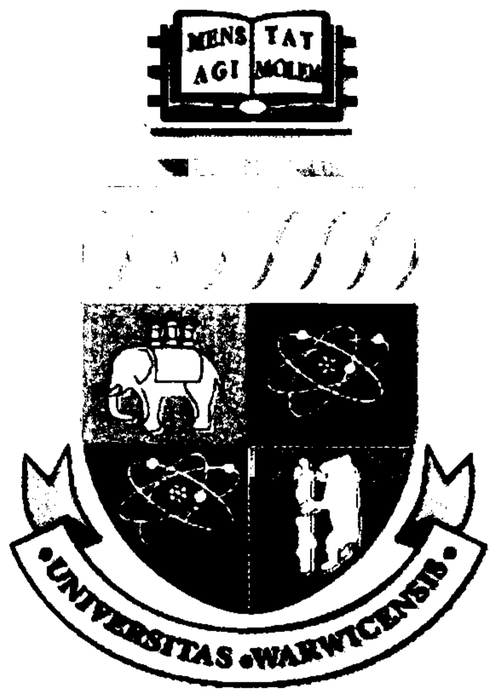
THE WEIGHT OF EXPERIENCE:

AN INVESTIGATION OF PROBABILITY WEIGHTING UNDER DECISIONS FROM EXPERIENCE

CHRISTOPH UNGEMACH

A thesis submitted for the Degree of Doctor of Philosophy to the

Department of Psychology at the University of Warwick



MARCH 2008

Contents

List of Tables

List of Figures

List of Abbreviations

DfD = decisions from description

DfXP = decisions from experience

EV = expected value theory

PT = prospect theory

Acknowledgments

I am grateful to the ESRC for funding this project. I would like to especially thank my supervisors, Nick Chater and Neil Stewart, for their continuous support and patience. They have been inspiring, motivating and intellectually stimulating at every stage and have helped make this process enjoyable as well as educational.

From my department at Warwick, I would also like to express my gratitude to my colleagues, particularly, William Jiménez-Leal, Duncan Guest, Chris Kent, Menelaos Apostolou, Caroline Morin, Luciano Buratto, James Adelman and Andrew Barnacle, who offered technical support during the tricky times. I would also like to thank all of the participants, at Warwick and beyond.

A number of others have also offered their personal and intellectual encouragement during the course of this thesis, especially, Ido Erev, Greg Barron, Robin Hau, Tim Rakow, Stein Reimers and Jean Hartley. Additionally, the team at Dectech, including Henry Stott, Richard Lewis, Benny Cheung and Greg Davies, gave me the opportunity to learn more about applied decision making research in a vibrant and humorous work environment.

I am unutterably grateful to Claire Westall for her tremendous support throughout the final stage of the preparation of this thesis. She helped me to have enjoyable breaks between work and kept me sane.

Finally, I wish to dedicate this work to my family. My mother and sister have been a constant source of loving support, kindness and consideration. I cannot express the sincerity of my thanks to them.

Declaration

I hereby declare that the research reported in this thesis is my own work, unless stated otherwise. No part of this thesis has been submitted for a degree at another University.

Some parts of this work have been presented as papers at various conferences and workshops or have been submitted for publication. They are listed as follows:

Experiment 1 and 2 were presented at the Symposium of the 21st Research Conference on Subjective Probability, Utility, and Decision Making (SPUDM21), 2007.

Experiment 2 and 7 were presented at the Annual Conference of the Society for Judgement and Decision Making, 2007.

Experiment 2, 3 and 7 and parts of Chapter 6 were presented at the One-Day Workshop 'Unravelling decisions from experience' of the European Association for Decision Making (EADM), 2008.

Substantial portions of Chapter 3 have been submitted for publication to Psychological Science.

Christoph Ungemach

March 2008

Coventry, UK

Abstract

In decisions from experience tasks objective information regarding payoffs and probabilities must be inferred from samples of possible outcomes. A series of recent experiments has revealed that people show deviating choice behaviour in such tasks, indicating underweighting of small probabilities instead of overweighting of small probabilities as in decisions from description. In a range of experiments, the research presented in this thesis provides a new direction by showing that such reversals from overweighting to underweighting in decisions from experience are very robust and can be replicated even if all the existing explanations – sampling error, recency weighting and judgement error – are experimentally controlled for. Furthermore, reversals were replicated within common decision making biases like the common ratio effect. An important, but unexpected, new finding has been the observation of a reversed reflection effect under decisions from experience. This suggests that the difference between choice behaviour may not be restricted to underlying transformations of probabilities, as suggested in the literature. Drawing from an extensive range of model tests and parameter estimations, it is also demonstrated that the differences are reflected in the best fitting parameter values for prospect theory under decisions from experience. However, it is also shown that simple reinforcement models, which provide a more intuitive rationale for experiential choice behaviour, can account for the data just as well, without any assumptions regarding the weighting of probabilities.

CHAPTER 1

DECISIONS FROM EXPERIENCE LITERATURE REVIEW

1.1 Introduction

This thesis investigates the choice phenomena observed in decisions from experience, a new strand of research that has received a lot of attention in the recent decision making literature. This first chapter provides a brief introduction to research on experience-based decision making and the concepts relevant to it. It will begin with an overview of the theoretical concepts established in decision making under risk in general and the experimental paradigm of decisions from description which has so far dominated this field. It will also present the empirical choice phenomena observed in risky choice together with the most prominent choice models that have been developed to account for these findings. Thereafter, Section 1.4 will introduce another family of decision making tasks that have been revisited in a number of recent publications, namely, experiential choice tasks that fall within the domains of decision making under risk and uncertainty. The paradigm of decisions from experience will be introduced and the literature on experience-based decision making will be summarised. The decisions from experience paradigm is the basis for the experimental work presented in this thesis. Research from related disciplines that have also investigated the impact of learning, feedback, and experience in decision making tasks will be drawn upon in the subsequent sections. Further, a review of the research on probability and frequency judgements will be offered as this will be important in order to understand the different processes that may be involved in experiential choice tasks. Finally, this

chapter will conclude with an outline of the motivation for the research of this thesis and a brief look forward to the following chapters.

1.2 Decision Making Under Risk

Before we can understand the significance of choice phenomena observed in experiential choice tasks we first need to look at the brief interdisciplinary history of decision making under risk in general and the choice paradigm that has been dominating this area of research. Since its beginning, the theoretical investigation of risky decision making has been preoccupied with the examination of various games and gambles. This is probably not a coincidence, as one of the pillars of the discipline, mathematical probability theory, is also said to originate from correspondence over a gambling problem[1] (see Gigerenzer, 1989): This tradition can be traced into the present where the problems in decision making under risk are still presented in the form of simple gambles, comprising a choice between pairs of prospects. In this context a prospect $(x_1, p_1; \ldots; x_n, p_n)$ is defined as a gamble yielding the outcome x_i with the probability p_i, where $\sum p_i = 1$. In the case of only two outcomes, of which one is zero, the notation can be simplified to the form (x, p) with $(x, p; 0, 1\text{-}p)$. The summary description of such a pair of prospects is usually presented to experimental participants in the form of a list.

---

[1] Blaise Pascal and Pierre Fermat are said to have founded modern probability theory over the discussion of the so-called *Problem of points* which was first mentioned in a 15th century textbook by the Italian mathematician Luca Pacioli.

For example,

A: Get $ x with probability *p*, $0 otherwise.

or

B: Get $ y for sure.

Another milestone in the progression of decision theory, the shift from expected value to expected utility (EU) theory, was also instigated by a game, the St. Petersburg lottery. Expected utility theory, originally proposed by Daniel Bernoulli (1738/1954), assumed that prospects like the above are evaluated by comparing the sums of the products of the subjective utility of the option's outcomes $u(x_i)$, which is a positive monotonic but decelerating function of the desirability of a monetary amount x_i, and the probability of obtaining these outcomes p_i,

$$EU = \sum p_i u(x_i) \tag{1.1}$$

and that people choose the prospect with the higher expected utility over the option with a lower expected utility. With its formal axiomatic derivation by von Neumann and Morgenstern (1947), EU was established as both a normative model providing a benchmark for how people should make decisions and a descriptive model of how people actually make choices between options.

The descriptive nature of the model was soon to be questioned though, as there was accumulating evidence showing that for some pairs of prospects the observed choice behaviour did actually violate EU axioms. The two most important violations were the Allais paradox (Allais, 1953) and the Ellsberg paradox (Ellsberg, 1961). I will not attempt to provide a comprehensive review of all the inconsistencies found at that time (for more details see Camerer, 1992; Kahneman

& Tversky, 1979; Starmer, 2000) but I will devote some space to the Allais paradox as it illustrates a set of choice phenomena that will also be relevant for later sections of this thesis. The first choice problem designed by Allais (1953) was of the following form:

A_1) \$1 million

or

B_1) \$5 million, .10;
\$1 million, .89;
\$0, .01.

In a second game the prospects to choose between were:

A_2) \$1 million, .11;
\$0, .89;

or

B_2) \$5 million, .10;
\$0, .90.

Each pair shares a common consequence (.89, \$1 million within the first pair and .89, \$0 within the second pair) and the second set can be derived from the first by subtracting the common consequence of .89, \$1 million. According to the independence axiom of EU theory the preferences within the two pairs should be independent of such common consequences and people should consistently choose either 'A_1' + 'A_2' or 'B_1' + 'B_2'. Instead, Allais (1953) discovered the predominant choices to be 'A' for the first and 'B' for the second pair which is called the *common consequences effect*.

In another example the following gambles were used:

C_1) \$3000

or

D_1) \$4000, .80;
\$0, .20.

By dividing both options by four we get a second set

C_2) \$3000, .25;
\$0, .75;

or

D_2) \$4000, .20;
\$0, .80.

Here, one usually finds the majority of subjects to prefer the sure option 'C_1' with the lower expected value within the first set and the more risky high expected value option 'D_2' in the second set. This is again a violation of the independence axiom of EU theory according to which people choosing 'C' or 'D' in the first gamble should choose the same option in the second problem. As the ratio of *p* and *x* for 'C' and 'D' is identical in both sets this phenomenon is referred to as the *common ratio effect*. Both of the paradoxes illustrate the *certainty effect* which describes the fact that changing the probability of an outcome by a constant factor has more impact when either the initial or the resulting probability involved certainty ($p = \{0,1\}$) than when it was and remained merely probable. This is commonly illustrated by Zeckhauser's paradox: When playing Russian roulette people would be willing to pay more for the removal of a single bullet from a gun with only one bullet in the six chambers than for the removal of one bullet from a gun with four bullets in the six chambers.

Another violation of EU, which will be of interest later is the *reflection effect* (Kahneman & Tversky, 1979) which can be observed when subjects are presented with the options C_1 and D_1 shown above and the same gambles consisting of losses of identical size instead of gains:

C_3) -\$3000

or

D_3) -\$4000, .80;
\$0, .20.

The inversion of the sign of the outcomes results in a preference reversal indicating a change in risk preference. People seem to be risk averse in the context of gains and risk seeking in the context of losses.

When we put these results together we can see a pattern emerges which has become known as the *four-fold pattern of risk attitudes* (Tversky & Kahneman, 1992), consisting of risk-averse behaviour for gains with medium to high probabilities or losses of small probability and risk-seeking for losses with medium to high probabilities or small probabilities in the domain of gains. Both the Allais paradox and the reflection effect will be discussed in more detail in Chapter 5.

In their seminal paper Kahneman and Tversky (1979) presented the powerful and formally appealing Prospect Theory (PT) that could explain the descriptive deficiencies mentioned above and which would dominate decision theory thereafter. One of the most substantial contributions of PT was the distinction between a *value function* and a *probability weighting function*. The value function $v(\cdot)$ is defined over changes in wealth (gains or losses) instead of absolute levels of wealth. In order to account for the changes in risk preference mentioned above the value function is assumed to be S-shaped with concavity within the domain of gains ($v''(x) < 0$ for $x > 0$) and convexity in the domain of losses ($v''(x) > 0$ for $x < 0$), reflecting diminishing sensitivity for both domains with increasing distance from the reference point. Furthermore the function is steeper for losses than for gains, resulting in losses looming larger than equivalent gains ($-v(-x) > v(x)$ for $x > 0$), see Figure 1.1 .

