheel hidden behind the panel was covered by brown and brown, and there were no other colors on the wheel. He then spinned the wheel three times and reported the resulting colors. These colors were written down on the white board so that all subjects could see during the instruction. Subjects then received a training questionnaire with all choice situations that they would face during the experiment. The instructor went through them with the subjects, and the subjects filled in the training questionnaires based on the sample information from the practice wheel as a practice.

After all subjects were familiarized with the experimental tasks, the instructor explained to the subjects how their final payment would be determined with an example envelope content. The oral instructions ended with the explanation of the structure of the experiment.

After the instructions and before the start of the actual experiment, each subject drew a sealed envelope and the implementer randomly drew a period number from 1 to 4. Then, the implementer randomly drew a card from the deck of four cards. The selected period number and the card were sealed in two envelopes and only revealed at the end of the experiment. The implementer then drew a color composition for the wheel. He confirmed to all subjects that the wheel was covered by two and only two colors: yellow and brown.

Before handing out the questionnaires for the first choice period, each subject could state his preference between betting on yellow proportion and betting on brown proportion during the experiment. He received questionnaires with that color throughout the experiment. The subjects were requested to write their subject IDs on every questionnaire that they filled in so that their choices could be tracked down over the periods. The questionnaires were collected at the end of every choice period, and the sampling period proceeded. The outcome of every spin were announced by the implementer, and written down on the white board by the experimenter. New questionnaires were handed out after each sampling period.

At the end of the experiment, the color composition of the wheel, the card suit, and the choice period drawn for the payment stage was revealed to the subjects by the implementer. The subjects were requested to open their envelopes, and to proceed to the payment desk, where they got paid according to the outcome of their preferred lottery in the choice question that came out of their envelopes.

Chapter 5

An Experimental Test of Reduction Invariance[20]

with Han Bleichrodt and Yu Gao

5.1 Introduction

Probability weighting is an important reason why people deviate from expected utility (Fox & Poldrack, 2014; Luce, 2000; Wakker, 2010). Prelec (1998) proposed a functional form for the probability weighting function that is widely used in empirical research and that usually gives a good fit to empirical data (Chechile & Barch, 2013; Sneddon & Luce, 2001; Stott, 2006).

Although other functional forms have also been used (e.g. Currim & Sarin, 1989; Goldstein & Einhorn, 1987; Karmarkar, 1978; Lattimore, Baker, & Witte, 1992; Tversky & Kahneman, 1992), Prelec was the first to give an axiomatic foundation for a form of the probability weighting function.[21] His central condition, compound invariance (defined in Section 2), is, however, complex to test empirically as it involves four indifferences and may be subject to error cumulation. To the best of our knowledge, it has not been tested yet.

Luce (2001) proposed a simpler condition, reduction invariance. Luce (2000, p. 278) identified testing reduction invariance as an important open empirical problem. The purpose of this paper is to follow up on Luce's suggestion and to test reduction invariance in an experiment. Our data support the validity of reduction invariance. At the aggregate level, we found evidence for the condition and at the individual level it was clearly the dominant pattern.

A special case of reduction invariance is the rational case of reduction of compound gambles, which implies that the probability weighting function is a power function. Our data on reduction of compound gambles are mixed. At the aggregate level reduction of compound gambles was clearly violated. However, 60% of our subjects behaved in line with it. The subjects who deviated, did so systematically and found compound gambles more attractive than simple gambles.

[20] This chapter appeared in Journal of Mathematical Psychology (2016) Vol. 75: 170-182

[21] For a more recent axiomatic analysis of probability weighting see Diecidue et al. (2009).

5.2 Background

Let (x, p) denote a *gamble* which gives consequence x with probability p and nothing otherwise. Consequences can be pure, such as money amounts, or they can be a gamble (y, q) where y is a pure consequence. The set of pure consequences is a nonpoint interval $\mathcal{X}$ in $\mathbb{R}^+$ that contains 0. Preferences $\succcurlyeq$ are defined over the set $\mathcal{C}$ of gambles. We identify preferences over simple gambles (x, p) from preferences over $\big((x, p), 1\big)$ and preferences over consequences x from preferences over $(x, 1)$.

A function U *represents* $\succcurlyeq$ if it maps gambles and pure consequences to the reals and for all gambles $(x, p), (x', p')$ in $\mathcal{C}$, $(x, p) \succcurlyeq (x', p') \Leftrightarrow U(x, p) \geq U(x', p')$. If a representing function U exists, then $\succcurlyeq$ must be a *weak order*: transitive and complete. The representing function U is *multiplicative* if there exists a function $W: [0,1] \rightarrow [0,1]$ such that:

i. $U(x, p) = U(x)W(p)$.
ii. $U(0) = 0$ and U is continuous and strictly increasing.
iii. $W(0) = 0$ and W is continuous and strictly increasing.

The functions U and W are unique up to different positive factors and a joint positive power: $U \rightarrow a_1 U^b$ and $W \rightarrow a_2 W^b, a_1, a_2, b > 0$. This uniqueness implies that we can always normalize W such that $W(1) = 1$.[22] Luce (1996, 2000) and Marley & Luce (2002) gave preference foundations for the multiplicative representation. A central condition in these results is consequence monotonicity, which we also assume here.[23]

The multiplicative representation is general and contains many models of decision under risk as special cases. Examples are expected utility, rank- and sign-dependent utility (Quiggin, 1981; 1982), prospect theory (Tversky & Kahneman, 1992), disappointment aversion theory (Gul, 1991), and rank-dependent utility (Luce, 1991; Luce & Fishburn, 1991; 1995).

[22] Aczél & Luce (2007) analyzed the case where $W(1) \neq 1$ to model non-veridical responses in psychophysical theories of intensity (Luce, 2002; 2004).

[23] Consequence monotonicity means that if two gambles differ only in one consequence, the one having the better consequence is preferred. As Luce (2000, p. 45) points out, it implies a form of separability for compound gambles. It also implies backward induction, where each simple gamble in a compound gamble can be replaced by its certainty equivalent. Von Winterfeldt, Chung, Luce, & Cho (1997) found few violations of consequence monotonicity for choice-based elicitation procedures, as used in our experiment, and what there was seemed attributable to the variability in certainty equivalence estimates.

Prelec (1998) axiomatized the following family of weighting functions:

DEFINITION 5.1: $W(p)$ is *compound-invariant* if there exist $\alpha > 0$ and $\beta > 0$ such that $W(p) = \exp(-\beta(-lnp)^{\alpha})$.

Prelec's compound-invariant weighting function has several desirable properties. First, it includes the power functions $W(p) = p^{\beta}$ as a special case. The class of power weighting functions is the only one that satisfies *reduction of compound gambles*, which is often considered a feature of rational choice:

$$\big((x,p),q\big) \sim (x,pq).$$

A second advantage of the compound-invariant family is that for $\alpha < 1$, it can account for inverse S-shaped probability weighting, which has commonly been observed in empirical studies (Fox & Poldrack, 2014; Wakker, 2010). Finally, the parameters α and β have an intuitive interpretation (Gonzalez & Wu, 1999). The parameter α reflects a decision maker's sensitivity to changes in probability, with higher values representing more sensitivity, while β reflects the degree to which a decision maker is averse to risk, with higher values reflecting more aversion to risk.

The compound-invariant family of weighting functions satisfies the following condition:

DEFINITION 5.2: Let N be any natural number. *N-compound invariance* holds if $(x,p){\sim}(y,q)$, $(x,r){\sim}(y,s)$, and $(x',p^N){\sim}(y',q^N)$ imply $(x',r^N){\sim}(y',s^N)$ for all nonzero consequences x, y, x', y' and nonzero probabilities p, q, and r.

Compound invariance holds if N-compound invariance holds for all N. Prelec (1998) showed that if compound invariance is imposed on top of the multiplicative representation then $W(p)$ is compound-invariant. Bleichrodt, Kothiyal, Prelec, & Wakker (2013) showed that compound invariance by itself implies the multiplicative representation and, consequently, that the assumption of a multiplicative representation is redundant.

Compound invariance is difficult to test empirically. It requires four indifferences and elicited values appear in later elicitations, which may lead to error cumulation. For example, we could fix x, p, q, r, and x'. The first indifference would then elicit y, the second s, and the third y'. If each of these variables is measured with some error then this will affect the final preference between (x', r^N) and (y', s^N).

To address the problem of error cumulation, Luce (2001) proposed a simpler condition.

DEFINITION 5.3: Let N be any natural number. *N-reduction invariance* holds if $((x,p),q)\sim(x,r)$ implies $((x,p^N),q^N)\sim(x,r^N)$ for all nonzero consequences x and nonzero probabilities p, q, and r.

Reduction invariance holds if N-reduction invariance holds for all N. Reduction invariance is easier to test than compound invariance as it requires only two indifferences. Luce (2001, Proposition 1) showed that if N-reduction invariance for $N = 2,3$ is imposed on top of the multiplicative representation then the weighting function $W(p)$ is compound-invariant. To the best of our knowledge, Bleichrodt et al.'s (2013) result cannot be generalized to reduction invariance and the multiplicative representation still has to be assumed in this case.

5.3 Experimental design

The purpose of our experiment was to test reduction invariance (for $N = 2,3$) to obtain insight into the descriptive validity of the compound-invariant weighting function. The simplest way to test reduction invariance would be to fix x, p, and q, to elicit the probability r such that a subject is indifferent between $((x,p),q)$ and (x,r), and then to check whether he is indifferent between $((x,p^N),q^N)$ and (x,r^N). However, as Luce (2001) pointed out, a danger of this procedure is that many subjects may realize that $r = pq$ is a sensible response. This may distort the results as empirical evidence suggests that subjects do not satisfy reduction of compound gambles (Abdellaoui et al., 2015; Bar-Hillel, 1973; Bernasconi & Loomes, 1992; Keller, 1985; Slovic, 1969). Luce (2001) suggested another approach for testing reduction invariance, which we adopted in our experiment. Instead of asking for probability equivalents, we elicited the certainty equivalents of $((x,p),q)$, denoted $CE((x,p),q)$, and several $CE(x,r)$ for a range of values of r

centered on pq. Using interpolation, we then determined the value r_1 for which $CE\big((x,p),q\big) = CE(x,r_1)$. We then elicited $CE\big((x,p^2),q^2\big)$ and $CE\big((x,p^3),q^3\big)$ and tested whether $CE\big((x,p^2),q^2\big) = CE(x,(r_1)^2)$ and $CE\big((x,p^3),q^3\big) = CE(x,(r_1)^3)$ where $CE(x,(r_1)^2)$ and $CE(x,(r_1)^3)$ were, again, determined using interpolation.

Procedure

The experiment was run on computers. Subjects were seated in cubicles with a computer screen and a mouse and could not communicate with each other. Once everyone was seated, the instructions were displayed, followed by three comprehension questions. Subjects could only proceed to the actual experiment when they had correctly answered all three comprehension questions. Copies of the instructions and the comprehension questions are in Appendix 5.1.

Table 5.1 Compound gambles used in the experiment

Compound gambles	**Gamble**	**Type**	**Reduced probability**	**Expected value**
C1	$\big((€200, 82\%), 67\%\big)$	Original	54.94%	€109.88
C2	$\big((€200, 45\%), 67\%\big)$	Original	30.15%	€60.30
C3	$\big((€200, 63\%), 90\%\big)$	Original	56.70%	€113.40
C4	$\big((€200, 82\%), 39\%\big)$	Original	31.98%	€63.96
C5	$\big((€200, 67\%), 45\%\big)$	Square of C1	30.15%	€60.30
C6	$\big((€200, 20\%), 45\%\big)$	Square of C2	9.00%	€18.00
C7	$\big((€200, 40\%), 81\%\big)$	Square of C3	32.40%	€64.80
C8	$\big((€200, 67\%), 15\%\big)$	Square of C4	10.05%	€20.10
C9	$\big((€200, 55\%), 30\%\big)$	Cube of C1	16.50%	€33.00
C10	$\big((€200, 9\%), 30\%\big)$	Cube of C2	2.70%	€5.40
C11	$\big((€200, 25\%), 73\%\big)$	Cube of C3	18.25%	€36.50
C12	$\big((€200, 55\%), 6\%\big)$	Cube of C4	3.30%	€6.60

We measured the certainty equivalents of 12 compound gambles and of 6 simple gambles. The order in which these gambles were presented was random. The winning amount was always €200. Table 5.1 displays the compound gambles that we used. Compound gambles C1-C4 were the original gambles, gambles C5-C8 were derived from C1-C4 by taking the squares of the probabilities, and gambles C9-C12 were derived from C1-C4 by taking the cubes of the probabilities. Because taking the square and the cube of probabilities usually does not give round numbers, we selected the probabilities in the compound gambles C1-C4 such that only little rounding was necessary in the derived compound gambles. We could have avoided rounding altogether by presenting fractions. However, we observed in the pilot sessions that subjects found complex fractions harder to handle than probabilities.

By comparing the certainty equivalents of C2 and C5 and (roughly) those of C4 and C7 we could test whether subjects preferred to have most of the uncertainty resolved in the first stage or in the second stage. Luce (1990, p. 228) already drew attention to modeling the order in which events are carried out and Ronen (1973) and Budescu & Fischer (2001) found that people prefer gambles with high first-stage probabilities and lower second-stage probabilities to gambles with high second-stage probabilities and lower first-stage probabilities. On the other hand, Chung, Von Winterfeldt, & Luce (1994) concluded that with a choice-based procedure most subjects were indifferent to the order in which events were carried out.

Table 5.2 Simple gambles used in the experiment

Simple gambles	**Gamble**	**Expected value**
S1	(€200, 3%)	€6
S2	(€200, 9%)	€18
S3	(€200, 17%)	€34
S4	(€200, 32%)	€64
S5	(€200, 57%)	€114
S6	(€200, 77%)	€154

Table 5.2 shows the simple gambles that we used in the experiment. The probabilities in the simple gambles were close to the reduced probabilities of the compound gambles.

To determine the certainty equivalents of the compound and the simple gambles, subjects made a series of choices between these gambles and sure amounts of money. Simple risk and compound risk were represented by urns containing colored balls. The color of the ball determined subjects' payoffs. We used one urn for the simple gambles and two urns for the compound gambles. Appendix 5.1 displays the way the simple and the compound gambles were presented.

All certainty equivalents were elicited using a choice-based iterative procedure, which is close to the PEST procedure used by, amongst others, Cho & Luce (1995) and Cho, Luce, & Von Winterfeldt (1994). We did not ask subjects directly for their certainty equivalents as this tends to lead to less reliable measurements (Bostic, Herrnstein, & Luce, 1990), but instead used a series of choices to zoom in on them. The iteration procedure is described in Appendix 5.2.

We included two types of consistency tests. First, we repeated the third choice in the iteration procedure for four randomly selected questions. Subjects were usually close to indifference in the third choice and, consequently, this was a rather strong test of consistency. Second, we repeated the entire elicitation of two certainty equivalents, one for a randomly selected simple gamble and one for a randomly selected compound gamble.

Subjects and incentives

The experiment was performed at the ESE-Econlab at Erasmus University in 5 group sessions. Subjects were 79 Erasmus University students from various academic disciplines (average age 23.4 years, 43 female). We paid the subjects a €5 participation fee. In addition, at the conclusion of each session we randomly selected two subjects who could play out one of their randomly drawn choices for real. If a subject had chosen the sure amount in that choice then we paid him that amount. If he had chosen the simple or the compound gamble then we created the relevant urn(s) and the subject drew the ball that determined his payoffs. The 10 subjects who played out one of their choices for real earned on average €49.60 per person. Sessions lasted 45 minutes on average including 10 minutes to implement payment.

Analysis

To test reduction invariance, we followed Luce's (2001) suggestion. We determined for each compound gamble $\big((€200, p), q\big)$ the probability r such that $CE\big((€200, p), q\big) = CE(€200, r)$ using the certainty equivalents of the simple gambles and linear interpolation. Subjects' certainty equivalents of the simple gambles did not always increase with the probability of winning €200 and, consequently, the value of r for which $CE\big((€200, p), q\big) = CE(€200, r)$ could not always be uniquely determined. If there were multiple values of r for which $CE\big((€200, p), q\big) = CE(€200, r)$ then we used the average of these values in our analysis. We also analyzed the results using only those responses for which r could be uniquely determined, but this did not affect our conclusions. Finally, we also estimated the weighting function by smoothing splines (Hastie, Tibshirani, &

Friedman, 2008, Section 5.4) and used this estimation to predict r.[24] We discuss the results of this nonparametric regression analysis in the subsection Robustness analysis.

People's preferences are typically stochastic and the elicited certainty equivalents are subject to noise. Moreover, the choice-based procedure determined certainty equivalents up to €1 precision and it was in theory possible that the absolute difference between $CE\big((€200, p^N), q^N\big)$ and $CE(€200, r^N)$, $N = 2{,}3$, was equal to 2 even though a subject satisfied reduction invariance exactly. For these reasons and because $CE(€200, r^N), N = 2{,}3$, had to be approximated, which introduced further imprecision, we considered a test of equality of the certainty equivalents too stringent. Instead, we followed Cho & Luce's (1995) approach in testing preference conditions and compared the proportions of respondents for whom $CE\big((€200, p^N), q^N\big) > CE(€200, r^N)$ with those for whom $CE\big((€200, p^N), q^N\big) < CE(€200, r^N), N = 2{,}3$. Under reduction invariance with random error, deviations from equality between $CE\big((€200, p^N), q^N\big)$ and $CE(€200, r^N)$ should be nonsystematic and we should observe that the proportion of subjects for whom $CE\big((€200, p^N), q^N\big) > CE(€200, r^N)$ does not differ systematically from the proportion for whom $CE\big((€200, p^N), q^N\big) < CE(€200, r^N)$. Because our elicitation method only determined certainty equivalents up to €1 precision we took $CE\big((€200, p^N), q^N\big)$ and $CE(€200, r^N)$ equal if $\big|CE\big((€200, p^N), q^N\big) - CE(€200, r^N)\big| \leq 2$.[25] We also analyzed the results using the exact equality. This did not affect our conclusions at the aggregate level but, obviously decreased support for reduction invariance at the individual level.[26]

Our null hypothesis is that reduction invariance holds, which involves testing the invariance $P\Big(CE\big((€200, p^N), q^N\big) > CE(€200, r^N)\Big) = P\Big(CE\big((€200, p^N), q^N\big) < CE(€200, r^N)\Big)$. As pointed out by Rouder, Speckman, Sun, Morey, & Iverson (2009) and Rouder, Morey, Speckman, & Province (2012) classic null-hypothesis significance tests are less suitable when testing for invariances for two reasons. First, they do not allow

[24] For these estimations we used the smooth splines function in R (R Core Team, 2015) which estimates prediction error by generalized cross-validation.
[25] Hence, we also defined $CE\big((€200, p^N), q^N\big) > CE(€200, r^N)$ if $CE\big((€200, p^N), q^N\big) - CE(€200, r^N) > 2$ and $CE\big((€200, p^N), q^N\big) < CE(€200, r^N)$if $CE\big((€200, p^N), q^N\big) - CE(€200, r^N) < -2$.
[26] In the consistency tests and the tests of reduction of compound gambles that we report in Section 4 we used Bayesian t-tests. In these tests we did not have to use interpolation and a substantial proportion of the subjects stated the same certainty equivalents. Using tests of proportions here would make the analysis less informative and underestimate the support for the null hypothesis.

researchers to state evidence for the null hypothesis and, second, they overstate the evidence against the null hypothesis. We therefore used Bayes factors to test our null hypotheses. The Bayes factors describe the relative probability of the observed data under the null and the alternative hypothesis. For example, a Bayes factor of 10 will indicate that the null is 10 times more likely than the alternative given the data. We used the package BayesFactor in R (Morey et al., 2015) to compute the Bayes factors. Following Jeffreys (1961) we interpret a Bayes factor larger than 3 as "some evidence" for the null, a Bayes factor larger than 10 as "strong evidence" for the null, and a Bayes factor larger than 30 as "very strong evidence" for the null. Similarly, a Bayes factor less than 0.33 [0.10, 0.03] is counted as some [strong, very strong] evidence for the alternative hypothesis.

In the individual subject analyses, we classified individual subjects based on the number of times they displayed the patterns $\big((€200, p^N), q^N\big) - CE(€200, r^N) < -2$, $-2 < CE\big((€200, p^N), q^N\big) - CE(€200, r^N) < 2$, and $CE\big((€200, p^N), q^N\big) - CE(€200, r^N) > 2$ for both $N = 2$ and $N = 3$. For 2-reduction invariance, we defined subjects who reported $CE\big((€200, p^2), q^2\big) - CE(€200, r^2) < -2$ more than twice as *Type compound<simple*. We only required them to display this pattern in a majority of tests to account for response error. Similarly, we defined subjects who reported $CE\big((€200, p^2), q^2\big) - CE(€200, r^2) > 2$ more than twice as *Type compound>simple*. The other subjects were assumed to behave in line with reduction invariance (plus some error) and were defined as *Type RI*. The classification for $N = 3$ was identical.

In the individual analyses of reduction of compound gambles we defined subjects who reported $CE\big((€200, p), q\big) - CE(€200, pq) < -2$ in a majority of tests (more than 6 times) as *Type compound<simple*. Subjects who reported $CE\big((€200, p), q\big) - CE(€200, pq) > 2$ more than 6 times were defined as *Type compound>simple* and the other subjects were assumed to behave in line with reduction of compound gambles plus error and were defined as *Type RCG*.

5.4 Results

We removed one subject from the analyses because her responses reflected confusion.[27] The results presented next used the responses of the remaining 78 subjects.

Consistency

Each subject repeated four choices and two complete elicitations. For each subject, the repeated choices were randomly selected (and hence differed across subjects) but they were always a choice that the subject had faced in the third step of the iteration procedure. Subjects made the same choice in 72.8% of the repeated choices. Reversal rates up to one third are common in the literature (Wakker, Erev, & Weber, 1994; Stott, 2006) and we, therefore, consider our reversal rates as satisfactory, especially if we take into account that subjects were usually close to indifference in the third iteration. Fifty-four subjects (69%) had one reversal at most. Six subjects (8%) had more than two reversals. We also analyzed the data without these subjects, but this led to similar results. The proportions of reversals were about the same in the simple gambles and in the compound gambles: 24% versus 29% and the Bayesian 95% credible intervals overlapped.

We also repeated two complete elicitations, one for a simple gamble and one for a compound gamble. Both gambles were randomly selected and, consequently, they differed across subjects. The data favored the null hypothesis of equality between the original and the repeated measurement (the Bayes factors (BFs) were 6.48 for simple gambles and 7.07 for compound gambles). The mean absolute deviation between the original and the repeated measurement was €15.38. The median was lower (€8) indicating that there were a few outliers with large differences, but for most subjects the differences were modest. The data favored the null hypothesis that the mean difference between the original and the repeated measurement was the same for the simple and for the compound gambles (BF = 7.96).