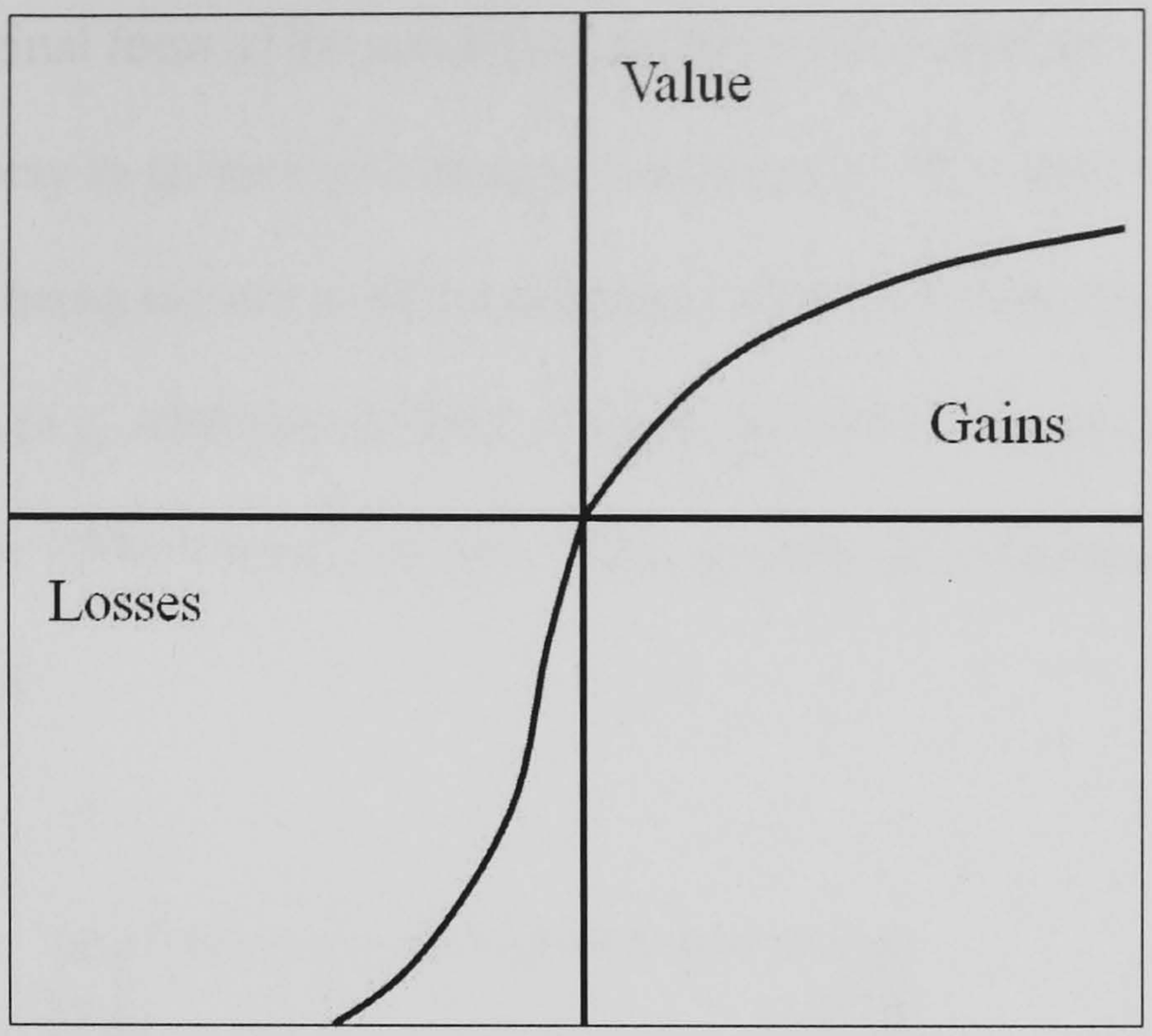


Figure 1.1. Example of a typical PT value function.

The more important and novel part of the model though, was the addition of the probability weighting function $w(\cdot)$ which captures the impact of p on the value of a prospect, see Figure 1.2. According to Kahneman and Tversky (1979) small probabilities are overweighted and large probabilities are underweighted. As a consequence $w(p) + w(1\text{-}p)$ do not necessarily add up to unity. In the context of the Allais paradox described above, this means that, if decision weights do not scale linearly, scaling down the probabilities by a common factor can change preferences and thus C_1 might be chosen over D_1 more than C_2 is chosen over D_2. In particular, if small probabilities are overweighted, the ratio between the weights of .2 and .25 may be higher than between .8 and 1, because the smallest probability (.2) gains an additional boost.

The original form of the weighting function was discontinuous[2] at both ends and later gave way to an inverse S-shaped version which has been consistently replicated after being subject to direct empirical tests involving descriptions of simple gambles (e.g. Abdellaoui, 2000; Bleichrodt, 2001; Camerer & Ho, 1994; Gonzalez & Wu, 1999; Tversky & Fox, 1995; Tversky & Kahneman, 1992; Wu & Gonzalez, 1996).

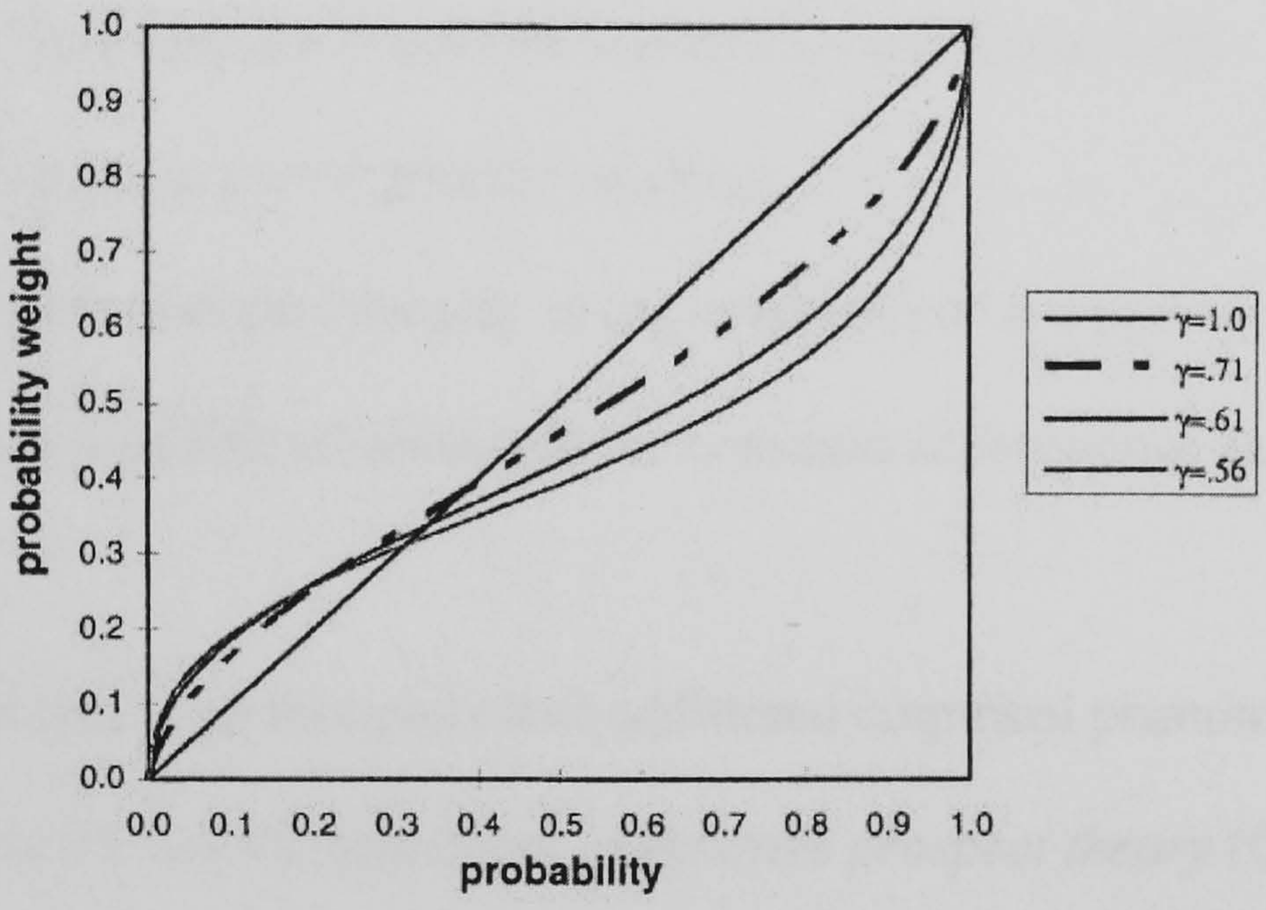


Figure 1.2. Tversky and Kahneman (1992) weighting function for various γ values (redrawn from Wu & Gonzalez, 1996)

The PT weighting function $w(\cdot)$, according to Tversky and Kahneman (1992), has the following functional form:

---

[2] The original prospect theory paper (Kahneman & Tversky, 1979, p. 283) explicitly mentions ambiguities regarding the behaviour of the weighting function at the extreme ends and predicts that "highly unlikely events are either ignored or overweighted, and the difference between high probability and certainty is either neglected or exaggerated".

$$w(p_j) = \begin{cases} \dfrac{p_j^{\gamma}}{\left(p_j^{\gamma} + \left(1-p_j\right)^{\gamma}\right)^{1/\gamma}}, \text{if } x_j \geq 0, \\ \dfrac{p_j^{\delta}}{\left(p_j^{\delta} + \left(1-p_j\right)^{\delta}\right)^{1/\delta}}, \text{if } x_j < 0. \end{cases} \tag{1.2}$$

There are two parameters (γ, δ) which determine the shape of the weighting function separately for gains (γ) and losses (δ). To determine the value of a gamble $(p, x; q, y)$, both functions are combined multiplicatively in the form

$$V(p, x\,;\, q, y) = w(p)v(x) + w(q)v(y). \tag{1.3}$$

This functional form and the interplay of the properties of $v(\cdot)$ and $w(\cdot)$ can explain the four-fold pattern of risk attitudes and the common consequence and common ratio effects.

It did not take long though before additional empirical phenomena were found that neither PT nor its refinement *cumulative prospect theory* (CPT) by Tversky and Kahneman (1992) could account for (see Camerer, 1992, 1995). A comprehensive review of all the latest paradoxes is provided by Birnbaum (2007). Alternative models have been proposed since then, including regret theory (Loomes & Sugden, 1982), rank-dependent utility theory (Quiggin, 1982, 1993), aspiration-level theory (Lopes, 1987), *decision by sampling* (Stewart, Chater, & Brown, 2006; Stewart & Simpson, in press) and the TAX model (Birnbaum, 2007), to name but a few. However, CPT still dominates the literature and is considered as the most influential theory (Bleichrodt, 2001; Starmer, 2000; Wu, Zhang, & Gonzalez, 2004).

We can see from this section why simple static lotteries have been so appealing. They provide clearly separable properties which can be controlled easily to identify specific properties of behaviour in risky choice or to test different model predictions. Whenever paradoxes are observed within these gambles they become important benchmarks for which subsequent models must account. This has led to a number of significant contributions that have enhanced our understanding of decision making under risk and it is therefore not surprising that the field has been dominated by this paradigm for more than half a century.

1.3 Choice under uncertainty

Although the simple gamble paradigm has been "as indispensable to research on risk as is the fruitfly to genetics" (Lopes, 1983, p. 137) it also has its limitations. Until now we have only been dealing with gambles that involved prospects for which all probabilities attached to the consequences have been known to the decision maker. These situations involve *risk* and risk should be distinguished from *uncertainty* (Knight, 1921). Uncertainty defines situations in which some of the probabilities are unknown (i.e., the decision maker must estimate or infer the probabilities, and may be able to do so very imprecisely). Most sports bets, for example a bet on a football team winning, are examples of uncertain outcomes. No objective probability exists for this event, only subjective assessments regarding its likelihood. In the context of real-life situations one can assume that most decisions are made under at least partial uncertainty (Busemeyer, 1985) which is not represented in textbook gambles (Wu & Gonzalez, 1999). The first theory put forward to deal with this alternative strand of decisions under uncertainty was Subjective Expected Utility (Savage, 1954), which introduced a subjective

probability measure in addition to a personal value function. Later, with CPT the basic PT was generalised to accommodate uncertain outcomes as well. The two-stage model (Fox & Tversky, 1998; Tversky & Fox, 1995) provided another extension of PT into the domain of uncertainty, assuming that the probability p of an uncertain event (E, x) is judged by the decision maker in a first stage, before the subjective probability $p(E)$ is then transformed using the probability weighting function $w(\cdot)$,

$$W(E)v(x) = w(p(E))v(x). \tag{1.4}$$

Empirical tests conducted by Wu and Gonzalez (1999) lend support to this model and demonstrate that the same inverse S-shaped pattern of risk preferences which Wu and Gonzalez (1996) found for risky choice can also be applied to uncertainty. As I will show in the following section, the theories of both domains, decisions under risk and uncertainty provide useful concepts that can be applied to a new category of decision problems, *decisions from experience*.