Because the questions that were repeated had different expected values, we also looked at the absolute difference as a percentage of the expected value. The mean of these percentages was 60%, the median was again much lower: 18%. The data supported the null that the means of these percentages were equal for the simple and the compound gambles

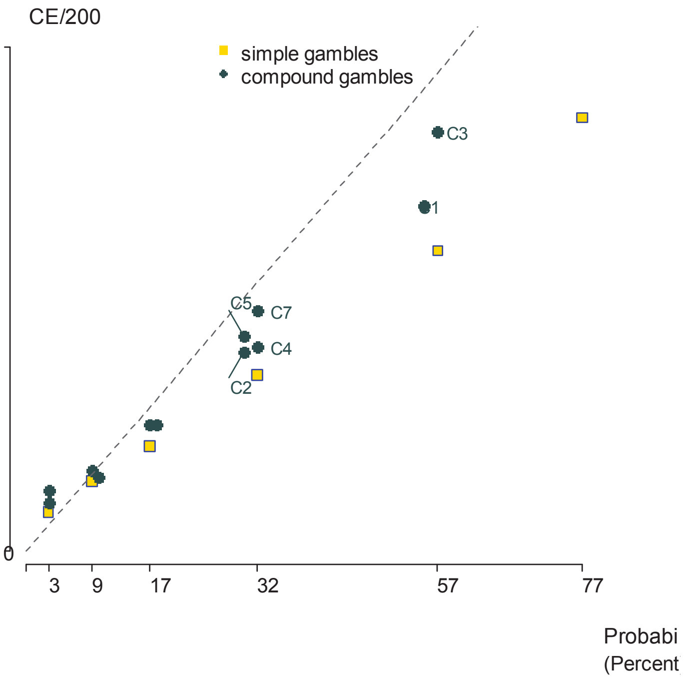
[27] In several choices, she chose 0 for sure over a gamble, which gave a positive probability of €200 and could not result in a payoff less than €0.

(BF = 6.96) and we had no indication that subjects made more errors or had less precise preferences in the, arguably, more complex compound gambles.

Certainty equivalents

Figure 5.1 displays the certainty equivalents of the simple and the compound gambles. We divided these certainty equivalents by 200 to give a visual impression of subjects' risk attitudes. For risk neutral subjects, the certainty equivalents of the simple gambles (the squares in the figure) will lie on the diagonal; points above the diagonal reflect risk seeking and points below the diagonal reflect risk aversion. The figure shows the usual pattern of risk seeking for small probabilities and risk aversion for moderate and large probabilities, which is equivalent to inverse S-shaped probability weighting if utility is linear.

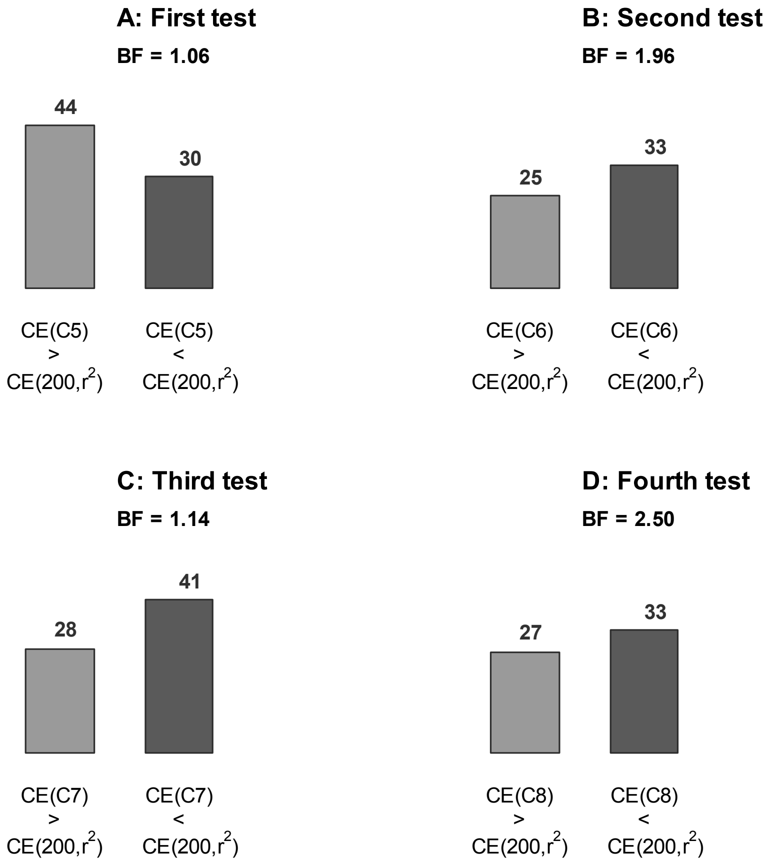
Figure 5.1: Mean certainty equivalents (divided by 200) of the simple and the compound gambles



Tests of reduction invariance

Figures 5.2.1 and 5.2.2 show the results of the eight tests of reduction invariance that we performed. Figures 5.2.1 shows the results of the four tests of 2-reduction invariance and Figure 5.2.2 those of the four tests of 3-reduction invariance. For each test we have indicated the BF-values.

Figure 5.2.1 Tests of 2-reduction invariance



Notes: The figure shows the number of subjects for whom the certainty equivalent of the compound gamble is greater than respectively smaller than the certainty equivalent of the simple gamble (taking into account the imprecision in our measurements). BF stands for Bayes factor with higher values indicating more support for the null hypothesis that reduction invariance holds.

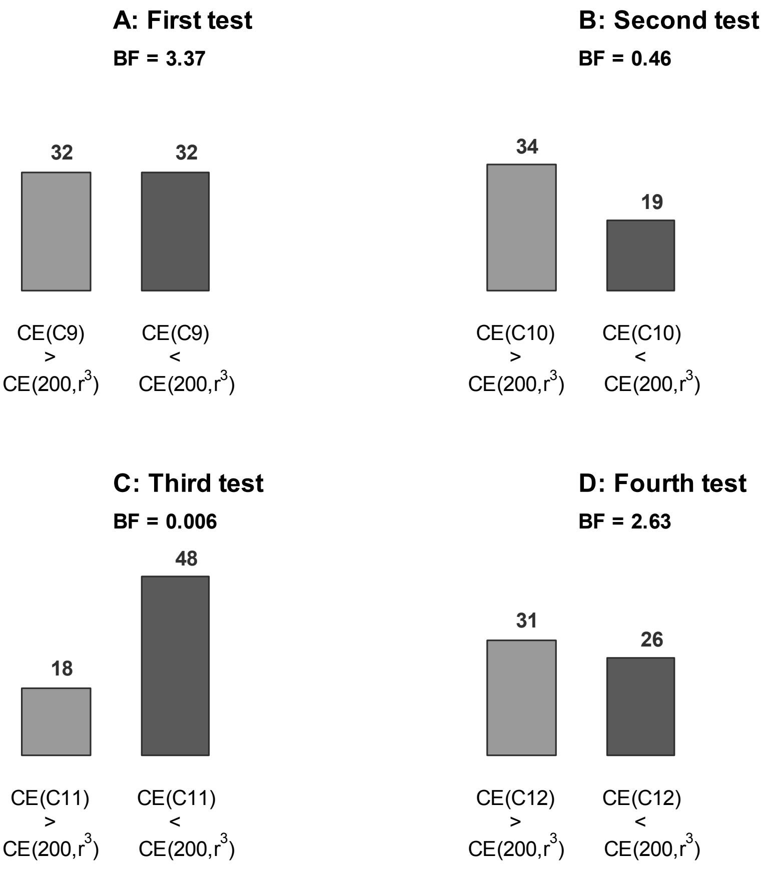


Figure 5.2.2 Tests of 3-reduction invariance

Notes: The figure shows the number of subjects for whom the certainty equivalent of the compound gamble is greater than respectively smaller than the certainty equivalent of the simple gamble (taking into account the imprecision in our measurements). BF stands for Bayes factor with higher values indicating more support for the null hypothesis that reduction invariance holds.

Pooled over all tests, the data supported the null hypothesis that reduction invariance held ($BF = 5.34$). This was also true if we look at the tests of 2-RI ($BF = 4.77$) and 3-RI ($BF = 5.12$). If we look at the eight tests separately, the data did not provide much support for either the null or the alternative. The exception was the third test of 3-RI which provided very strong evidence for the alternative that reduction invariance did not hold and the first test of 3-RI which provided some evidence for reduction invariance.

Table 5.3 shows the classification of the subjects. Reduction invariance was the dominant type with 45% of the subjects satisfying it in both tests. No other type was close to reduction invariance. Both in the tests of 2-RI and in the tests of 3-RI around 60% of the subjects satisfied reduction invariance. Two thirds of the subject could be classified the same way in both the 2-RI and the 3-RI tests. The data support the hypothesis that amongst the subjects who could be classified the same way those who behaved according to reduction invariance were more common than those who did not behave according to reduction invariance ($BF = 3.81$).

Table 5.3 Classification of subjects in the 2-reduction invariance (2-RI) and the 3—reduction invariance (3-RI) tests

Type		**2-RI**			Total
		compound > simple	RI	compound < simple	
3-RI	compound > simple	6	8	0	14
	RI	8	35	5	48
	compound < simple	1	4	11	16
Total		15	47	16	78

Tests of reduction of compound gambles

The general picture that emerges from our results is that reduction invariance was supported. This poses the question whether the special, rational case of reduction invariance, reduction of compound gambles, also held. Our results indicate that it did not hold at the aggregate level. Figure 5.1 gives a visual impression. The circles show the certainty equivalents of the compound gambles plotted against the reduced probabilities. If reduction of compound gambles held the circles and the squares should overlap. It is clear from the Figure that they did not. Bayesian tests revealed very strong evidence for the alternative hypothesis that reduction of compound gambles did not hold ($BF = 1.14e-23$). [28]

However, at the individual level we observed that around 47 (60%) of the subjects behaved in line with reduction of compound gambles (taking account of preference imprecision). The subjects who deviated from it, deviated overwhelmingly in the direction of higher certainty equivalents for the compound gambles than for the corresponding simple gambles (according to the Bayes factors the posterior probability that a subject who deviated from reduction of compound gambles had a higher certainty equivalent for the compound gamble was 5642 times as high as the probability that he had a higher certainty equivalent for the corresponding simple gamble).

Robustness

We used linear interpolation in the analysis of reduction invariance to determine $CE(200, r^2)$ and $CE(200, r^3)$. A problem in this analysis was that we could not always determine r uniquely. We, therefore, also used interpolation by smoothing splines, a nonparametric regression technique which smoothens out response errors. The fit was good for most subjects.

The figures for this robustness check are in Appendix 5.3. Overall, the robustness check led to the same conclusions as the analysis using linear interpolation. Based on the pooled data, the support for reduction invariance increased compared to the analysis using linear interpolation ($BF = 8.63$). The results of the separate tests were largely similar to those under linear interpolation except that in the third test of 2-RI we now also observed some evidence that reduction invariance did not hold. The support against reduction invariance in the third test of 3-RI decreased from very strong evidence to some evidence.

---

[28] The pairwise tests supported the alternative hypothesis that reduction of compound gambles did not hold with Bayes factors less than .33 except for the differences between C6 and S2 ($BF = 3.14$) and between C8 and S2 ($BF = 5.70$) where the data gave some evidence for reduction of compound gambles and the differences between C11 and S3 ($BF = 0.68$), C2 and S4 ($BF = 1.07$), and C4 and S4 ($BF = 0.94$) where the data supported neither the null nor the alternative hypothesis.

At the individual level, reduction invariance was still clearly the dominant pattern and the numbers were close to those observed under linear interpolation.

5.5 Discussion

Our data largely supported reduction invariance, the central condition underlying Prelec's (1998) compound invariant weighting function. At the aggregate level our data provided some evidence in favor of reduction invariance and at the individual level reduction invariance was clearly the dominant pattern. The only test in which we found strong evidence for the alternative hypothesis that reduction invariance did not hold was the third test of 3-RI. We do not know why this happened. The reduced probability in the third test of 3-RI was similar to that in the first test of 3-RI where we found evidence for reduction invariance. The fact that in the third test of 3-RI p was less than q cannot explain the observed violation of reduction invariance either as this was also true in, for example, the second test of 2-RI where the null of reduction invariance was supported over the alternative.

Our tests of reduction invariance require the use of measured certainty equivalents. Luce (2000) argues that certainty equivalents may lead to biased estimations of the subjective values of gambles due to inherently different attitudes towards gambles (multi-dimensional entities) and certain money amounts (one-dimensional entities). Von Nitzsch and Weber (1988) demonstrated empirical evidence of this bias. This problem could be avoided by matching gambles with gambles, i.e. by directly eliciting r such that $((x,p),q)\sim(x,r)$ and then checking whether $((x,p^N),q^N)\sim(x,r^N), N=2{,}3$. As Luce (2001) pointed out, this test carries the risk that subjects will give the salient answer $pq = r$ in spite of the many observed empirical violations of reduction of compound gambles. We, therefore, followed Luce's (2001) suggestion to use certainty equivalents in the tests of reduction invariance. To reduce possible distortions, we used a choice-based procedure to determine the certainty equivalents. Previous evidence suggests that observed anomalies are substantially reduced when choice-based certainty equivalents are used instead of judged certainty equivalents (Bostic et al., 1990; von Winterfeldt el., 1997). The procedure we used is close to the PEST procedure used by Luce in his experimental research (Chung et al., 1994; Cho et al., 1994, Cho & Luce, 1995).

We used several ways to account for the stochastic nature of people's preferences. Rather than testing equality of certainty equivalents we followed Cho & Luce (1995) and tested whether the proportion of subjects for whom $CE\big((200, p^N), q^N\big)$ exceeded $CE(200, r^N)$ was the same as the proportion of subjects for whom $CE\big((200, p^N), q^N\big)$ was less than $CE(200, r^N)$. Moreover, we accounted for the imprecision in our measurements and in the individual analyses we only required preference patterns to hold in a majority of cases. There exist different and more sophisticated procedures to model choice errors. For example, Davis-Stober (2009) derived statistical tests based on order-constrained inference techniques, which were applied, amongst others in Regenwetter, Dana, & Davis-Stober (2011) to test transitivity and in Davis-Stober, Brown, & Cavagnaro (2015) to compare models based on strict weak order representations with those based on lexicographic semiorder representations. It is interesting to repeat our analysis using these methods, but it should be realized that they are, to the best of our knowledge, not yet applicable to matching tasks and that they require each choice to be repeated many times. In our experiment subjects made around 100 choices, but if we were to use the same amount of repetitions as Regenwetter et al. (2011) or Regenwetter & Davis-Stober (2012) did, subjects would have to make more than 2000 choices, which might reduce accuracy.

We found mixed support for reduction of compound gambles, the rational special case of reduction invariance. The condition was clearly violated at the aggregate level, but 60% of the subjects behaved in line with it. The violations of reduction of compound gambles that we observed indicate that subjects generally preferred compound gambles to simple gambles giving the same reduced probability. This compound risk seeking is consistent with Friedman (2005) and Kahn & Sarin (1988). It could be explained by a utility of gambling (Luce & Marley, 2000; Luce, Ng, Marley, & Aczél, 2008) as the compound gambles offer the possibility to gamble twice. On the other hand, Abdellaoui, Klibanoff, & Placido (2015) observed that their subjects were compound risk averse and preferred simple gambles with the same reduced probability. They also observed that subjects became more compound risk averse for higher probabilities, while we observed the opposite pattern. The range of probabilities Abdellaoui et al. explored is larger than the range we explored. Moreover, the compound gambles for which they found compound risk aversion were more complex than the compound gambles we used and it was more difficult for their subjects to compute the reduced probabilities. Complexity aversion may have contributed to compound risk aversion in their study.

We obtained some evidence that when choosing between two gambles with the same expected value, subjects preferred the gamble with the higher second-stage probability to the gamble with the higher first-stage probability. This is consistent with a preference to have most uncertainty resolved at the first stage and violates event commutativity (Luce, 2000). We found very strong evidence that the certainty equivalent of C7, which offered a higher probability at the second stage, was higher than the certainty equivalent of C4, which offered approximately the same reduced probability but a higher first-stage probability (according to the Bayes factors, the posterior probability that CE(C7) > CE(C4) was 471 times as high as the probability that CE(C7) < CE(C4)). More support for a preference to have the high probability resolved later comes from a comparison of compound gambles C1 and C3, which were also close in reduced probability. We found very strong evidence that the certainty equivalent of C3, which offered a larger second-stage probability exceeded that of C1, which offered a larger first-stage probability (odds 56.93). On the other hand, we also found strong evidence that the certainty equivalent of gamble C5 exceeded the certainty equivalent of gamble C2 (odds 20.41), which is inconsistent with a preference to have the high probability resolved later. As mentioned above, Ronen (1973) and Budescu & Fischer (2001) obtained clear evidence to have the high probability resolved first. Budescu & Fischer (2001) observed that hope was an important reason why their subjects preferred higher initial probabilities. A typical reason subjects gave was that "the progress from one stage to the other means something, it's better to lose at a later stage". Apparently, such considerations played no role in our study or they were offset by other considerations such as disappointment aversion which predicts that the high probability will be resolved later.

5.6 Conclusion

Prelec's (1998) compound-invariant family provides a simple way to model deviations from expected utility. It has a preference foundation, its parameters are intuitive, and it has often been used in empirical research. Luce (2001) gave an elegant simplification of Prelec's central condition and our study showed evidence in support of Luce's central condition, reduction invariance. This implies that Prelec's function provides an accurate description of the way people weight probabilities and endorses its use in empirical research. Reduction of compound gambles, a special case of reduction invariance, which is often considered rational, was rejected at the aggregate level, even though 60% of the subjects behaved in line with it implying that the power probability weighting function, which depends on reduction of compound gambles, should be used with caution.

Appendix 5.1

Instructions and comprehension questions

Instructions

Welcome!

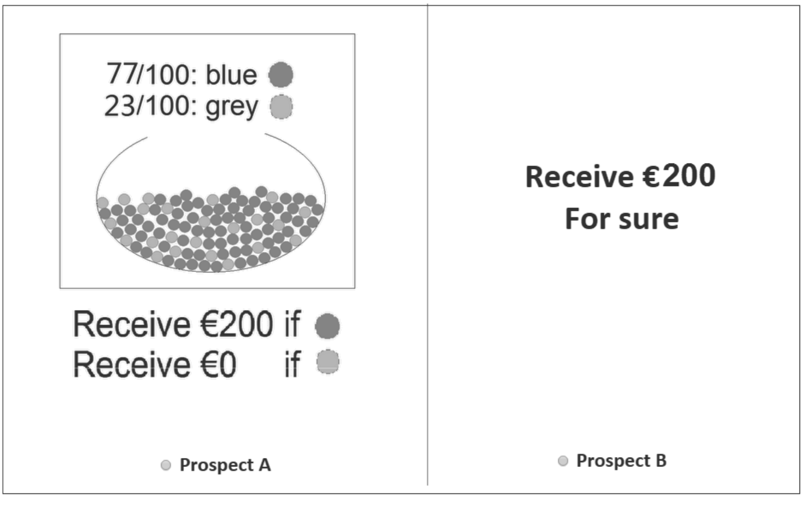
During this experiment, you will face different choice situations involving risk. In each situation, you are asked to choose between two prospects:

- Prospect A gives you an amount of money contingent on the color of a ball drawn from an urn.
- Prospect B gives you an amount of money for sure.

The outcome of Prospect A can depend on a single draw from an urn or on two consecutive draws from two different urns.

Figure A5.1 shows an example of the first scenario where the outcomes of Prospect A is determined by a single draw from an urn.

Figure A5.1



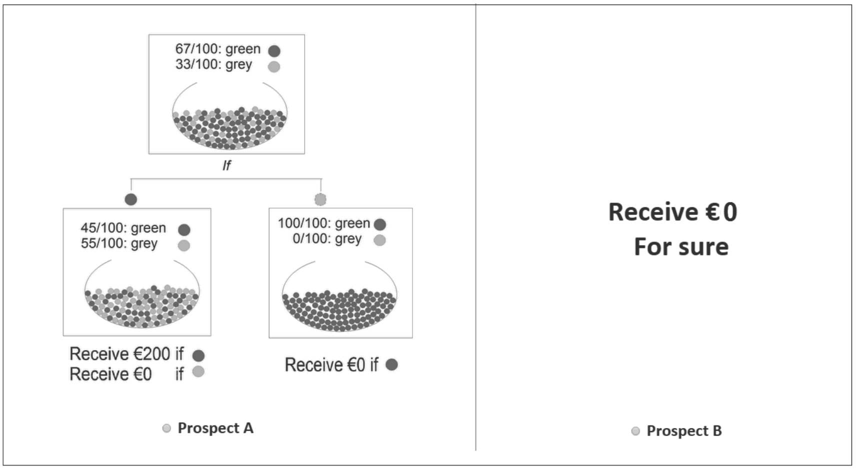
In this choice situation, there are 100 balls in the urn, of which 77 are blue, and 23 are grey. If the drawn ball is blue, you receive €200; if it is grey, you receive €0.

On the other hand, Prospect B gives you €200 for sure.

In this example, you would prefer Prospect B, because it gives you €200 for sure whereas receiving

the same amount is not certain in Prospect A.
Figure A5.2 presents an example of the second scenario where the outcomes of Prospect A depends on two draws.

Figure A5.2



In this choice situation, the first draw is made from the urn displayed on the top which contains 67 green and 33 grey balls. If the ball is green, then a second ball will be drawn from the left urn below; otherwise it will be drawn from the right urn below.

The final outcome will be determined by the second ball. For instance, if the second ball is drawn from the left urn, then a green ball will result in €200, and a grey ball will result in €0. If the second ball is drawn from the right urn, then the outcome will be €0 for sure because all balls in the right urn are green.

On the other hand, Prospect B gives you €0 for sure.

In this example, you would prefer Prospect A, because it gives you a positive chance of receiving €200 whereas Prospect B gives you €0 for sure.

Once you have made your choice between two prospects, a confirm button will appear. If you agree with your choice, please click on it to go to the next question. You will not be able to change your choice after you click on the "Confirm" button.

Payment

To thank for your participation, you will receive a €5 show-up fee.

In addition, two participants in this room will play out one of her choices for real. They will be selected randomly at the end of the experiment. For each of the selected participants, one of the choice situations that she faced during the experiment will be randomly selected, and his/her choice in that choice situation will be played for real.

We will now test your understanding of the instructions.
Assume that you have been selected as one of the two participants who can play a question for real and that the question below was randomly selected.

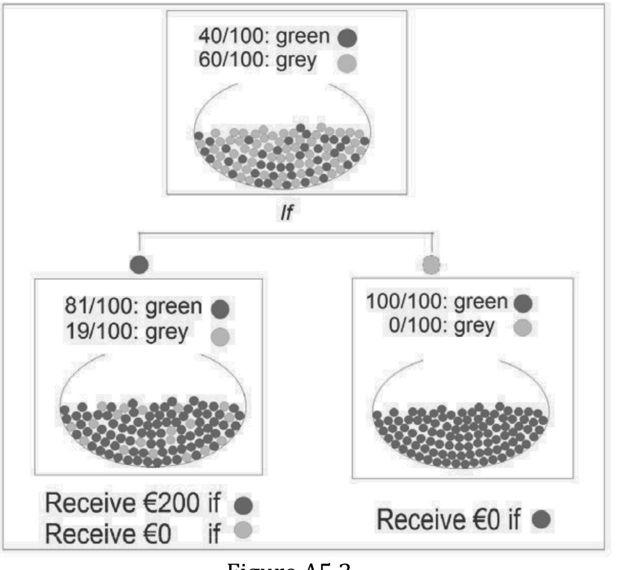


Figure A5.3

Please answer the following questions.

Question 1
How many balls are there in each urn?

- 40
- 60
- 100
- 81

Question 2
In the top urn, how many green balls are there?

- 40
- 60
- 100
- 81

Question 3
In which case will you receive €200?

- Draw a grey ball in the top urn, OR draw a green ball in the bottom left urn.
- Draw a green ball in the top urn, OR draw a green ball in the bottom left urn.
- Draw a grey ball in the top urn, AND draw a green ball in the bottom left urn.
- Draw a green ball in the top urn, AND draw a green ball in the bottom left urn.

Appendix 5.2

Iteration procedure

Subjects always chose between a gamble and a sure amount x.