1.4 Descriptions vs. Experience

From the previous sections we have seen that a lot experimental evidence has been accumulated in experiments using simple descriptions of gambles, supporting the notion of an inverse S-shaped probability weighting function across the domains of both risk and uncertainty. Consequently, it came as a surprise when a series of recent studies reported choice behaviour that differs sharply from that observed in decision making tasks using classical lottery descriptions: crucially, choices appeared to indicate an *underweighting* of small probabilities rather than an overweighting (e.g., Barron & Erev, 2003; Hertwig, Barron, Weber, & Erev, 2004,

2006; March, 1996; Weber, Shafir, & Blais, 2004). What all these studies have in common is the use of a different type of choice tasks. As pointed out by Busemeyer (1985) decisions under uncertainty can either be based on past experience with similar situations or on information about the outcomes actively collected prior to making a decision. Whereas Tversky and Fox (1995) examine the former, the experience-based decision tasks discussed here investigate choices that are based on novel, actively sampled information. To understand the potential causes of the reversal from over- to underweighting it is important to examine these tasks more systematically.

1.4.1 Small feedback-based decisions

The early experiential paradigms reporting a pattern of underweighting of small probabilities are repeated choice problems consisting of several hundred trials of consequential decisions whose accumulated outcomes constitute the final score (e.g., Barron & Erev, 2003; Erev & Barron, 2005; March, 1996). This strand of work goes back to the animal learning tradition initiated by the work of Thorndike (1898) and the law of effect, after which options that have led to good outcomes in the past are more likely to be chosen again. Models incorporating this principle in different ways had later a renaissance in the probability learning literature, which I will also discuss in one of the following sections. Furthermore, these learning models have been shown to provide a useful framework to investigate repeated binary choices (Erev & Haruvy, in preparation).

One specific type of repeated decision making problems referred to as *small feedback-based decisions* (Barron & Erev, 2003), provided the first evidence for apparent underweighting of small probabilities in experiential choice. In small

feedback-based decisions participants have to choose between the same two options repeatedly over hundred of trials. Although they start off with no information about the outcomes of the two options they receive feedback in the form of payoffs from the chosen distribution after every single trial. The decisions in this context are referred to as 'small' because the expected consequence of each individual choice comprises very small payoffs of only a few cents. Significant amount are only accumulated over time. The learning curves obtained from this data also allow the examination of the changes in choice behaviour across trials with the accumulation of feedback.

The method is similar to the one used by Busemeyer (1985) with the difference that participants sample from binary events instead of normally distributed outcomes. However, this kind of task has to be distinguished from the repeated play of gambles (see Keren & Wagenaar, 1987; Lopes, 1981; Wedell & Bockenholt, 1990) where participants are typically asked to select which of two gambles they would prefer to play repeatedly (e.g. 10 times) after a one-shot decision, whereas small feedback-based decisions actually involve repeated play with a new decision in each trial.

In addition to differences in the maximising choice proportions compared to decisions from description, Barron and Erev (2003) also found mirrored common ratio and reflection effects, which are all reversals of phenomena usually reported in descriptive choice tasks. The authors see the explanation for the observed deviation as relating to the possibility that people rely excessively on recent outcomes. This only emerges in the context of feedback-based decisions where information has to be accumulated over time. With the participants' final payment being contingent on the accumulated total score this procedure poses potential

problems that exacerbate the interpretation of the results. Firstly, there is an overlap of two different, potentially opposing, strategies within each trial. The accumulation of information regarding the potential payoffs (exploration) coincides with the selection of the option that appears to offer the highest expected value in order to maximise the total score (exploitation). Or, as Lee (1971, p. 248) describes it, a participants "learns while he earns" (see also Berry & Fristedt, 1985; Erev & Barron, 2005; March, 1996). An alternative with a rare but high payoff, for example, comes with the cost of receiving inferior payoffs in most of the trials resulting in a restrained information search for this option (see also Denrell, 2007). This adds additional dynamics to the problem and makes the task structurally different from a descriptive one shot decision making problem. Secondly, it remains unclear whether the behaviour shown is due to the impact of experience itself or whether the preference reversals are driven by repeatedly playing the gambles.

1.4.2 Decisions from Experience

The necessary disentanglement of these confounding variables was achieved by Weber, Shafir and Blais (2004) and Hertwig et al. (2004, 2006) who introduced an improved experimental design. In an initial sampling phase participants could freely explore a pair of lotteries without cost (or reward) by drawing samples (with replacement) from the options' underlying outcome distributions (e.g., a participant might sample the sequence {4, 4, 0, 4, 0, 4} from a distribution with a .8 chance of winning 4 points and 0 points otherwise). Only after this sampling phase can participants decide which lottery to play once for real. Hertwig et al. (2004) refer to this paradigm as *decisions from experience* (hereafter called DfXP) and distinguish

it from d*ecisions from description* (DfD), the paradigm described above. By separating the sampling phase from the actual one-shot decision this design gets rid of the exploration-exploitation trade-off and distinguishes itself from *small feedback-based decisions*.

Hertwig et al. (2004) found reversed choice proportions in six different gambles (see Table 1.1) between DfD (showing apparent probability overweighting), and DfXP (showing apparent underweighting).

TABLE 1.1

Summary of the choice proportions reported by Hertwig et al. (2004)

Decision Problem	H	L	Percentage choosing H		
			Rare event	DfD	DfXP
1	4, .8	3, 1.0	0, .2	36	88
2	4, .2	3, .25	4, .2	64	44
3	-3, 1.0	-32, .1	-32, .1	64	28
4	-3, 1.0	-4, .8	0, .2	28	52
5	32, .1	3, 1.0	32, .1	48	20
6	32, .025	3, .25	32, .025	64	12

In choice problem 1, for example, only 36% of the participants in the DfD Condition chose the more risky option (H option) with the higher expected value, whereas the same option was chosen by 88% in the DfXP Condition. Similar results were reported by Weber, Shafir and Blais (2004). Another reversal towards more H choices under DfXP is observed for choice problem 4. However, it is important to note though, that there are also reversals in the other direction, for example for problems 2, 3, 5, and 6, with lower proportions of H choices under DfXP. According to Hertwig et al. (2006) the difference between the two formats can be attributed to a differences with regard to the psychological impact of the rare events. As the impact of rare events under DfD is usually described by

overweighting of small probabilities the reversed pattern would indicate apparent underweighting of small probabilities under DfXP. Such a dependence on the impact of the rare event is illustrated in the Table above. If a rare event makes an H option unattractive, like the 20% chance of winning nothing in choice problem 1, the underweighting of this rare event will make this H option more attractive and increase the proportion of participants choosing it. On the other hand, if the rare event is the part of an H option that makes it attractive then underweighting of its probability will have the opposite effect by penalising this option, for example in problems 5 and 6 in Table 1.1.

Although the participants were free to sample as often as they wanted Hertwig et al. (2004) report that the participants' information search prior to choice was based on a rather small number of samples. Crucially, in such a sampling task, when there is a low probability event and the sample is small, the number of times the event occurs in a given sample is positively skewed. Using the previous example (.8 chance of winning 4 points and 0 points otherwise), if we have 100 people who only draw 10 samples each from this distribution, on average 38 will experience 0 points fewer than twice, and will thus underestimate the true probability of receiving 0 points, including 11 who will not even experience the zero outcome once; 30 will experience 0 points exactly twice and will correctly estimate the probability; and 32 will experience the zero outcome more than twice and will overestimate the probability. This positive skew pattern occurs for all rare binomially distributed events and is also confirmed by Hertwig et al.'s data which shows that in some cases participants did not encounter the rare event at all and stayed completely ignorant of its existence. Although the asymmetry between underestimation and overestimation is small, Hertwig et al. (2004) argued that it is

one of the sources of the underweighting of low probability events. While the implementation of the DfXP paradigm helped excluding repeated choice as a potential cause it came at the cost of introducing sampling error as a new confound.

Hertwig et al. (2004) also tested whether an overemphasis of outcomes from more recent samples may explain the apparent underweighting of rare events in DfXP (as shown by Barron & Erev, 2003). They split the sequence of draws from each option into two halves and found that the expected values of samples from the second half of the sequences predicted participants' performances much better than the expected values of samples from the first half.

In summary, Hertwig et al. (2004) suggested that reliance on small samples of information (perception of a lack of variability) and overweighting of recently sampled information (recency effect) are the possible explanations for the underweighting of rare events under DfXP. However, there is also a potential interaction between the two as having less exposure to a rare event also reduces the probability of encountering it in one of the recent trials.

Generally, prospect theory (Kahneman & Tversky, 1979) does not seem to predict choice satisfactorily within decisions from experience as it assumes overweighting, not underweighting of small probabilities. Hertwig et al. (2006) consequently tried to develop a model that can account for the data by incorporating recency weighting in a Bush and Mosteller (1955) type stochastic learning model, which is an example of that old learning theory tradition, in combination with PT's value function and a recency parameter. This *value-updating model* provides choice proportions that approximate the observed proportions quite well and will be explored more fully in Chapter 6 where it will be compared with other potential modelling solutions.

Similar to the reliance on small samples explanation, Fox and Hadar (2006) argue that sampling error in DfXP reconciles an apparent overweighting of small probabilities in decision from description (as embodied in prospect theory, Kahneman & Tversky, 1979), with an apparent underweighting of small probabilities in decisions from experience. That is, underweighting occurs because, in small samples, rare events are simply not sampled very frequently. Hence, the phenomenon is not psychological, but results from the statistical properties of small samples. Hertwig et al. (2004) had not fully considered this as they evaluated the choice behaviour using the experimenter defined probabilities.

Furthermore, Fox and Hadar (2006) also raise the possibility of a distortion through judgement error, which might arise in the mapping from experienced frequencies to probabilities (i.e. at the first stage of the two-stage model of decisions under uncertainty (Fox & Tversky, 1998; Tversky & Fox, 1995), described above). However, they provide evidence against this possibility. By adding an explicit probability judgement task to the design, they show that probability judgements in their experiment were well calibrated. Moreover, they successfully applied prospect theory value- and weighting-function parameters reported by Tversky and Kahneman (1992), originally fitted to descriptive problems, to individual probability judgements, and found a good fit with the observed choices. Fox and Hadar (2006) therefore argue that apparent evidence for underweighting of probability in decisions from experience is, in reality, entirely consistent with overweighting of probabilities, as found in descriptive choice problems.

The existence of a sampling error, however, highlights the actual difference between the two paradigms. Whereas, a summary description in DfD is

standardised, providing the same information for every participant, in DfXP participants' experienced probabilities are likely to differ from the objective probability depending on the sequence of outcomes actually observed. The initial DfXP design is thus not an adequate method to investigate the question whether people exhibit a different choice pattern when faced with experiential choice problems that are structurally identical to gamble descriptions.

One of the goals of this thesis is to provide an answer to this question. The thesis will take up these points after the introduction has further explored the links between decision making and learning, feedback and experience as they have been investigated in other areas of psychology or related research fields like economics, animal learning and foraging behaviour.

1.5 Other related decision making tasks

Although feedback, experience and learning have not been investigated intensively in the decision making literature other domains have investigated these issues more closely. This section will provide a brief summary of the type of problems studied elsewhere and will highlight important findings that could provide insights for the decision problems discussed here.

1.5.1 The 1950's: Probability Learning

A task very similar to the type of experiential problems introduced above, especially with small feedback-based decisions, is the probability learning paradigm, which, as I already pointed out earlier, has its roots in the animal learning tradition that dates back to Thorndike (1898), investigating formalisations of principles inherent to human learning phenomena. Probability learning has been

studied extensively within experimental psychology in the context of mathematical learning theories (Estes, 1950) during the 50's and 60's right into the late 70's. Reviews of this literature can be found in Estes (1976), Fiorina (1971) and Myers (1976). More recent surveys are provided by Shanks, Tunney, and McCarthy (2002) as well as Vulkan (2000).