1. The initial value of x was the even number closest to the expected value of the gamble.
2. X was decreased when it was chosen over the gamble and increased when the lottey was chosen.
3. The initial step size was 4, 8, 16, or 32. By choosing powers of 2 we ensured that subsequent changes were also integers. The initial step size was the number in the set {4, 8, 16, 32} that was closest to half the initial value.
4. The step size remained constant until the subjects switched. Then it was halved.
5. The minimum step size was 2. The switching point was the midpoint between the largest value of x for which the gamble was preferred and the smallest value of x for which x was preferred.
6. If a subject had to choose between 200 for sure and the gamble or between 0 for sure and the gamble. If subjects chose the dominated option, a warning message appeared: “Please reconsider your choice”. The subject was asked to

choose again. If the subject continues to choose the dominated choice, we proceeded to the next elicitation.

Table A5.1 shows the initial values and the initial step sizes for the eighteen gambles in the experiment.

Table A5.1 Initial values and initial step sizes for the gambles in the experiment

Gamble	Expected value	Initial value	Initial step size
C1	109.88	110	32
C2	60.30	60	32
C3	113.40	114	32
C4	63.96	64	32
C5	60.30	60	32
C6	18	18	8
C7	64.80	64	32
C8	20.10	20	8
C9	33	32	16
C10	5.40	6	4
C11	36.50	36	16
C12	6.60	6	4
S1	6	6	4
S2	18	18	8
S3	34	34	16
S4	64	64	32
S5	114	114	32
S6	154	154	32

Appendix 5.3

Tests of reduction invariance under fitting of the certainty equivalents by smoothing splines

Figure A5.4.1 Tests of 2-reduction invariance

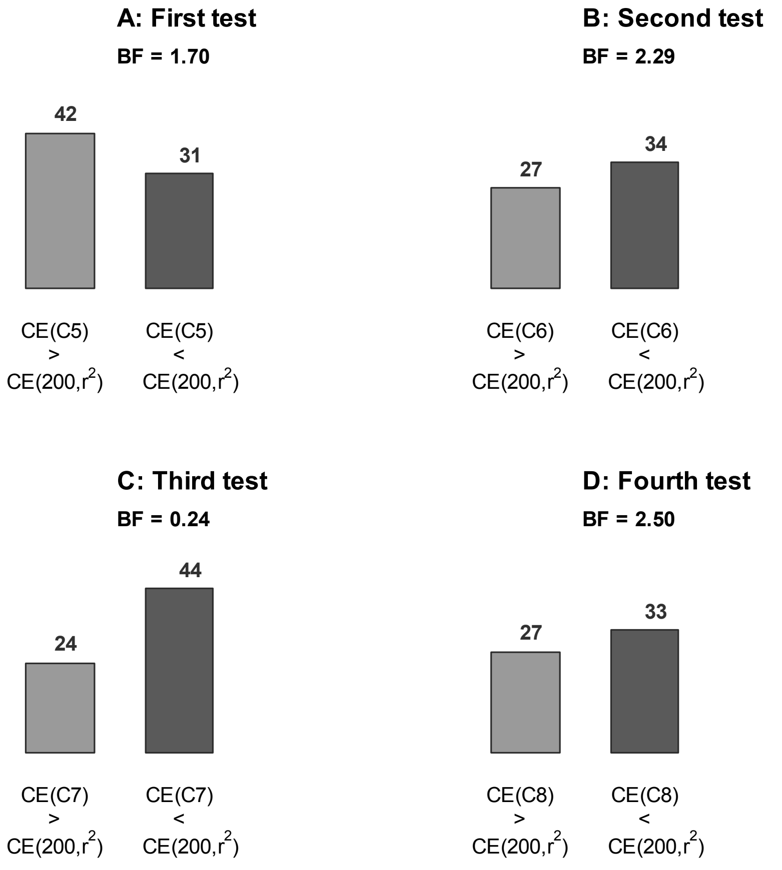


Figure A5.4.2 Tests of 3-reduction invariance

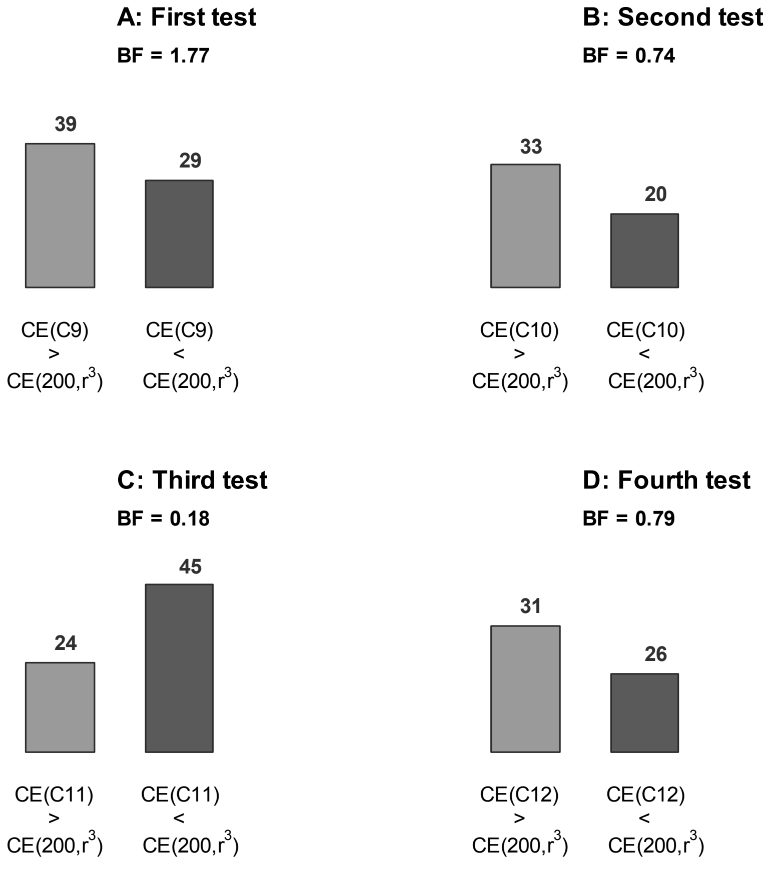


Table A5.2 Classification of subjects in the 2-reduction invariance (2-RI) and the 3-reduction invariance (3-RI) tests

Type		**2-RI**			Total
		compound > simple	RI	compound < simple	
3-RI	compound > simple	8	6	1	15
	RI	7	30	10	47
	compound < simple	1	4	11	16
Total		16	40	22	78

Summary and Conclusions

The predominant view in the literature on risk and uncertainty is that people overweight rare outcomes. This view was challenged by Barron & Erev (2003) and Hertwig et al. (2004). These studies claimed that the opposite of overweighting is found when people make decisions from experience, and that decisions from experience (DFE) is a more realistic representation of decisions that we often make in life than the traditional paradigm of decisions from description (DFD). However, the issue of when and why rare outcomes receive too little or too much attention has been controversial.

To understand the intriguing behavioural gap between the two choice paradigms, this dissertation examines decisions from experience and from description by investigating the role of probability weighting and subjective beliefs. In Chapter 2 the role of probability weighting under DFE is clarified by controlling for the impact of beliefs, ambiguity, and utility curvature. The results of the experiment indicate a clear de-biasing effect of sampling experience when there is no sampling error or ambiguity: learning from experience attenuates, but does not reverse, overweighting of rare outcomes.

In Chapter 3, the importance of prior beliefs in understanding DFE is discussed. Here, it is claimed that the sources of uncertainty that the subjects face in the original DFE experiments is a case of *ambiguity,* rather than of *complete ignorance*. Therefore, every subject has his own subjective prior beliefs about the decision environment. The findings reported in this chapter indicate that the DFD-DFE gap has two components: (1) regressive probability estimations due to the updating of prior beliefs and (2) overweighting of small probabilities due to the impact of ambiguity. The findings of Chapter 2 and Chapter 3 suggest that the underweighting of rare and extreme outcomes, or the so-called black swans neglect, reported in the previous DFE studies is a problem of knowledge: when people come to know black swans by observation, or by anticipation, they do not neglect them.

The problem of probabilistic inference is investigated further in Chapter 4. The non-Bayesian model of updating formulated in this chapter provides a natural interpretation of well-known biases in belief updating. The structural estimations of the model parameters indicate that subjects were conservative and acted as if they missed (on average) 65% of the signals generated by a random mechanism, although this rate was lower when the signals were more informative. The subjects also exhibited confirmatory bias by misreading 17% of the signals contradicting their prior beliefs.

Chapter 5 reports an empirical test of Prelec's (1998) compound invariance family. Luce (2000) writes "As a meta-theoretical principle, I hold that any descriptive theory of decision making should always include as special cases any locally rational theory for the same topic" (p. 108). Thus, he gives a meta-theoretical argument for Prelec's (1998) theory of probability weighting as it accommodates the normative condition of *reduction of compound lotteries* as a special case. The experiment in this chapter tests Luce's (2001) *reduction invariance* condition characterizing Prelec's family, which is easier to test empirically than Prelec's *compound invariance* condition. The data supports the *reduction invariance* condition both at the aggregate and at the individual level. Hence, the results provide an empirical justification for Prelec's compound invariance family.

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., & Wakker, P. P. (2011). The rich domain of uncertainty: Source functions and their experimental implementation. *The American Economic Review, 101*(2), 695-723.

Abdellaoui, M., Bleichrodt, H., & Paraschiv, C. (2007). Loss aversion under prospect theory: A parameter-free measurement. *Management Science, 53*(10), 1659-1674.

Abdellaoui, M., Bleichrodt, H., Kemel, E., & L'Haridon, O. (2014). Beliefs and attitudes for natural sources of uncertainty. *Working Paper*.

Abdellaoui, M., Klibanoff, P., & Placido, L. (2015). Experiments on compound risk in relation to simple risk and to ambiguity. *Management Science, 61*(6), 1306-1322.

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

Abdellaoui, M., Vossmann, F., & Weber, M. (2005). Choice-based elicitation and decomposition of decision weights for gains and losses under uncertainty. *Management Science, 51*(9), 1384-1399.

Aczél, J., & Luce, R. D. (2007). A behavioral condition for Prelec's weighting function on the positive line without assuming W(1)=1. *Journal of Mathematical Psychology, 51*(2), 126-129.

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'ecole Americaine. *Econometrica, 21*, 503-546.

Andrews, R. J., Logan, T. D., & Sinkey, M. J. (2015). Indentifying confirmatory bias in the field evidence from a poll of experts. *Journal of Sports Economics*, 1527002515617511.

Ashby, N. J., & Rakow, T. (2014). Forgetting the past: Individual differences in recency in subjective valuations from experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(4), 1153.

Ayton , P., & Fischer, I. (2004). The hot hand fallacy and the gambler's fallacy: Two faces of subjective randomness? *Memory & Cognition, 32*(8), 1369-1378.

Baillon, A. (2008). Eliciting subjective probabilities through exchangeable events: An advantage and a limitation. *Decision Analysis, 5*(2), 76-87.

Baillon, A., Bleichrodt, H., Keskin, U., L'Haridon, O., & Li , C. (in press). The effect of learning on ambiguity attitudes. *Management Science*.

Balcombe, K., & Fraser, I. (2015). Parametric preference functionals under risk in the gain domain: A Bayesian analysis. *Journal of Risk and Uncertainty, 50*(2), 161-187.

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

Bar-Hillel, M. (1973). On the subjective probability of compound events. *Organizational Behavior and Human Performance, 9*(3), 396-406.

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

Barron, G., & Ursino, G. (2013). Underweighting rare events in experience based decisions: Beyond sampling error. *Journal of Economic Psychology, 39*.

Barron, G., & Yechiam, E. (2009). The coexistence of overestimation and underweighting of rare events and the contingent recency effect. *Judgment and Decision Making, 4*(6), 447.

Bénabou, R., & Tirole, J. (2002). Self-confidence and personal motivation. *The Quarterly Journal of Economics, 117*(3), 871-915.