In a typical experiment involving probability learning subjects have to predict repeatedly which one of two binary outcomes is going to appear in a long series of trials. The event to predict could be the appearance of a light on the left or right side of a screen, or the colour of the next card drawn from a deck with blue and yellow cards. The probabilities attached to the different events are fixed and usually independent of previous responses and outcomes. After each prediction the participants receive feedback by revealing the actual outcome.

The striking observation in these probability learning tasks was that the asymptotic probabilities of predicting the different outcomes were equal to the outcomes' relative frequencies of occurrence. This indicates that subjects are able to learn the probabilities with which events appear quite accurately. Furthermore, it implies that people use a sub-optimal strategy and fail to maximise expected value. To illustrate this point, imagine a pair of stimuli of which one (A) appears 70% of the time the other one (B) 30% of the time. By probability matching and predicting 'A' in 70% of the cases and 'B' in 30% of the cases the rate of correct predictions will be only 58% ($.7^2 + .3^2$). Instead, the optimal strategy would be to consistently predict the event that has been identified as the more likely one, resulting in a correction prediction rate of 70% (1.0 *.7). Sub-optimal behaviour in the form of asymptotic probability matching has been replicated in numerous experiments with humans, rats, pigeons and monkeys. It has also been found in multiple-cue choice

tasks (e.g. Estes, Campbell, Hatsopoulos, & Hurwitz, 1989; Friedman & Massaro, 1998; Myers, 1976; Shanks, 1990). Birnbaum and Wakcher (2002) have been able replicate the classical results with data collected over the Web by implementing a button design that is comparable with the one used in experiential choice tasks.

However, there have also been results that complicate the picture further. A famous deviation from probability matching is the *gamblers fallacy* or *negative recency* (Anderson, 1960; Anderson & Whalen, 1960) which describes the tendency of consistently predicting the non-reinforced option based on the belief that a run without successes will be balanced out by consecutive future successes, thereby misinterpreting the law of large numbers and neglecting the independency of the events. Other departures from probability matching in the direction of more rational behaviour, so called overmatching or overshooting, have been observed under conditions with larger incentives, meaningful feedback and extensive training (e.g. Edwards, 1956, 1961; Shanks et al., 2002; Siegel, 1959; Siegel & Goldstein, 1959). Although all these different manipulations indicate that the effect is less robust than initially thought (Friedman & Massaro, 1998), they do not seem to eliminate it completely (Myers, 1976).

The most successful models predicting similar asymptotic learning curves have been stochastic learning models (e.g., Bush & Mosteller, 1955; Estes, 1950) which incorporate different updating mechanisms upon the received reinforcement. A review of the predictions and progression of Estes' model can be found in Bower (1994). Alternative models have been built on assumptions regarding the usage of runs and patterns that people might have identified within sampled outcome sequences (Edwards, 1956; Goodnow, 1955; Nicks, 1959; Restle, 1961). These approaches are much more difficult to formalise (Vulkan, 2000), as there is a vast

number of potential hypothesis that can be formulated. Yet another strand of explanation for the gap between matching and optimisation was the assignment of utilities to switching between options in late trials. This includes utility gained through the reduction of boredom or an increased utility for successful prediction of the less frequent outcome (Brackbill & Bravos, 1962; Siegel, 1959).

Probability matching tendencies have also been found in the context of the experiential decision problems. Erev and Barron (2005), for example, reported deviations from maximisation in small feedback-based decisions that can be approximated by probability matching. Underweighting of small probabilities in small feedback-based decisions on the other hand cannot be explained by probability matching (see Barron & Erev, 2003). However, DfXP describes a slightly different paradigm that has to be distinguished from probability learning tasks in a number of ways. One of the differences lies in the reinforcement. In probability matching and small feedback-based decisions people are facing the trade-off between exploration and exploitation mentioned earlier. Related to that, DfXP is a single shot decision task with a sampling or learning phase prior to choice whereas probability learning tasks are repeated choice tasks. Thus, in DfXP it is not possible to test for probability matching as it does not provide sampling trials long enough to reveal stable asymptotic performance. There is also a difference in terms of the degree of practice to which participants are exposed. In the case of probability learning this comprises hundreds of trials or even a thousand (Edwards, 1961). Subjects therefore base their choice on much more data than under DfXP, which is characterised by an extremely short information search of less than 20 samples per choice problem.

Although both tasks are quite distinct, the findings from the probability matching literature could still prove to be relevant for the identification of the properties causing the apparent underweighting of small events. The processes or choice strategies behind the two phenomena could overlap and the models introduced in this literature could provide alternative approaches to understand the DfXP phenomenon. Of special interest in this context are the Bush and Mosteller type stochastic learning models that have already been tested in the domain of DfXP in the form of the value-updating model (Hertwig et al., 2006). Furthermore, the probability learning literature has put forth some additional approaches to explore such trial by trial data which has not yet been utilized or tested in DfXP. This includes models that try to capture the impact of sequence effects, runs and patterns (Restle, 1961), which will be discussed in more detail later in Chapter 6.

1.5.2 The 1970's and 1980's: Experience in the Judgement literature

A similarly pessimistic view regarding the possibility of learning from experience and feedback has been discussed within the Judgement literature during the 70's and 80's. The research at that time was strongly influenced by the emergence of the heuristics and biases programme in decision-making which led to the discussion of conditions under which learning rules and heuristics can be acquired through experience (Einhorn, 1980). It was found that actual outcome feedback provided in real world environments is not adequate to learn heuristics and complex rules as it does not necessarily provide the relevant information (Brehmer, 1980). This might be due to the lack of the necessary schemata to make use of the information provided (Brehmer, 1980) or due to scarcity of feedback (Klayman, 1988). In an evaluation of a job selection procedure, for example, one might not have

information about incorrectly rejected candidates (false negatives) which would be important to judge the validity of the chosen strategy (Einhorn, 1980; Einhorn & Hogarth, 1978).

Instead it was shown that learning from experience in the form of feedback can actually reinforce normatively poor decision rules and result in a range of biases (Brehmer, 1980; Einhorn & Hogarth, 1978; Hammond, Summers, & Deane, 1973) in both deterministic and probabilistic tasks, without the agent being aware of it (Einhorn, 1980). The biases discussed included the tendency to focus on confirmatory evidence (Wason, 1960) and the inability to incorporate negative or disconfirming information (Einhorn & Hogarth, 1978; Klayman, 1988).

The predominant tasks used to investigate outcome feedback were single- and multiple-cue probability learning paradigms (Hogarth, McKenzie, Gibbs, & Marquis, 1991; Klayman, 1988). In these tasks complex probabilistic linear or nonlinear relationships between cue variables and a criterion have to be learnt through repeated outcome feedback. People show difficulties in learning these functions, especially when they are nonlinear and negative (a more detailed review is provided by Klayman, 1988). By looking at the acquisition of appropriate strategies in rather complex tasks like single- and multiple-cue probability learning experiments it is not surprising that performance was behind the predictions of normative approaches, such as Bayesian learning. The outcome-irrelevant learning structures referred to in this literature seem to be more complex real-life problems and do not necessarily overlap with more controlled laboratory tasks like DfXP, which provides immediate unambiguous feedback embedded in a repeated and identical context. It is therefore difficult to apply these results in the context of DfXP. As I will show in the following sections, there are alternative approaches

from other disciplines that are better suited to account for the type of experiential choice problems investigated here.

1.5.3 Repeated gambles in the economics literature

I have so far focussed on the psychological literature, but there is closely related work conducted in a different tradition: experimental economics. Due to their interest in learning processes experimental economists have a long tradition of investigating behaviour in decision making tasks with repeated trials rather than one-shot decision making tasks. First of all repetition helps the participants to familiarise themselves with the environment and the task (Binmore, 1999). In the context of complex games with multiple players it can also facilitate the understanding of one's own strategic options, the strategies used by the opponent, and the interaction of the two (Hertwig & Ortmann, 2001). Another motivation for economists to use repeated trials is their focus on equilibrium solutions. In the prisoner's dilemma or the ultimatum game, for example, people's one-shot decision behaviour does not necessarily fit with equilibrium predictions. However, when assuming an equilibrium to be established in the long-run, as asymptotic behaviour (Camerer, 1997; Fudenberg & Levine, 1998) and through an interactive process of trial-and-error learning (Binmore, 1999), people's behaviour changes, conforming to the Nash equilibrium. Some economists therefore share the view that within simply framed problems, providing adequate incentives and allowing for sufficient learning participants will get to equilibriums (Binmore, 1999) and expected utility maximisation will work as an appropriate descriptive approximation of individuals' true preferences (Plott, 1996). Chu and Chu (1990), for example, showed that the occurrence of preference reversals decreased when

participants repeated the experiment many times. Even the mere prospect of playing a gamble repeatedly seems to be sufficient to increase expected value maximisation in individual choice behaviour (Keren & Wagenaar, 1987). However, repetition does not always increase the tendency to use maximising strategies, as shown in probability learning tasks.

A maximisation problem in economics that is similar to the experiential choice problems mentioned above are stationary replications of one-shot decisions called *multi-armed bandit problems.* Bandit problems, first mentioned by Robbins (1952) got their name from the analogy with a slot machine. In an *n*-armed bandit problem one can choose from *n* different options (similar to an arm or lever of a slot machine) which provide a reward drawn from its underlying distribution. The player iteratively chooses one of the arms each round and observes the associated reward with the objective to maximise the sum of rewards collected. Small feedback-based decision problems can be formalised as *two-armed bandit problems* with the two arms A and B, each associated with a probability of receiving a reward (p(A) and p(B)) where $p(A) \neq 1 - p(B)$. As in the probability learning paradigm an optimal strategy would be to consistently choose the arm that maximises expected utility, after both p(A) and p(B) have been established in a series if trials. Instead of such rational behaviour probability matching is observed quite frequently which led to the application of stochastic learning models (e.g. Bush & Mosteller, 1955; Erev & Roth, 1999). This illustrates once more the obvious overlap between the different literatures on probability learning, bandit problems, reinforcement learning, and small feedback based decisions (Barron & Erev, 2003; Roth & Erev, 1995; Vulkan, 2000).

1.5.4 Risk sensitivity in the literature on animal decision making

Another research area that has been looking into learning processes involved in decision making is the field of foraging behaviour and animal decision making. As in human decision making, risk-sensitive foraging behaviour in animals is compared with a benchmark in the form of a rational model, for example, maximisation of expected utility. The literature seems to document a great number of similarities between risky decision making in animals and humans including the fact that animals show analogous deviations from expected utility maximisation, for example intransitivity of preferences (Shafir, 1994) and violations of the independence axiom in standard Allais-type common ratio manipulations (Battalio, Kagel, & MacDonald, 1985; MacDonald, Kagel, & Battalio, 1991; Real, 1996).

Yet there has also been experimental evidence for non-linear transformations in subjective probabilities that differ from the common finding in human risky choice. Instead of distortions of subjective probabilities in the direction of underweighting of high probabilities and overweighting of small probabilities, Real (1991) found over-representation of common events and under-representation of rare events similar to the finding in DfXP. The link between the two phenomena could be the learning mechanism shared by the experimental paradigms. With animals it is obviously not possible to directly communicate information regarding the risk attached to the available outcomes in the form of symbolic representations (Real, 1996). Instead the variability of the outcomes (e.g. magnitudes of food) can only be experienced by allowing the animal to explore the options in learning trials preceding choice.

One of the models that has been proposed to explain risk-sensitivity data in animal choice is scalar utility theory (SUT) by Marsh and Kacelnik (2002) which

postulates a cognitive representation of outcomes based on an internal scale which is governed by Weber's Law. An alternative explanation using associative learning mechanisms has been found to predict risk-sensitivity under situations where risk is operationalised in the form of delay of reward (Kacelnik & Bateson, 1996), instead of variability of reward size. According to Weber, Shafir and Blais (2004), both approaches are using measures of risk sensitivity that are proportional to the coefficient of variation (CV), the ratio of the standard deviation of outcomes and their expected value. Unlike the variance or SD, the CV is dimensionless and allows comparisons across domains. More importantly, the CV proved to be a better predictor of risk sensitivity than variance or SD in both human and non-human choice data when information about the risky option is obtained through repeated sampling and personal experience (Weber et al., 2004). They also show that associative learning models similar to the fractional adjustment model (Bush & Mosteller, 1955) correlates strongly with the CV, which could explain the successful application of Hertwig et al.'s (2006) value-updating model in the context of DfXP.

We can see from these sections on research from other fields that there are similarities with the problems under investigation here. Most of them resemble small feedback-based decision type tasks. The approaches that have been put forward to model the choices in experiential contexts on the other hand are quite different. The only overlap can be seen in the application of reinforcement learning models which seem to provide an appropriate description of the updating process on the basis of newly accumulated experience in experiential problems that incorporate an exploration-exploitation trade-off. How useful these insights are going to be in the context of the DfXP-type problems investigated here remains

unclear though, as the properties of these choice paradigms do not fully overlap. I will come back to this question when I discuss specific applications later. Although not explicitly mentioned a pre-condition for all the different tasks mentioned here is the ability to keep track of the actual outcome frequencies observed. Research investigating whether people are actually able to utilise this type of information will be presented in the next section.

1.6 Proportion and Probability Judgements

Although working with gambles presented in descriptive rather than experiential terms, Kahneman and Tversky (1979) already emphasised the distinction between probability weighting, which reflects the impact of an event, and the probability estimation, which is the perceived likelihood of an event, and noted that both processes may shape the perceived impact of rare events in real-life situations independently. This distinction is indeed very important in the context of experience-based decisions where the risk attached to an outcome has to be inferred from the frequency of occurrences, bearing the danger of an increasing incidence of deviations between objective and estimated probabilities. An alternative explanation for the apparent underweighting in DfXP could be a systematic underestimation of small probabilities. Overestimation, on the other hand, would have the opposite effect, increasing the impact of rare events and resulting in overweighting (Kahneman & Tversky, 1979). But how does underestimation fit in with what we already know about the processing of frequency information? This section will review research that has been investigating people's ability to judge frequencies and probabilities and explore how this could be linked with the observed choice behaviour in DfXP.

In general, the literature acknowledges that people show an innate sensitivity towards relative frequency information within a range of different tasks (Zacks & Hasher, 2002) and that people show great ability in synthesizing, storing, and accurately retrieving occurrences of event attributes (Howell, 1973). Zacks and Hasher (2002) describe frequency of occurrence as "… a fundamental aspect of the information that people code about their experiences in the world …"(p. 21). Provided estimates do mirror observed relative frequencies of the presented items quite well and their relationship can be described by an identity function (Peterson & Beach, 1967). Accuracy varies though and shows typically a small distortion with low frequencies being overestimated and high frequencies being underestimated (regression towards the mean) which Lichtenstein, Slovic, Fischhoff, Layman and Combs (1978) referred to as the *primary bias*. This is reported for both proportions (e.g., Erlick, 1964; Stevens & Galanter, 1957) and frequency judgements (e.g., Hertwig, Pachur, & Kurzenhauser, 2005; Peterson & Beach, 1967; Zacks & Hasher, 2002) of verbal and non-verbal stimuli and across different domains (language, statistical reasoning, and consumer decision making). One variable that has been found to be related to accuracy is sample size, with judgements of relative frequencies and proportions becoming more accurate with increasing sample size (e.g., Erlick, 1964; Sedlmeier, 1999; Shanks, 1995). Hintzman (1976) showed that frequency judgements were also independent of temporal recency of presentation and duration of presentation. Furthermore, this ability to encode frequency information has also proven to be remarkably stable across different age groups (e.g., Hasher & Chromiak, 1977; Hasher & Zacks, 1979), underlining the early acquisition and the reliability of this skill which has led to the hypothesis that the ability to encode frequencies is innate and automatic

(Hasher & Zacks, 1979, 1984). Sedlmeier (1999) reports evidence for the equivalence of probability and frequency judgements, especially when the input consists of serially encountered frequencies of events (Dougherty & Franco-Watkins, 2002). The apparent ability to accurately assess frequencies in our environment is also one of the arguments for the so-called *frequentist hypothesis* (Cosmides & Tooby, 1996) which is based on the idea that frequencies provide a basic input for a wide range of reasoning processes and that the presentation of uncertainty in a frequency format instead of probabilities can eliminate biases that have been found in judgement under uncertainty (e.g. the base rate neglect in Bayesian reasoning, Christensen-Szalanski & Beach, 1982; Gigerenzer & Hoffrage, 1995).

There are only a few studies that provide counter examples for the high accuracy of frequency judgements. Lichtenstein et al. (1978), for example, report rather poor accuracy for direct estimates of various causes of death. This difference to previous work might be due to the fact that subjects had to estimate the frequencies of events (e.g. rare diseases) they were not very familiar with (Shanteau, 1978). Other exceptions are reported in more naturalistic studies involving respondents' reports on frequencies of their own behaviour and episodic recall. Suboptimal performance in these tasks seems to be due to the usage of different heuristics (Schwarz & Wänke, 2002), like the availability heuristic (Tversky & Kahneman, 1973). Very rarely do studies report a pattern of underestimation of low proportions and overestimation of high proportions (e.g., Nash, 1964; Pitz, 1966; Shuford, 1961; Simpson & Voss, 1961) that would explain choice behaviour in the direction of underweighting of small probability events as found in DfXP.

Thus, in the light of all this evidence judgement error in the direction of underestimation of low frequencies seems a rather unlikely phenomenon and would contradict the experimental findings already established. Instead, one should expect accurate judgements with slight deviations from the true values in a typical inverse-S shaped pattern, indicating overestimation of low-frequency outcomes and underestimation of high-frequency outcomes. This is also what is observed in the context of tasks that are quite similar to DfXP. As we have seen in the section on probability learning tasks, the accurate assessment of frequencies is one of the requirements for the ability to *probability match* (Estes, 1976). Experiential choice tasks provide similar environments with sequentially experienced outcome frequencies and one should therefore expect similar effects. A first test of the judgement error hypothesis was conducted by Fox and Hadar (2006), who added a probability judgement task to their DfXP design. Participants' estimates were well adjusted though and in line with the results from the literature summarised here. I will come back to this question in Chapter 3 where I will provide a series of experiments testing the effect of judgement error in DfXP. Several of these experiments were implemented in a Web-based format, a methodology that I introduce briefly in the next section.

1.7 Web-based experimenting

An extensive part of my research that I am going to present in the experimental chapters of my thesis has been implemented in the form of Web-based experiments. I therefore want to provide a brief discussion of Web-based experimentation, its advantages, its potential problems, and research on its validity before I outline why I have chosen to make use of this technology.

With the development of the Internet to a mainstream communication platform that is reaching into all areas of life there has also been an increased interest of psychologists in conducting research on and through the Internet (e.g. Birnbaum, 2004; Schmidt, 2001; Skitka & Sargis, 2006). The technological advancements in CPU performance, modern Web browsers and the increased speed of the available network connections, that now allow the streaming of whole TV programs, have made it possible to carry out more complex and graphics-intense Web-based experiments that go beyond the capabilities of the usual HTML surveys. Today Web-based research has also found its way into the most important publications in the field (Skitka & Sargis, 2006).

My main motivations for using Web experimenting were the general advantages including low cost, uncomplicated participant recruitment, increased efficiency (Birnbaum, 2001; Fraley, 2004) and access to a wider and more heterogeneous sample which goes beyond the typical college students population (Skitka & Sargis, 2005). There is also evidence suggesting that participants respond more naturally when they are in familiar contexts like their homes instead of a laboratory environment (Skitka & Sargis, 2006). Nevertheless, there are also potential problems including the precision with which specific stimuli can be delivered (Krantz, 2001), uncontrolled contextual variations that can introduce additional error variance (Skitka & Sargis, 2006), differences between Web users and nonusers that may limit the generalisability of the findings (Birnbaum, 2004), non-response error due to low response rates (Skitka & Sargis, 2005) and repeated participation or fraud (Schmidt, 1997). A more detailed discussion of these advantages and potential limitations of Web-based research are provided in recent

reviews by Birnbaum (2004), Reips (2000; 2001) and Skitka and Sargis (2005; 2006).

There is substantial evidence regarding the equivalence of results obtained through Web and lab research for a wide range of designs and psychometric properties (Birnbaum, 1999; Buchanan, 2000; Buchanan & Smith, 1999; Krantz & Dalal, 2000). More importantly, a series of successful replications using Web-based research have been provided in the field of decision making using tasks with properties similar to the ones in DfXP tasks. This includes studies on violations of stochastic dominance and event-splitting effects in decision making under risk (Birnbaum, 1999, 2000; Birnbaum & Martin, 2003), probability learning experiments (Birnbaum & Wakcher, 2002) and medical decision making (Waters, Weinstein, Colditz, & Emmons, 2006). Furthermore, Birnbaum (1999) found that the data quality, tested in the form of direct and indirect monotonicity, was actually better for the Web sample.

There are many different ways of implementing Web experiments using server-side and client-side processing methods (for a review of these methods see Birnbaum, 2000; 2004; Fraley, 2004; Schmidt, 2001). A method that has only recently been utilised by psychologists in order to implement Web-based experiments, and which I have used to implement my experimental work, is Adobe Flash (e.g. Reimers & Maylor, 2005; Reimers & Stewart, 2007). Flash movies are displayed in the Web browser using a Flash plug-in which comes preinstalled with most of the available browsers on all major operating systems including Windows, Linux, MacOS as well as for handheld devices and the latest generation mobile phones. According to a survey conducted by Millward Brown, published by Adobe Systems Incorporated (n.d.), the flash player is "the world's most pervasive

software platform … reaching 99% of Internet-enabled desktops in mature markets as well as a wide range of devices". Flash facilitates the development of ergonomic content combined with a timely presentation format and allows greater methodological creativity.

With regard to experiments involving experience-based decision making tasks most of the reported experiments have already been implemented in a computerised format (e.g. Barron & Erev, 2003; Hertwig et al., 2004) which makes the transition to a Web-based experiment an easy step. It also has to be emphasised that the implemented design has been successfully used in the laboratory environment prior to the data collection on the web (see Chapter 2 and 3) and that the necessary steps have been conducted following the available guidelines (e.g., Birnbaum, 2001; Fraley, 2004; Schmidt, 1997). Given the reasons outlined here, I therefore believe that this is an appropriate technique to collect data on the experimental problem under investigation.

1.8 Motivation for thesis

The goal of this first chapter was to introduce the research on experiential choice tasks by providing the necessary theoretical background and the related concepts that have been identified in different research domains. It also reviewed the available studies on DfXP and their findings, indicating underweighting of small probabilities instead of overweighting of small probabilities in DfD. Furthermore, I have laid out the various hypotheses that have been put forward by a number of authors and the experimental work that has been conducted to test them. Notably, with the work of Fox and Hadar (2006) the focus has shifted from recency

weighting to sampling error. At the same time the question under examination seems also to have changed. Whereas the original studies conducted on experiential decision tasks were examining the effect of the format itself, the discussion now centres upon the inappropriateness of a specific design which obviously does not capture an experiential context that is structurally identical to a gamble description. Thus, the title of Fox and Hadar's (2006) paper, "Decisions from experience = sampling error + prospect theory", is therefore premature because it suggests that the underlying processes are the same for both descriptions and experience formats. However, as I have shown, there remain several confounds which make such interpretations difficult and these issues have not yet been explored appropriately or thoroughly.

As a consequence, my own research seeks to make a directional contribution to the field and its continuing development by providing a series of experiments that investigate the equivalence of experience-based decision making and decision from description under more appropriate experimental conditions. Firstly, the thesis offers a comprehensive review of the different strands of research that have been conducted and will indicate where the current discussion can be profit from contributions in related areas of research. By specifically testing the sampling error hypothesis as an explanation for the established choice phenomena I will show that none of the variables that has been put forward so far – neither sampling error, recency weighting nor judgement error – is actually sufficient to explain the effect. This will help bring to a close the discussion on sampling error whilst simultaneously reopening the debate about the actual overlap of cognitive processes involved in DfXP and DfD. Further, I will provide evidence suggesting

that established choice models fail to account for the DfXP phenomena indicating that they are not able to capture the relevant properties involved in DfXP. Thus, an additional contribution will be an effort to identify the cognitive processes constituting the differences between description-based decisions and decisions from experience by investigating the impact of different properties of the actually experienced sequences. The main focus of the additional experimental work will therefore be an examination of sampling order effects including sub-sequences resulting from switching between options. Other sequential properties, like runs of outcomes, will be explored theoretically by testing a variety of models on the collected data, drawing on modelling work that has been developed in the context of the probability learning paradigm.