Bénabou, R., & Tirole, J. (2006). Belief in a just world and redistributive politics. *The Quarterly Journal of Economics, 121*(2), 699-746.

Bénabou, R., & Tirole, J. (2016). Mindful economics: The production, consumption, and value of beliefs. *The Journal of Economic Perspectives, 30*(3), 141-164.

Bernasconi, M., & Loomes, G. (1992). Failures of the reduction principle in an Ellsberg-type problem. *Theory and Decision, 32*(1), 77-100.

Bleichrodt, H., & Pinto, J. L. (2000). A parameter-free elicitation of the probability weighting function in medical decision analysis. *Management Science, 46*(11), 1485-1496.

Bleichrodt, H., Cillo, A., & Diecidue, E. (2010). A quantitative measurement of regret theory. *Management Science, 56*(1), 161-175.

Bleichrodt, H., Kothiyal, A., Prelec, D., & Wakker, P. (2013). Compound invariance implies prospect theory for simple prospects. *Journal of Mathematical Psychology, 57*(3), 68-77.

Booij, A. S., van Praag, B. M., & van de Kuilen, G. (2010). A parametric analysis of prospect theory's functionals for the general population. *Theory and Decision, 68*(1-2).

Bostic, R., Herrnstein, R. J., & Luce, R. D. (1990). The effect on the preference-reversal phenomenon of using choice indifferences. *Journal of Economic Behavior & Organization, 13*(2), 193-212.

Bruhin, A., Fehr-Duda, H., & Epper, T. (2010). Risk and rationality: Uncovering heterogeneity in probability distortion. *Econometrica, 78*(4), 1375-1412.

Budescu, D. V., & Fischer, I. (2001). The same but different: An empirical investigation of the reducibility principle. *Journal of Behavioral Decision Making, 14*(3), 187-206.

Camilleri, A. R., & Newell, B. R. (2009). The role of representation in experience-based choice. *Judgment and Decision Making, 4*(7), 518.

Camilleri, A. R., & Newell, B. R. (2011). Description-and experience-based choice: Does equivalent information equal equivalent choice? *Acta Psychologica, 136*(3), 276-284.

Camilleri, A. R., & Newell, B. R. (2011). When and why rare events are underweighted: A direct comparison of the sampling, partial feedback, full feedback and description choice paradigms. *Psychonomic Bulletin & Review, 18*(2), 377-384.

Carnap, R. (1945). On inductive logic. *Philosophy of Science, 12*(2), 72-97.

Carnap, R. (1950). *Logical Foundations of Probability.* University Press, Chicago.

Carnap, R. (1952). *The Continuum of Inductive Methods.* University Press, Chicago.

Chateauneuf, A., Eichberger, J., & Grant, S. (2007). Choice under uncertainty with the best and the worst in mind: Neo-additive capacities. *137*(1), 538-567.

Chechile, R. A., & Barch, D. H. (2013). Using logarithmic derivative functions for assessing the risky weighting function for binary gambles. *Journal of Mathematical Psychology, 57*(1), 15-28.

Chew, S. H., & Sagi, J. S. (2008). Small worlds: Modeling attitudes toward sources of uncertainty. *Journal of Economic Theory, 139*(1), 1-24.

Cho, Y., & Luce, R. D. (1995). Tests of hypotheses about certainty equivalents and joint receipt of gambles. *Organizational Behavior and Human Decision Processes, 64*(3), 229-248.

Cho, Y., Luce, R. D., & Von Winterfeldt, D. (1994). Tests of assumptions about the joint receipt of gambles in rank-and sign-dependent utility theory. *Journal of Experimental Psychology: Human Perception and Performance, 20*(5), 931.

Christandl, F., Fetchenhauer, D., & Hoelzl, E. (2011). Price perception and confirmation bias in the context of a VAT increase. *Journal of Economic Psychology, 32*(1), 131-141.

Chung, N.-K., Von Winterfeldt, D., & Luce, R. D. (1994). An experimental test of event commutativity in decision making under uncertainty. *Psychological Science, 5*(6), 394-400.

Coutts, A. (2016). Good news and bad news are still news: Experimental evidence on belief updating. *7th International Conference of the French Association of Experimental Economics (ASFEE).*

Cubitt, R., Kopsacheilis, O., & Starmer, C. (2016). Decomposing description-experience gap. *Manuscript in preparation*.

Currim, I. S., & Sarin, R. K. (1989). Prospect versus utility. *Management Science, 35*(1), 22-41.

Davis-Stober, C. P. (2009). Analysis of multinomial models under inequality constraints: Applications to measurement theory. *Journal of Mathematical Psychology, 53*(1), 1-13.

Davis-Stober, C. P., Brown, N., & Cavagnaro, D. R. (2015). Individual differences in the algebraic structure of preferences. *Journal of Mathematical Psychology, 66*, 70-82.

De Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. *Annales de l'institut Henri Poincaré*, *7*, pp. 1-68.

Denrell, J., & March, J. G. (2001). Adaptation as information restriction: The hot stove effect. *Organization Science, 12*(5), 523-538.

Diecidue, E., Schmidt, U., & Zank , H. (2009). Parametric weighting functions. *Journal of Economic Theory, 144*(3), 1102-1118.

Dimmock, S. G., Kouwenberg, R., & Wakker, P. (2015). Ambiguity attitudes in a large representative sample. *Management Science, 62*(5), 1363-1380.

Dimmock, S. G., Kouwenberg, R., Mitchell, O. S., & Peijnenburg, K. (2015). Estimating ambiguity preferences and perceptions in multiple prior models: Evidence from field. *Journal of Risk and Uncertainty, 51*(3), 219-244.

Edwards, W. (1968). Conservatism in human information processing. *Formal representation of human judgment, 17*, 51.

Eil, D., & Rao, J. M. (2011). The good news-bad news effect: Asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics, 3*(2), 114-138.

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

Epstein, L. G. (2006). An axiomatic model of non-Bayesian updating. *The Review of Economic Studies, 73*(2), 413-436.

Erev, I., & Barron, G. (2005). On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological Review, 112*(4), 912.

Erev, I., Ert, E., Plonsky, O., Cohen, D., & Cohen, O. (2015). From anomalies to forecasts: A choice prediction competition for decisions under risk and ambiguity. *Mimeo*, 1-56.

Erev, I., Ert, E., Roth, A. E., Haruvy, E., Herzog, S. M., Hau, R., . . . Lebiere, C. (2010). A choice prediction competition: Choice from experience and from description. *Journal of Behavioral Decision Making, 23*(1), 15-47.
Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over-and underconfidence: The role of error in judgment processes. *Psychological Review, 101*(3), 519.
Ert, E., & Trautmann, S. (2014). Sampling experience reverses preferences for ambiguity. *Journal of Risk and Uncertainty, 49*(1), 31-42.
Ertac, S. (2011). Does self-relevance affect information processing? Experimental evidence on the response to performance and non-performance feedback. *Journal of Economic Behavior & Organization, 80*(3), 532-545.
Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin, 53*(2), 134.
Estes, W. K. (2002). Traps in the route to models of memory and decision. *Psychonomic Bulletin & Review, 9*(1), 3-25.
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.
Etchart-Vincent, N. (2009). Probability weighting and the 'level' and 'spacing' of outcomes: An experimental study over losses. *Journal of Risk and Uncertainty, 39*(1), 45-63.
Fehr-Duda, H., De Gennaro, M., & Schubert, R. (2006). Gender, financial risk, and probability weights. *Theory and Decision, 60*(2-3), 283-313.
Fiedler, K., & Unkelbach, C. (2014). Regressive judgment implications of a universal property of the empirical world. *Current Directions in Psychological Science, 23*(5), 361-367.
Fiedler, K., Unkelbach, C., & Freytag, P. (2009). On splitting and merging categories: A regression account of subadditivity. *Memory & Cognition, 37*(4), 383-393.
Fox, C. R., & Hadar, L. (2006). "Decisions from experience" = sampling error + prospect theory: Reconsidering Hertwig, Barron, Weber & Erev (2004). *Judgment and Decision Making, 1*(2), 159.
Fox, C. R., & Poldrack, R. A. (2014). Prospect theory and the brain. In P. Glimcher, & F. Ernst, *Handbook of Neuroeconomics* (2nd ed., pp. 533-567). New York: Elsevier.
Fox, C. R., & Tversky, A. (1998). A belief-based account of decision under uncertainty. *Management Science, 44*(7), 879-895.
Fox, C. R., Rogers, B. A., & Tversky, A. (1996). Options traders exhibit subadditive decision weights. *Journal of Risk and Uncertainty, 13*(1), 5-17.
Frey, R., Mata, R., & Hertwig, R. (2015). The role of cognitive abilities in decisions from experience: Age differences emerge as a function of choice set size. *Cognition, 142*, 60-80.
Friedman, Z. G. (2005). Evaluating risk in sequential events: Violations in expected utility theory. *Washington University Undergraduate Research Digest, 1*, pp. 29-39.
Ghirardato, P., Maccheroni, F., & Marinacci, M. (2004). Differentiating ambiguity and ambiguity attitude. *Journal of Economic Theory, 118*(2), 133-173.
Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102*(4), 684.
Gilboa, I., & Schmeidler, D. (1989). Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics, 18*(2), 141-153.

Glöckner, A., Fiedler, S., Hochman, G., Ayal, S., & Hilbig, B. (2012). Processing differences between descriptions and experience: A comparative analysis using eye-tracking and physiological measures. *Frontiers in Psychology, 3*, 173.

Glöckner, A., Hilbig, B. E., Henninger, F., & 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.

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

Gonzalez, C., & Dutt, V. (2011). Instance-based learning: Integrating sampling and repeated decisions from experience. *Psychological Review, 118*(4), 523.

Gonzalez, R., & Wu, G. (1999). On the shape of the probability weighting function. *Cognitive Psychology, 38*(1), 129-166.

Gottlieb, D. A., Weiss, T., & Chapman, G. B. (2007). The format in which uncertainty information is presented affects decision biases. *Psychological Science, 18*(3), 240-246.

Grether, D. M. (1980). Bayes' rule as a descriptive model: The representativeness heuristic. *The Quarterly Journal of Economics, 95*, 537-557.

Griffin, D., & Tversky, A. (1992). The weighting of evidence and the determinants of confidence. *Cognitive Psychology, 24*(3), 411-135.

Gul, F. (1991). A theory of disappointment aversion. *Econometrica, 59*(3), 667-686.

Hadar, L., & Fox, C. R. (2009). Information asymmetry in decision from description versus decision from experience. *Judgment and Decision Making, 4*(4), 317.

Harrison, G. W. (2008). Maximum likelihood estimation of utility functions using Stata. *Working Paper*.

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The elements of statistical learning: Data mining, inference and prediction* (2nd ed.). Berlin: Springer.

Hau, R., Pleskac, T. J., & 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., & 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., & Erev, I. (2009). The description-experience gap in risky choice. *Trends in Cognitive Sciences, 13*(12), 517-523.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science, 15*(8), 534-539.

Hilbig, B. E., & Glöckner, A. (2011). Yes, they can! Appropriate weighting of small probabilities as a function of information acquisition. *Acta Psychologica, 138*(3), 390-396.

Hogarth, R. M., & Soyer, E. (2011). Sequentially simulated outcomes: Kind experience versus nontransparent description. *Journal of Experimental Psychology: General, 140*(3), 434.

Hogarth, R. M., Lejarraga, T., & Soyer, E. (2015). The two settings of kind and wicked learning environments. *Current Directions in Psychological Science, 24*(5), 379-385.

Holt, C. A. (2007). *Markets, games, and strategic behavior: recipes for interactive learning.* Pearson Addison Wesley.

Humphrey, S. J. (2006). Does learning diminish violations of independence, coalescing and monotonicity? *Theory and Decision, 61*(2), 93-128.