The thesis is organised into a further six chapters. Chapters 2 and 3 will present initial experimental work conducted on the effect of sampling error. First, Chapter 2 will examine the contribution of sample size which has been identified as one of the sources for sampling error in the form of underrepresentation of rare events. Chapter 3 will go one step further by introducing a novel experimental approach which eliminates sampling error and thereby remedies the mentioned confounds in the original methodology. Additional questions arising from this work, including the effect of sampling order, will be the focus of Chapter 4. Thereafter, Chapter 5 will review the collected data in the context of established biases. Chapter 6 then tests how far different choice models that have been developed to account for decision making under risk from both description-based and experience-based choice tasks are actually able to explain the various strands of experimental evidence the thesis presents. The set of models considered will

include adaptations of models already established and one of my own modelling approaches which draws upon knowledge from other relevant areas of research within and outside the field of decision making in psychology but which, as discussed above, are pertinent to the problem constituted by DfXP. The thesis concludes in Chapter 7 with a summary of the findings, a discussion of the conclusions that can be drawn from the research presented and concrete suggestions for the different avenues of exploration that are crucial for the development of future research in this field.

CHAPTER 2

DECISIONS FROM EXPERIENCE UNDER COMPREHENSIVE-SAMPLING (EXPERIMENT 1)

2.1 Introduction

In Chapter 1, I noted that there remain a number of methodological shortcomings in the original DfXP design. The most significant of these is the difference between the objective and the experienced probabilities due to sampling error. We have seen that, in small samples, rare events can be underrepresented. In light of these differences in terms of the statistical properties of experienced samples, the research to date has overlooked the reality that DfXP tasks are not structurally similar to their descriptive counterparts. One way of resolving this issue is to increase the number of samples that participants have to draw. According to Fox and Hadar (2006), the apparent underweighting of probabilities should disappear when sampling error is reduced or eliminated. This chapter will provide an initial experiment testing this hypothesis. The motivation behind the design of the experiment presented here is two-fold. Firstly, it provides a replication of the original study testing the stability of the findings. Secondly, it investigates the impact of sample size on the apparent underweighting of small probabilities in DfXP.

Before I describe the experiment, I wish to briefly establish two methodological concepts that are relevant to the sampling process. In the original DfXP design it was left to the participants to decide how many samples to draw from the two distributions. I will refer to this type of design as *Free Sampling*. One

way of reducing the impact of undersampling is the introduction of a higher, fixed number of samples that participants have to draw before they can choose between the two options. Such a *Comprehensive Sampling* design comes with the advantage of standardising the information search for all the participants across all choice problems, another methodological problem not accounted for in previous sampling tasks. The differences in choice behaviour between the Free-Sampling Condition and the Comprehensive-Sampling Condition are examined in the following experiment.

2.2 Method

2.2.1 Participants

51 participants were recruited from students and members of staff of the University of Warwick through advertisements and a subject panel. All participants received £2 for their participation.

2.2.2 Stimuli

The six decision problems presented were taken from the same set of gambles used in the original Hertwig et al. (2004) experiment. The options have different expected values. Four of the choice problems provide gains; the remaining two consist of losses (see Table 2.1).

TABLE 2.1

Summary of the decision problems used in the experiment

Decision Problem	Options		Expected value		Rare event
	H	L	H	L	
1	4, .8; 0, .2	3, 1.0	3.2	3	0, .2
2	4, .2; 0, .8	3, .25; 0, .8	0.8	0.75	4, .2
3	-3, 1.0	-32, .1; 0, .9	-3	-3.2	-32, .1
4	-3, 1.0	-4, .8; 0, .2	-3	-3.2	0, .2
5	32, .1; 0, .9	3, 1.0	3.2	3	32, .1
6	32, .025; 0, .975	3, .25; 0, .75	0.8	0.75	32, .025

2.2.3 Design and procedure

The experiment was set up as a between-subjects design with two conditions, a Free-Sampling Condition, as in the original DfXP design, and a Comprehensive-Sampling Condition, with a fixed number of samples per option. Assignment to the two conditions was random, as was the presentation order of the six choice problems. 26 participants were assigned to the Free Sampling Condition. The remaining 25 ran the Comprehensive Sampling Condition. Both conditions provided a sampling phase in which participants explored the two options represented by two buttons, 'A' and 'B', on a computer screen, followed by a final decision phase where they had to choose the option they preferred.

Each click on a button in the exploration phase revealed an outcome from the option's underlying payoff distribution. The outcomes in both conditions were determined randomly (with replacement) for each participant and displayed for one second on top of the button pressed. While the outcome was displayed both buttons were ghosted out and inactive to prevent participants from clicking through the samples too quickly. The prerequisites to go to the following decision phase differed between the two conditions. The Free-Sampling Condition followed

Hertwig et al.'s (2004) paradigm: participants could stop the exploration of the buttons as soon as they felt confident enough to make a decision for real. However, in the Comprehensive-Sampling Condition participants had to sample 39^3 outcomes from each option, in any order, before proceeding to the decision phase. After the 39^{th} sample from one option its button turned shaded and further sampling from it was stopped. Figure 2.1 shows a screenshot of both exploration and the decision phase.

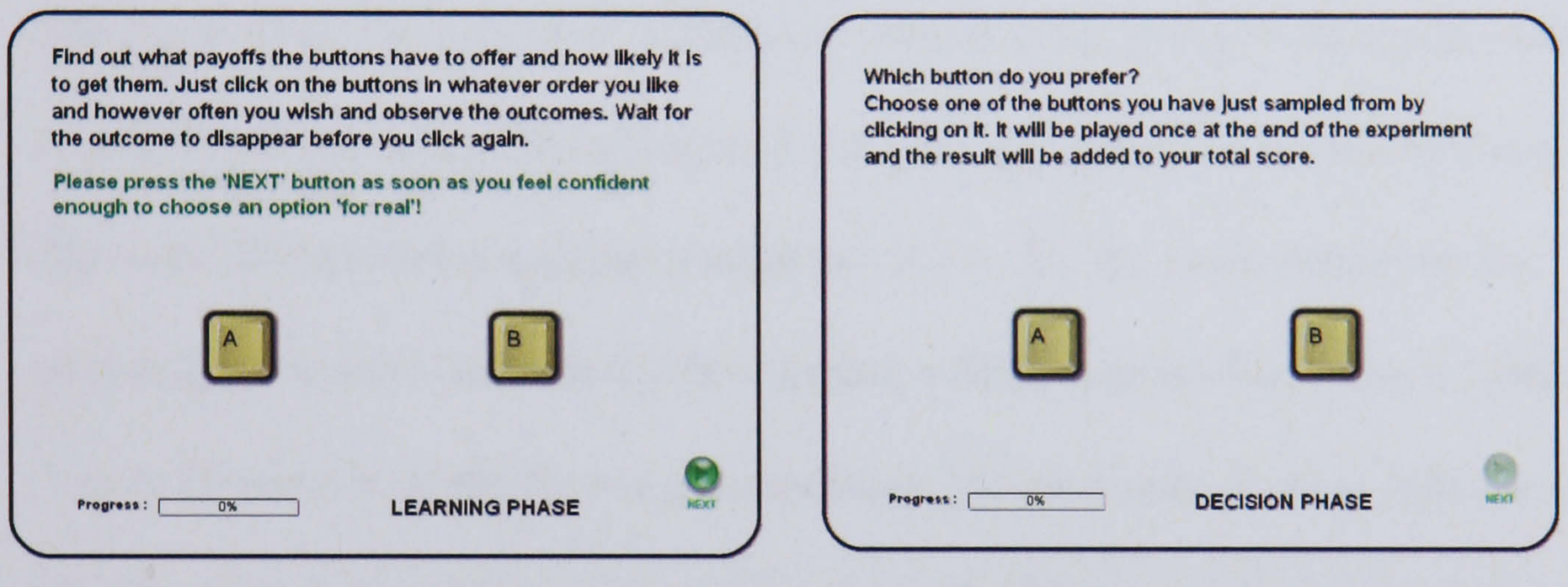


Figure 2.1. Screenshots from the sampling/learning phase (left) and the decision phase (right).

In the decision phase participants had to choose between the two options they had previously sampled from. Whereas the outcomes from the learning phase did not affect their score, participants were informed that the option chosen in the final decision phase would be played once at the end of the experiment and the result would be set against their total. Participants were instructed to maximise the number of points they accumulate within the six choice problems. At the end of the experiment, all six lotteries were played randomly for each participant before they were informed of their points total.

---

[3] The reason for the 39 samples was a programming error, the original number was 40.

2.3 Results

2.3.1 Free Sampling

2.3.1.1 Information search

One of the important observations in the original DfXP design was the limited information search resulting in sampling error. Within the Free-Sampling Condition a similar pattern could be observed. The number of draws was rather low with a median of 16.5 ($M = 21.12$) draws per choice problem. This matches the data reported by Hertwig et al. (2004) and Weber et al. (2004) who report values of 15 and 17, respectively. Sample sizes were not very stable across the different choice problems with a median correlation of $r = .58$. An examination of the sampling symmetry between the two options within choice problems revealed that in only 28 percent of all the cases participants had obtained an equal number of samples from both buttons. The median absolute difference between sample sizes for buttons A and B across all choice problems was 1 ($M = 3.25$). It can be seen in Figure 2.2 that participants sampled similar numbers of outcomes from options with the high and lower expected values ($t(25) = 0.76$, $p = .452$).

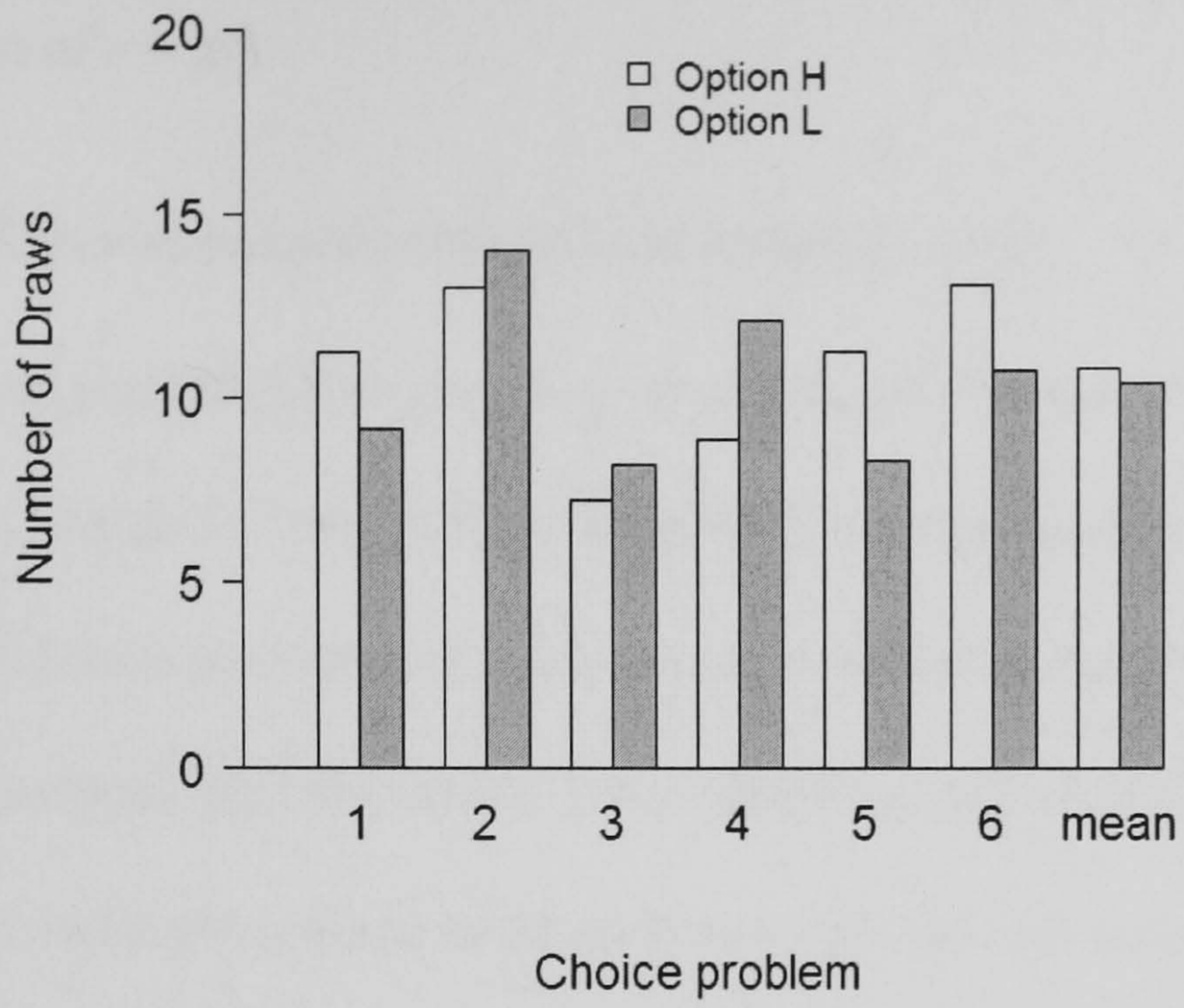


Figure 2.2. Number of draws from the high and low expected value option across the six choice problems

An examination of the sampling sequences showed that a number of different strategies were used to explore the two buttons. In 28% of all sequences the sampling process for one button was finished before the second button was explored, resulting in only one alternation between the two buttons. The opposite approach, continuous alternation between the two buttons was observed in 19% of the sequences. The median number of alternations was 3 ($M = 6$). To test the stability of the transitions between buttons the ratio of the actual number of switches and the potential maximum number of transitions (n-1) given the actual sample size n was calculated. Values close to one indicate a high rate of alternations whereas values approaching zero are a sign for a more separated exploration of the two options. A median switch ratio of .26 ($M = .44$) shows that there was a considerable amount of switching between options. Across all the

decision problems the switching ratios were relatively stable with a median correlation of $r = .68$.