Hurwicz, L. (1951). Some specification problems and applications to econometric models. *Econometrica, 19*(3), 343-44.
Jeffreys, H. (1961). *Theory of probability.* New York: Oxford University Press.
Jessup, R. K., Bishara, A. J., & Busemeyer, J. R. (2008). Feedback produces divergence from prospect theory in descriptive choice. *Psychological Science, 19*(10), 1015-1022.
Johnson, C., Baillon, A., Bleichrodt, H., Li, Z., Van Dolder, D., & Wakker, P. P. (2014). Prince: An improved method for measuring incentivized preferences. *Available at SSRN.*
Kahn, B. E., & Sarin, R. K. (1988). Modeling ambiguity in decisions under uncertainty. *Journal of Consumer Research, 15*(2), 265-272.
Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. In *The concept of probability in psychological experiments* (pp. 25-48). Springer.
Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*, 263-291.
Karmarkar, U. A. (1978). Subjectively weighted utility: A descriptive extension of the expected utility model. *Organizational Behavior and Human Performance, 21*(1), 61-72.
Karni, E. (2009). A mechanism for eliciting probabilities. *Econometrica, 77*(2), 603-606.
Kellen, D., Pachur, T., & Hertwig, R. (2016). How (in)variant are subjective representations of described and experienced risk and rewards? *Cognition, 157*, 126-138.
Keller, L. R. (1985). Testing of the 'reduction of compound alternatives' principle. *Omega, 13*(4), 349-358.
Kemel, E., & Travers, M. (2016). Comparing attitudes toward time and toward money in experience-based decisions. *Theory and Decision, 80*(1), 71-100.
Keren, G., & Wagenaar, W. A. (1987). Violation of utility theory in unique and repeated gambles. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*(3), 387.
Klayman, J. (1995). Varieties of confirmation bias. *Psychology of learning and motivation, 32*, 385-418.
Krawczyk, M. W. (2015). Probability weighting in different domains: The role of effect, fungibility, and states. *Journal of Economic Psychology, 51*, 1-15.
Kruschke, J. K. (2011). *Doing Bayesian Data Analysis.* New York: Academic Press.
Kvam, P. D., & Pleskac, T. J. (2016). Strength and weight: The determinants of choice and confidence. *Cognition, 152*, 170-180.
Laplace, P. S. (1825). *Essai philosophique sur les probabilites.* Bachelier.
Lattimore, P. K., Baker, J. R., & Witte, A. D. (1992). The influence of probability on risky choice: A parametric examination. *Journal of Economic Behavior & Organization, 17*(3), 377-400.
Lejarraga, T., & Gonzalez, C. (2011). Effects of feedback and complexity on repeated decisions from description. *Organizational Behavior and Human Decision Processes, 116*(2), 286-295.
Lejarraga, T., Pachur, T., Frey, R., & Hertwig, R. (2016). Decisions from experience: From monetary to medical gambles. *Journal of Behavioral Decision Making, 29*(1), 67-77.

Lopes, L. L. (1982). Doing the impossible: A note on induction and the experience of randomness. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8*(6), 626-636.
Lopes, L. L. (1985). Averaging rules and adjustment processes in Bayesian inference. *Bulletin of the Psychonomic Society, 23*(6), 509-512.
Luce, R. D. (1959). *Individual Choice Behavior.* New York: John Wiley & Sons.
Luce, R. D. (1990). Rational versus plausible accounting equivalences in preference judgments. *Psychological Science, 1*(4), 225-234.
Luce, R. D. (1991). Rank-and sign-dependent linear utility models for binary gambles. *Journal of Economic Theory, 53*(1), 75-100.
Luce, R. D. (1996). When four distinct ways to measure utility are the same. *Journal of Mathematical Psychology, 40*(4), 297-317.
Luce, R. D. (2000). *Utility of Gains and Losses: Measurement-theoretical and Experimental Approaches.* Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.
Luce, R. D. (2001). Reduction invariance and Prelec's weighting functions. *Journal of Mathematical Psychology, 45*(1), 167-179.
Luce, R. D. (2002). A psychophysical theory of intensity proportions, joint presentations, and matches. *Psychological Review, 109*(3), 520.
Luce, R. D. (2004). Symmetric and asymmetric matching of joint presentations. *Psychological Review, 111*(2), 446.
Luce, R. D., & Fishburn, P. C. (1991). Rank-and sign-dependent linear utility models for finite first-order gambles. *Journal of Risk and Uncertainty, 4*(1), 29-59.
Luce, R. D., & Fishburn, P. C. (1995). A note on deriving rank-dependent utility using additive joint receipts. *Journal of Risk and Uncertainty, 11*(1), 5-16.
Luce, R. D., & Marley, A. A. (2000). On elements of chance. *Theory and Decision, 49*(2), 97-126.
Luce, R. D., & Raiffa, H. (1957). *Games and Decisions: Introduction and Critical Survey.* Wiley, New York.
Luce, R. D., Ng, C., Marley, A., & Aczél, J. (2008). Utility of gambling II: Risk, paradoxes, and data. *Economic Theory, 36*(2), 165-187.
Machina, M. J., & Schmeidler, D. (1992). A more robust definition of subjective probability. *Econometrica, 60*(4), 745-780.
March, J. G. (1996). Learning to be risk averse. *Psychological Review, 103*(2), 309.
Marinacci, M. (2015). Model uncertainty. *Journal of European Economic Association, 13*(6), 1022-110.
Marley, A., & Luce, R. D. (2002). A simple axiomatization of binary rank-dependent utility of gains (losses). *Journal of Mathematical Psychology, 46*(1), 40-55.
Massey, C., & Wu, G. (2005). Detecting regime shifts: The causes of under - and overreaction. *Management Science, 51*(6), 932-947.
Morey, R. D., Rouder, J. N., Jamil, T., & R Core Team. (2015). *BayesFactor.* Vienna, Austria: R Foundation for Stastical Computing.
Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many disguises. *Review of General Psychology, 2*(2), 175.
Nilsson, H., Rieskamp, J., & Wagenmakers, E. J. (2011). Hierarchical Bayesian parameter estimation for cumulative prospect theory. *Journal of Mathematical Psychology, 55*(1), 84-93.

Oswald, M. E., & Grosjean, S. (2004). Confirmatory bias. In *Cognitive illusions: A handbook on fallacies and biases in thinking, judgment and memory.* Psychology Press.
Phillips, L. D., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology, 72*(3), 346.
Pleskac, J. T., & Ralph, H. (2014). Ecologically rational choice and the structure of the environment. *Journal of Experimental Psychology: General, 143*(5), 2000-2019.
Prelec, D. (1998). The probability weighting function. *Econometrica, 66*, 497-527.
Qiu, J., & Steiger, E. -M. (2011). Understanding the two components of risk attitudes: An experimental analysis. *Management Science, 57*(1), 193-199.
Quiggin, J. (1981). Risk perception and risk aversion among Australian farmers. *Australian Journal of Agricultural Economics, 25*(2), 160-169.
Quiggin, J. (1982). A theory of anticipated utility. *Journal of Economic Behavior and Organization, 3*(4), 323-343.
R Core Team. (2015). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.
Rabin, M. (2013). An approach to incorporating psychology into economics. *The Americal Economic Review, 103*(3), 617-622.
Rabin, M., & Schrag, J. L. (1999). First impressions matter: A model of confirmatory bias. *The Quarterly Journal of Economics, 114*(1), 37-82.
Raiffa, H. (1968). *Decision analysis: Introductory lectures on choices under uncertainty.* Addison-Wesley.
Rakow, T., & Newell, B. R. (2010). Degrees of uncertainty: An overview and framework for future research on experience-based choice. *Journal of Behavioral Decision Making, 23*(1), 1-14.
Rakow, T., Demes, K. A., & 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.
Ramsey, F. P. (1931). Truth and probability. *The foundations of mathematics and other logical essays*, 156-198.
Regenwetter, M., & Davis-Stober, C. P. (2012). Behavioral variability of choices versus structural inconsistency of preferences. *Psychological Review, 119*(2), 408.
Regenwetter, M., Dana, J., & Davis-Stober, C. P. (2011). Transitivity of preferences. *Psychological Review, 118*(1), 42.
Ronen, J. (1973). Effects of some probability displays on choices. *Organizational Behavior and Human Performance, 9*(1), 1-15.
Rottenstreich, Y., & Tversky, A. (1997). Unpacking, repacking, and anchoring: Advances in support theory. *Psychological Review, 104*(2), 406.
Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review, 72*, 621-642.
Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology, 56*(5), 356-374.
Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin and Review, 16*(2), 225-237.
Savage, L. J. (1954). *The Foundations of Statistics.* Wiley, New York.

Scheibehenne, B., & Pachur, T. (2015). Using Bayesian hierarchical parameter estimation to assess the generalizability of cognitive models of choice. *Psychonomic Bulletin & Review, 22*(2), 391-407.
Schotter, A., & Trevino, I. (2014). Belief elicitation in the laboratory. *Annual Review of Economics, 6*(1), 103-128.
Schunk, D., & Betsch, C. (2006). Explaining heterogeneity in utility functions by individual differences in decision modes. *Journal of Economic Psychology, 27*(3), 386-401.
Shanteau, J. C. (1970). An additive model for sequential decision making. *Journal of Experimental Psychology, 85*(2), 181.
Sidman, M. (1952). A note on functional relations obtained from group data. *Psychological Bulletin, 49*(3), 263.
Sinkey, M. (2015). How do experts update beliefs? Lessons from a non-market environment. *Journal of Behavioral and Experimental Economics, 57*, 55-63.
Slovic, P. (1969). Manipulating the attractiveness of a gamble without changing its expected value. *Journal of Experimental Psychology, 79*, 139.
Sneddon, R., & Luce, R. D. (2001). Empirical comparisons of bilinear and nonbilinear utility theories. *Organizational Behavior and Human Decision Processes, 84*(1), 71-94.
Spetzler, C. S., & Stael von Holstein, C.-A. S. (1975). Probability encoding in decision analysis. *Management Science, 22*(3), 340-358.
Stott, H. P. (2006). Cumulative prospect theory's functional menagerie. *Journal of Risk and Uncertainty, 32*(2), 101-130.
Sundali, J., & Croson, R. (2006). Biases in casion betting: The hot hand and the gambler's fallacy. *Judgment and Decision Making, 1*(1), 1.
Trautmann, S. T., & Van de Kuilen, G. (2015). Ambiguity Attitudes. In *The Wiley Blackwell Handbook of Judgment and Decision Making* (Vol. 1, pp. 89-116). John Wiley & Sons, Ltd.
Tversky, A., & Bar-Hillel, M. (1983). Risk: The long and the short. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9*(4), 713-717.
Tversky, A., & Fox, C. R. (1995). Weighing risk and uncertainty. *Psychological Review, 102*(2), 269.
Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76*(2), 105.
Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 82*(6), 1124-1131.
Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty, 5*(4), 297-323.
Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review, 101*(4), 547.
Tversky, A., & Wakker, P. (1995). Risk attitudes and decision weights. *Econometrica, 63*(6), 1255-1280.
Ungemach, C., Chater, N., & Stewart, N. (2009). Are probabilities overweighted or underweighted when rare outcomes are experienced (rarely)? *Psychological Science, 20*(4), 473-479.
Van de Kuilen, G. (2009). Subjective probability weighting and the discovered preference hypothesis. *Theory and Decision, 67*(1), 1-22.
Van de Kuilen, G., & Wakker, P. (2006). Learning in the Allais paradox. *Journal of Risk and Uncertainty, 33*(3), 155-164.

Van de Kuilen, G., & Wakker, P. (2011). The midweight method to measure attitudes toward risk and ambiguity. *Management Science, 57*(3), 582-598.
Von Neumann, J., & Morgenstern, O. (1944). *Theory of Games and Economic Behavior.* Princeton University Press.
Von Nitzsch, R., & Weber, M. (1988). Utility function assessment on a micro-computer: An interactive procedure. *Annals of Operations Research, 16*(1), 149-160.
Von Winterfeldt, D., Chung, N.-K., Luce, R. D., & Cho, Y. (1997). Tests of consequence monotonicity in decision making under uncertainty. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(2), 406.
Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p-values. *Psychonomic Bulletin & Review, 14*(5), 779-804.
Wakker, P. (2002). Decision-principles to justify Carnap's updating method and to suggest corrections of probability judgments. *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence* (pp. 544-551). Morgan Kaufmann Publishers Inc.
Wakker, P. (2004). On the composition of risk preference and belief. *Psychological Review, 111*(1), 236.
Wakker, P. (2008). Explaining the characteristics of the power (CRRA) utility family. *Health Economics, 17*(12), 1329-1344.
Wakker, P. (2010). *Prospect Theory: For Risk and Ambiguity.* Cambridge University Press.
Wakker, P., & Deneffe, D. (1996). Eliciting von Neumann-Morgenstern utilities when probabilities are distorted or unknown. *Management Science, 42*(8), 1131-1150.
Wakker, P., Erev, I., & Weber, E. U. (1994). Comonotonic independence: The critical test between classical and rank-dependent utility theories. *Journal of Risk and Uncertainty, 9*(3), 195-230.
Weber , E. U., Shafir, S., & 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.
Wilks, S. S. (1962). *Mathematical Statistics.* Wiley, New York.
Wilson, A. (2014). Bounded memory and biases in information processing. *Econometrica, 82*(6), 2257-2294.
Winkler, R. L. (1972). *An Introduction to Bayesian Inference and Decision.* New York: Holt, Rinehart and Winston.
Wulff, D. U., & Pachur, T. (2016). Modeling valuations from experience: A comment on Ashby and Rakow (2014). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(1), 158.
Wulff, D. U., Canseco, M. M., & Hertwig, R. (2016). A meta-analytic review of two modes of learning and the description-experience gap. *Manuscript submitted for publication*.
Wullf, D. U., Hills, T. T., & Hertwig, R. (2015). How short-and long-run aspirations impact search and choice in decisions from experience. *Cognition, 144*, 29-37.
Yechiam, E., & Busemeyer, J. R. (2006). The effect of foregone payoffs on underweighting small probability events. *Journal of Behavioral Decision Making, 19*(1), 1-16.
Zabell, S. L. (1982). WE Johnson's "sufficientness" postulate. *The Annals of Statistics, 10*(4), 1090-1099.
Zabell, S. L. (1989). The rule of succession. *Erkenntnis, 31*(2-3), 283-321.

Zeisberger, S., Vrecko, D., & Langer, T. (2012). Measuring the time stability of prospect theory preferences. *Theory and Decision, 72*(3), 359-386.

Samenvatting

Summary in Dutch

Het overheersende idee in de literatuur over risico en onzekerheid is dat mensen zeldzame uitkomsten te zwaar wegen. Dit idee is betwist door Barron en Erev (2003) en Hertwig et al. (2004). Deze auteurs beweren dat het tegenovergestelde van het te zwaar wegen van kansen wordt gevonden wanneer mensen besluiten nemen op basis van ervaring, en dat het nemen van besluiten op basis van ervaring (BOE) een realistischere voorstelling is van de besluiten die we vaak nemen in ons leven dan het traditionele paradigma van het nemen van besluiten op basis van een omschrijving (BOO). Het vraagstuk van wanneer en waarom er te veel of te weinig aandacht wordt besteed aan zeldzame uitkomsten is echter controversieel.

Om het intrigerende verschil in gedrag tussen de twee besluitparadigma's te begrijpen, onderzoekt deze dissertatie besluiten op basis van ervaring en op basis van een omschrijving door de rol van het kansenweging en subjectieve opvattingen over kansen te bestuderen. In Hoofdstuk 2 wordt de rol van kansenweging bij BOE verduidelijkt door te controleren voor het effect van opvattingen, ambiguïteit, en de kromming van de nutsfunctie. De resultaten van het experiment tonen een duidelijke afname in onzuiverheid door ervaring op basis van een steekproef als er geen steekproeffout of ambiguïteit is: het leren op basis van ervaring vermindert het te zwaar wegen van zeldzame uitkomsten maar keert dit niet om.

In Hoofdstuk 3 wordt het belang van a-priori-opvattingen op het begrijpen van BOE besproken. Hier wordt gesteld dat de bron van onzekerheid waarmee de proefpersonen geconfronteerd worden in de oorspronkelijke BOE experimenten een geval is van *ambiguïteit*, in plaats van *totale onwetendheid*. Iedere proefpersoon heeft daarom zijn eigen subjectieve a-priori-opvattingen over de besluitomgeving. De bevindingen beschreven in dit hoofdstuk tonen dat de BOO-BOE leemte twee onderdelen heeft: (1) regressieve kansenramingen door het bijwerken van a-priori-opvattingen en (2) het te zwaar wegen van kleine kansen door het effect van ambiguïteit. De bevindingen van Hoofdstuk 2 en Hoofdstuk 3 suggereren dat het te licht wegen van zeldzame en extreme uitkomsten, ofwel het zogeheten negeren van zwarte zwanen, zoals beschreven in eerdere

BOE onderzoeken een kennisprobleem is: wanneer mensen zwarte zwanen leren kennen door observatie, of door anticipatie, negeren ze die niet.

Het probleem van probabilistische inferentie wordt verder onderzocht in Hoofdstuk 4. Het niet-Bayesiaanse model van het bijwerken van opvattingen geformuleerd in dit hoofdstuk biedt een natuurlijke interpretatie van bekende onzuiverheden in het bijwerken van opvattingen. De structurele vergelijkingen van de modelparameters suggereren dat proefpersonen conservatief waren en zich gedroegen alsof ze (gemiddeld) 65% van de signalen opgewekt door een mechanisme voor willekeurige selectie niet opmerkten, hoewel dit percentage lager lag wanneer de signalen informatiever waren. De proefpersonen vertoonden ook tunnelvisie door het verkeerd interpreteren van 17% van de signalen die hun a-priori-opvattingen weerspraken.

Hoofdstuk 5 bevat een empirische test van Prelec's (1998) samenstellings-constantheidsfamilie. Luce (2000) schrijft: "As a meta-theoretical principle, I hold that any descriptive theory of decision making should always include as special cases any locally rational theory for the same topic" (p. 108)[29]. Hij geeft daarmee een metatheoretisch argument voor Prelec's (1998) theorie van kansenweging aangezien het de normatieve conditie van het reduceren van samengestelde loterijen als speciaal geval bevat. Het experiment in dit hoofdstuk test Luce's (2001) conditie van *reductieconstantheid* die Prelec's familie karakteriseert, welke makkelijker empirisch te testen is dan Prelec's conditie van *samenstellingsconstantheid*. De data staaft de conditie van *reductieconstantheid* op zowel het geaggregeerde als op het individuele niveau. De resultaten bieden daarmee een empirische rechtvaardiging van Prelec's samenstellingsconstantheidsfamilie.

[29] "Als een metatheoretisch principe, meen ik dat elke descriptieve theorie van besluitvorming altijd als speciaal geval een lokaal rationele theorie voor hetzelfde onderwerp moet bevatten"

The Tinbergen Institute is the Institute for Economic Research, which was founded in 1987 by the Faculties of Economics and Econometrics of the Erasmus University Rotterdam, University of Amsterdam and VU University Amsterdam. The Institute is named after the late Professor Jan Tinbergen, Dutch Nobel Prize laureate in economics in 1969. The Tinbergen Institute is located in Amsterdam and Rotterdam. The following books recently appeared in the Tinbergen Institute Research Series:

653 T.V. WANG, *The Rich Domain of Decision Making Explored: The Non-Triviality of the Choosing Process*
654 D.A.R. BONAM, *The Curse of Sovereign Debt and Implications for Fiscal Policy*
655 Z. SHARIF, *Essays on Strategic Communication*
656 B. RAVESTEIJN, *Measuring the Impact of Public Policies on Socioeconomic Disparities in Health*
657 M. KOUDSTAAL, *Common Wisdom versus Facts; How Entrepreneurs Differ in Their Behavioral Traits from Others*
658 N. PETER, *Essays in Empirical Microeconomics*
659 Z. WANG, *People on the Move: Barriers of Culture, Networks, and Language*
660 Z. HUANG, *Decision Making under Uncertainty-An Investigation from Economic and Psychological Perspective*
661 J. CIZEL, *Essays in Credit Risk, Banking, and Financial Regulation*
662 I. MIKOLAJUN, *Empirical Essays in International Economics*
663 J. BAKENS, *Economic Impacts of Immigrants and Ethnic Diversity on Cities*
664 I. BARRA, *Bayesian Analysis of Latent Variable Models in Finance*
665 S. OZTURK, *Price Discovery and Liquidity in the High Frequency World*
666 J. JI, *Three Essays in Empirical Finance*
667 H. SCHMITTDIEL, *Paid to Quit, Cheat, and Confess*
668 A. DIMITROPOULOS, *Low Emission Vehicles: Consumer Demand and Fiscal Policy*
669 G.H. VAN HEUVELEN, *Export Prices, Trade Dynamics and Economic Development*
670 A. RUSECKAITE, *New Flexible Models and Design Construction Algorithms for Mixtures and Binary Dependent Variables*
671 Y. LIU, *Time-varying Correlation and Common Structures in Volatility*
672 S. HE, *Cooperation, Coordination and Competition: Theory and Experiment*
673 C.G.F. VAN DER KWAAK, *The Macroeconomics of Banking*
674 D.H.J. CHEN, *Essays on Collective Funded Pension Schemes*
675 F.J.T. SNIEKERS, *On the Functioning of Markets with Frictions*

676 F. GOMEZ MARTINEZ, *Essays in Experimental Industrial Organization: How Information and Communication affect Market Outcomes*
677 J.A. ATTEY, *Causes and Macroeconomic Consequences of Time Variations in Wage Indexation*
678 T. BOOT, *Macroeconomic Forecasting under Regime Switching, Structural Breaks and High-dimensional Data*
679 I. TIKOUDIS, *Urban Second-best Road Pricing: Spatial General Equilibrium Perspectives*
680 F.A. FELSŐ, *Empirical Studies of Consumer and Government Purchase Decisions*
681 Y. GAO, *Stability and Adaptivity: Preferences over Time and under Risk*
682 M.J. ZAMOJSKI, *Panta Rhei, Measurement and Discovery of Change in Financial Markets*
683 P.R. DENDERSKI, *Essays on Information and Heterogeneity in Macroeconomics*
684 U. TURMUNKH, *Ambiguity in Social Dilemmas*
685 U. KESKIN, *Essays on Decision Making: Intertemporal Choice and Uncertainty*
686 M. LAMMERS, *Financial Incentives and Job Choice*
687 Z. ZHANG, *Topics in Forecasting Macroeconomic Time Series*
688 X. XIAO, *Options and Higher Order Risk Premiums*
689 D.C. SMERDON, *'Everybody's doing it': Essays on Trust, Norms and Integration*
690 S. SINGH, *Three Essays on the Insurance of Income Risk and Monetary Policy*
691 E. SILDE, *The Econometrics of Financial Comovement*
692 G. DE OLIVEIRA, *Coercion and Integration*
693 S. CHAN, *Wake Me up before you CoCo: Implications of Contingent Convertible Capital for Financial Regulation*
694 P. GAL, *Essays on the role of frictions for firms, sectors and the macroeconomy*
695 Z. FAN, *Essays on International Portfolio Choice and Asset Pricing under Financial Contagion*
696 H. ZHANG, *Dealing with Health and Health Care System Challenges in China: Assessing Health Determinants and Health Care Reforms*
697 M. VAN LENT, *Essays on Intrinsic Motivation of Students and Workers*
698 R.W. POLDERMANS, *Accuracy of Method of Moments Based Inference*
699 J.E. LUSTENHOUWER, *Monetary and Fiscal Policy under Bounded Rationality and Heterogeneous Expectations*
700 W. HUANG, *Trading and Clearing in Modern Times*
701 N. DE GROOT, *Evaluating Labor Market Policy in the Netherlands*

702 R.E.F. VAN MAURIK, *The Economics of Pension Reforms*