2.3.1.2 Experienced probabilities and sampling error

As in the original DfXP experiment, the short information search in the Free-Sampling Condition came with substantial underrepresentation of the rare events. Across all choice problems the rare events were encountered less frequently than expected in more than two thirds of the sampling sequences (69%). More significant is the percentage of participants who did not encounter the rare event at all, thus remaining ignorant of its existence. This applied to 49% of all sequences, which is in good agreement with the 44% mentioned in Hertwig et al. (2006). Only 5% experienced the rare event as often as expected from its objective probability, leaving only 26% for cases of overrepresentation. Figure 2.3 provides further evidence for this interpretation, illustrating the positive skew of the distribution of differences between experienced and objective probabilities.

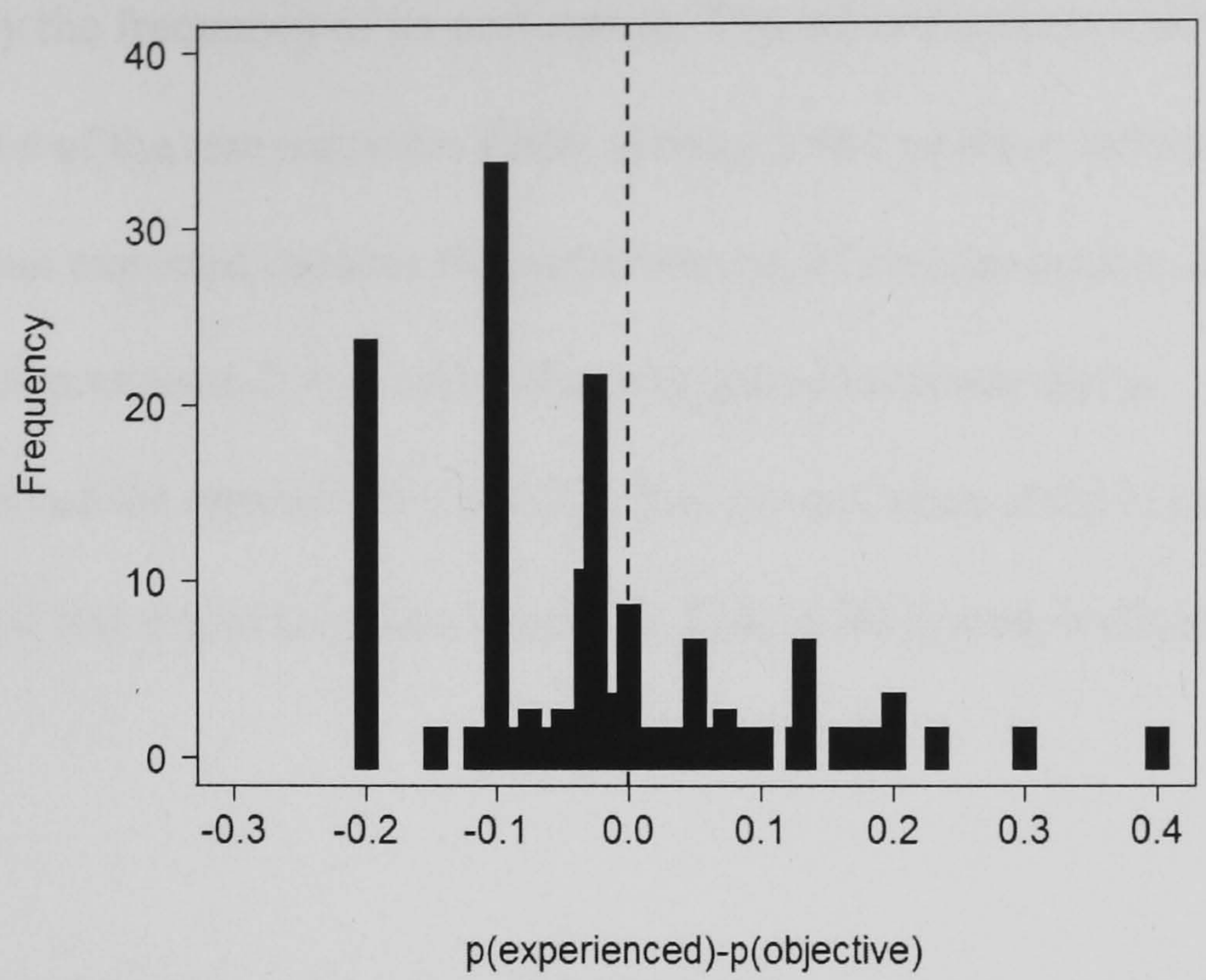


Figure 2.3. Histogram with the distribution of differences between experienced and objective probabilities. The bars to the left of 0 mark underrepresentation of the objective probabilities and the bars to the right mark overrepresentation of the objective probabilities.

In summary, all of the properties of the sampling behaviour and information search found in the Free-Sampling Condition very closely resemble the patterns found in the original DfXP experiment. This includes sample size, sample symmetry and deviations between experienced and objective probabilities. Again, there is a tendency to draw small samples which comes with a systematic underrepresentation of small probabilities.

2.3.1.3 Choice behaviour

To clarify whether this underrepresentation of small probability outcomes has an impact on choice, I calculated the number of times the option with the rare event was chosen depending on the number of times the rare event was actually observed. It is evident from the differences between the last two columns in Table 2.2 that the decision to choose the option with the rare outcome was partly

influenced by the frequency of its occurrence. This relationship is mediated though by the valence of the rare outcome. Encountering a rare positive outcome less frequently than expected reduces the attractiveness of such an option, as can be seen in choice problems 2, 4, 5 and 6. For a negative outcome that is underrepresented the opposite is observed; it is chosen more often in cases where it is experienced less frequently than expected. This is illustrated in choice problems 1 and 3.

TABLE 2.2

Choice behaviour depending on encounters with the rare event within the Free-Sampling Condition. Due to the small number of cases the category of encountering the rare event according to the objective probability was combined with the category of overrepresentation (last column).

	Options				Percentage choosing option with rare event		
Decision Problem	H	L	Rare event	Impact of the rare event	Rare event not encountered	Encountered less frequently than expected	Encountered as frequently as or more frequently than expected
1	4, 0.8	3, 1.0	0, .2	-	88 (7/8)	87 (13/15)	45 (6/11)
2	4, 0.2	3, 0.25	4, .2	+	14 (1/7)	42 (8/19)	86 (6/7)
3	-3, 1.0	-32, 0.1	-32, .1	-	89 (16/18)	81 (17/21)	60 (3/5)
4	-3, 1.0	-4, 0.8	0, .2	+	13 (1/8)	21 (3/14)	67 (8/12)
5	32, 0.1	3, 1.0	32, .1	+	0 (0/14)	0 (0/17)	44 (4/9)
6	32, 0.025	3, 0.25	32, .025	+	19 (4/21)	19 (4/21)	40 (2/5)

With regard to the proportions of participants choosing the option with the high expected value (H option), they were found to be nearly identical to the proportions in the original Free Sampling Group of Hertwig et al's (2004) experiment, which is also reflected in a high correlation between the proportions ($r = .94$, $p = .006$). In order to test whether the choice proportions within the six choice problems differed between the two Free-Sampling Conditions, six separate contingency tables were constructed from the observed raw proportions. The

results of the Fisher's exact tests for these tables are summarised in Table 2.3. None of the six proportions differed significantly.

TABLE 2.3

Percentage of participants in the Free Sampling Groups who selected the H option

	Percentage choosing H		
Decision Problem	Free Sampling Hertwig et al. (2004)	Free Sampling Experiment 1	p-value (2-Tail) Fisher's exact test
1	88	73	.291
2	44	54	.579
3	28	23	.755
4	56	58	1.00
5	20	15	.726
6	12	23	.465

However, when comparing the proportions of participants choosing "H" under the Free-Sampling Condition in Experiment 1 with the proportions found in Hertwig et al's Description Group, a negative correlation was found, $r = -.57$, $p = .229$. This is also reflected on the level of individual choice problems. The p-values of the Fisher's exact tests on the comparisons within the six contingency tables are summarised in Table 2.4. Only the proportions of decision problems 2 did not differ significantly from the proportions in the description format. Across all 6 choice problems, the mean difference between the proportions under Free Sampling and descriptive choice in this experiment was 32%, which is close to the 36% found by Hertwig et al. (2004).

TABLE 2.4

Percentage of H choices under descriptive choice and under Free Sampling

	Percentage choosing H		
Decision Problem	Descriptive Choice Hertwig et al. (2004)	Free Sampling Experiment 1	p-value (2-Tail) Fisher's exact test
1	36	73	.012
2	64	54	.572
3	64	23	.005
4	28	58	.048
5	48	15	.017
6	64	23	.005

A different summary of this choice pattern is provided in Figure 2.4. The chart shows the differences in choice proportions between the DfXP and DfD formats across the six decision problems for both the Free-Sampling Condition in Experiment 1 and the Hertwig et al. data. The differences have been transformed so that the orientation of the bars can be interpreted in terms of the underlying difference in probability weighting. Positive bars indicate less overweighting of small probabilities in DfXP whereas negative bars indicate more overweighting in DfXP. The juxtaposition of the differences from the Hertwig et al. (2004) data shows that the direction in all six choice problems is identical to previous findings, implying less overweighting of rare events.

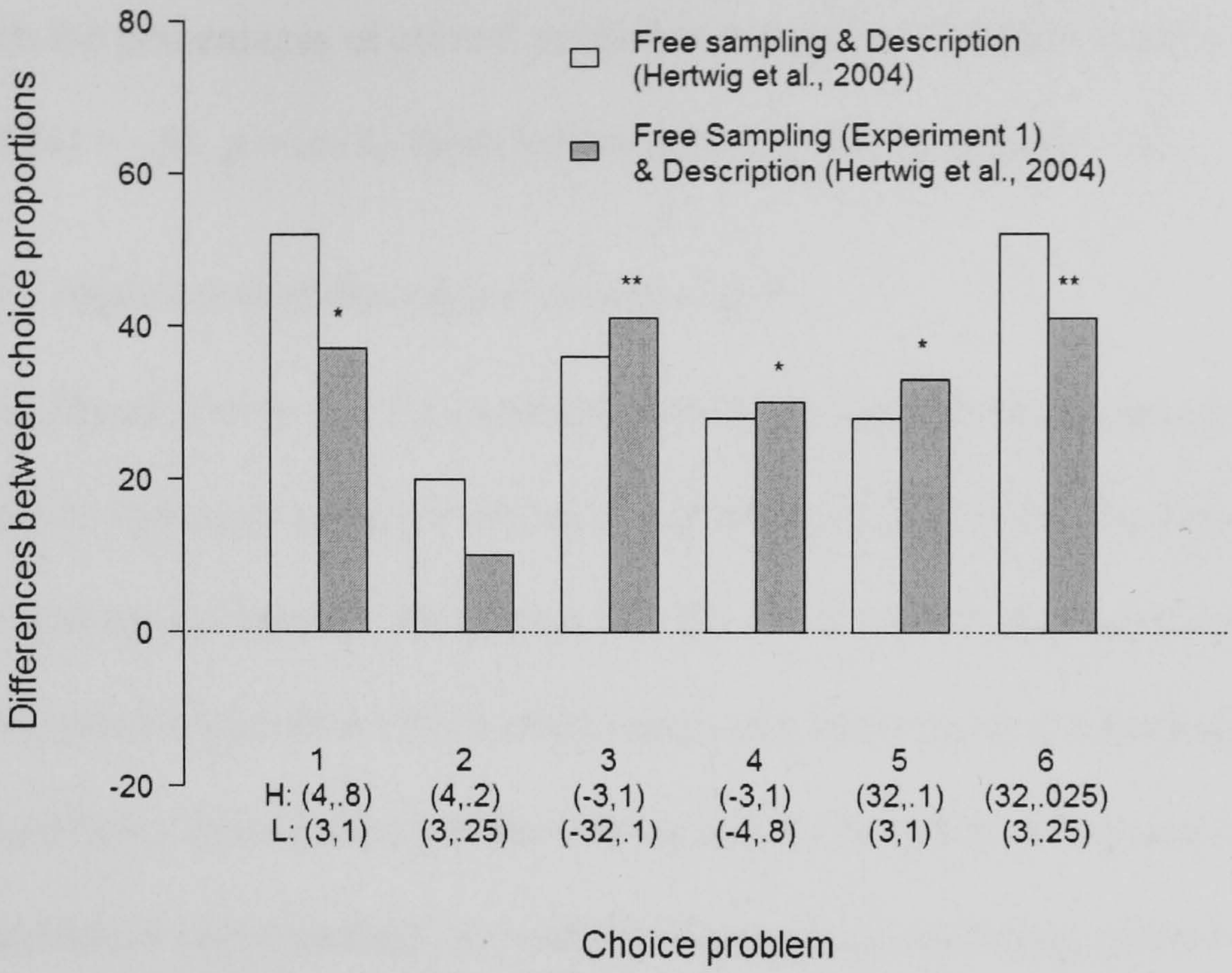


Figure 2.4. Differences in choice proportions between the descriptive and experiential choice proportions. The values are transformed so that positive values represent deviations in the direction of less overweighting. The white bars provide the original differences reported by Hertwig et al. (2004).

2.3.1.4 Recency weighting

As pointed out in the introduction, one of the main candidate explanations for the choice pattern has been recency weighting which assumes that the most recently encountered outcomes have more impact on choice than outcomes that have been observed earlier in the sequence. If this is true then recency weighting should be observable by comparing the predictive power of different parts of the experienced sequence. Following the analysis conducted by Hertwig at al. (2004), I split the outcome sequences of each button into two halves and calculated their expected values. The option with the highest expected value was predicted to be chosen, separately for the first and second half. A comparison of the predicted choices with

the actually observed choices showed that there was no significant difference between the percentages of correct predictions of earlier and later samples (74% vs. 76%, $t(24) = -.51$, $p = .612$). Both halves predicted choice equally well.

2.3.1.5 Application of descriptive choice models

Thus far I have shown that the choice behaviour deviates from the one usually observed in descriptive choice, replicating Hertwig at al. (2004). The question now is how well established choice models like EV or PT, which have been developed on the basis of descriptive choice phenomena, can account for the findings presented here. This section will provide an initial test of the performance of expected-value based models. A more detailed analysis including additional models used in related task and a comprehensive comparison will be presented in Chapter 6. Given the obvious deviations between objective and experienced probabilities due to the skewed distribution of the rare events, it is not appropriate to determine any model fits based on the objective probabilities. Instead, I tested the predictions of (a) a simple EV model and (b) prospect theory (PT) based on the probabilities that the participants have actually experienced within the sampled outcome sequences. When using a simple EV calculation for each sequence, 76% of the choices could be predicted correctly, which is in agreement with the 74% reported in the original data (Hertwig et al., 2006). A PT model, which requires additional parameters for the value and weighting function, could only account for 65% of the choices correctly when using the median value- and weighting-function parameters reported by Tversky and Kahneman (1992). Across participants this advantage of the EV model's performance was significant ($t(25) = 2.46$, $p = .021$, two-sided).

2.3.2 Comprehensive Sampling

2.3.2.1 Information search

Due to the nature of the Comprehensive-Sampling task (39 samples for each option) there was no difference between participants in terms of the number of samples drawn from each option. The switching strategies did differ and a pattern analogous to the one reported earlier was observed. Again, the majority of the sequences (68%) were explored with a few shifts between options. Less frequently observed were pure strategies like consistent alternation (7%) or exhaustive sampling of the 39 samples from one option before exploring the remaining option (25%). With a median (mean) switch ratio of .06 (M = .19) participants showed less switching than in the Free-Sampling Condition. The transition stability across all choice problems was also slightly higher than in the Free-Sampling Condition (median correlation of r = .84).

2.3.2.2 Experienced probabilities and sampling error

With the higher number of 39 samples the binomial distribution should overall be less skewed and therefore allow a more accurate assessment of the underlying probabilities. This is supported by the data. The mean absolute difference between the objective probabilities and the probabilities experienced during the 39 samples was much lower than under Free Sampling with only 4% (SD = 3.7). However, because of the odd number of samples people were also less likely to experience probabilities identical to the objective probabilities. As a result participants were shifted towards under- or overrepresentation of the rare events. This is also reflected in the observed frequencies of the rare events. In the majority of the cases (62%) participants encountered the rare event more frequently than expected.

Underestimation only applied to 38% of the cases. More importantly, due to the increased sample size there were only ten cases (7%) in which the rare event was not encountered a single time. All of these cases occurred in choice problem 6 where the probability of the rare event was only 2.5%. None of the sequences though allowed a completely accurate assessment of the underlying probability. The overall distribution of the differences between experienced and objective probabilities was therefore much slimmer than in the Free-Sampling Condition and also less skewed ($M = 0.5\%$), as illustrated in Figure 2.5. As a result of the different sampling mechanism in the Comprehensive-Sampling Condition underrepresentation is much less frequent. If sampling error in the form of underrepresentation is supposed to be the main reason for the apparent underweighting of small probabilities the effect should be less pronounced here.

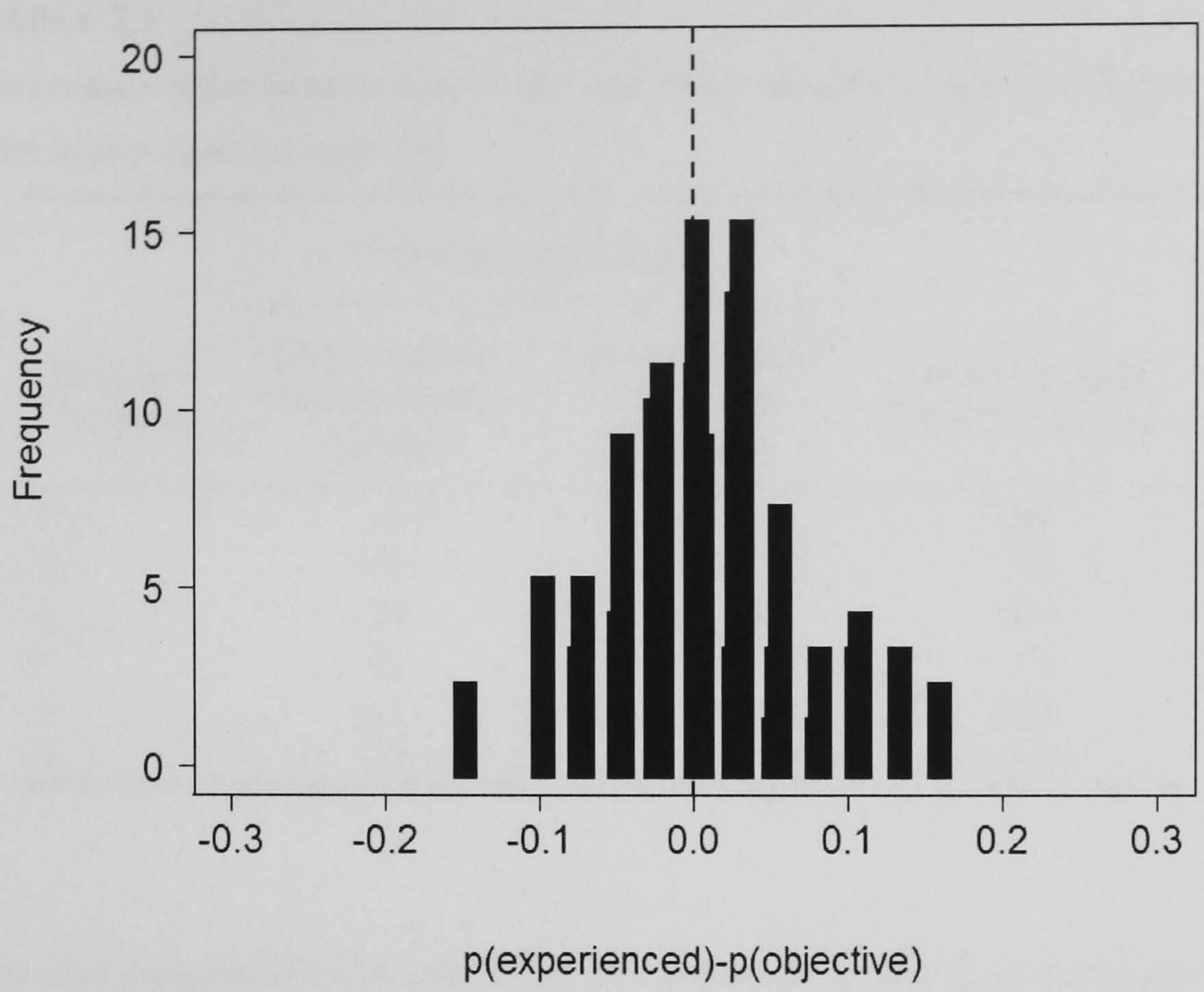


Figure 2.5. Histogram with the distribution of differences between experienced and objective probabilities under Comprehensive Sampling. The bars to the left of 0 mark underrepresentation of the objective probabilities and the bars to the right mark overrepresentation of the objective probabilities.

2.3.2.3 Choice behaviour

Although the sampling behaviour prior to choice differed, the actual choice behaviour within the Comprehensive-Sampling Condition was comparable to the original Free-Sampling Condition of Hertwig et al. (2004), which is shown in a high correlation between the choice proportions of the two conditions, $r = .81$, $p = .049$. Moreover, Table 2.5 shows that there are also no significant differences between the proportions of H choices of the two sampling conditions within the individual choice problems.

TABLE 2.5

Percentage of participants under Free- and Comprehensive-Sampling selecting the option with higher expected value (H)

Decision Problem	Percentage choosing H: Free-Sampling Hertwig et al. (2004)	Percentage choosing H: Comprehensive-Sampling Experiment 1	p-value (2-Tail) Fisher's exact test
1	88	64	.095
2	44	52	.778
3	28	40	.551
4	56	64	.773
5	20	48	.072
6	12	16	1.000

The direct comparison of both experiential conditions of the first experiment exhibits very similar results. Although based on different sampling modes comparable proportions of H choices were obtained for both conditions, $r = .73$, $p = .099$. Within the individual choice problems the only significant difference was found in decision problem 5 (see Table 2.6).

TABLE 2.6

Percentage of H choices under Free- and Comprehensive-Sampling

Decision Problem	Percentage choosing H: Free Sampling Experiment 1	Percentage choosing H: Comprehensive-Sampling Experiment 1	p-value (2-tail) Fisher's exact test
1	73	64	.555
2	54	52	1.00
3	23	40	.237
4	58	64	.776
5	15	48	.017
6	23	16	.726

However, when comparing the choice proportions under Comprehensive Sampling with the descriptive choice proportions reported by Hertwig et al. considerable differences were found between the two formats, $r = -.75$, $p = .084$ (see Table 2.7). The average (absolute) difference between the percentages was 25%, which is still high, but less extreme than in the Free-Sampling Condition.

TABLE 2.7

Percentage of participants choosing the H option under descriptive choice and under Comprehensive-Sampling

	Percentage choosing H		
Decision Problem	Descriptive Choice Hertwig et al. (2004)	Comprehensive-Sampling Experiment 1	p-value (2-Tail) Fisher's exact test
1	36	64	.089
2	64	52	.567
3	64	40	.156
4	28	64	.022
5	48	48	1.00
6	64	16	.001

The relative comparison in Figure 2.6 shows again the transformed differences between the choice proportions of the two formats. Once more, the direction of the differences for five out of six bars is the same as in previous DfXP experiments and points in the direction of less overweighing of rare events under DfXP. The only exception is decision problem 5 for which identical proportions of maximising choices have been obtained in the descriptive and experiential choice format. However, this anomaly appears to be random and does not follow a specific pattern, which will become more evident throughout the following experiments.

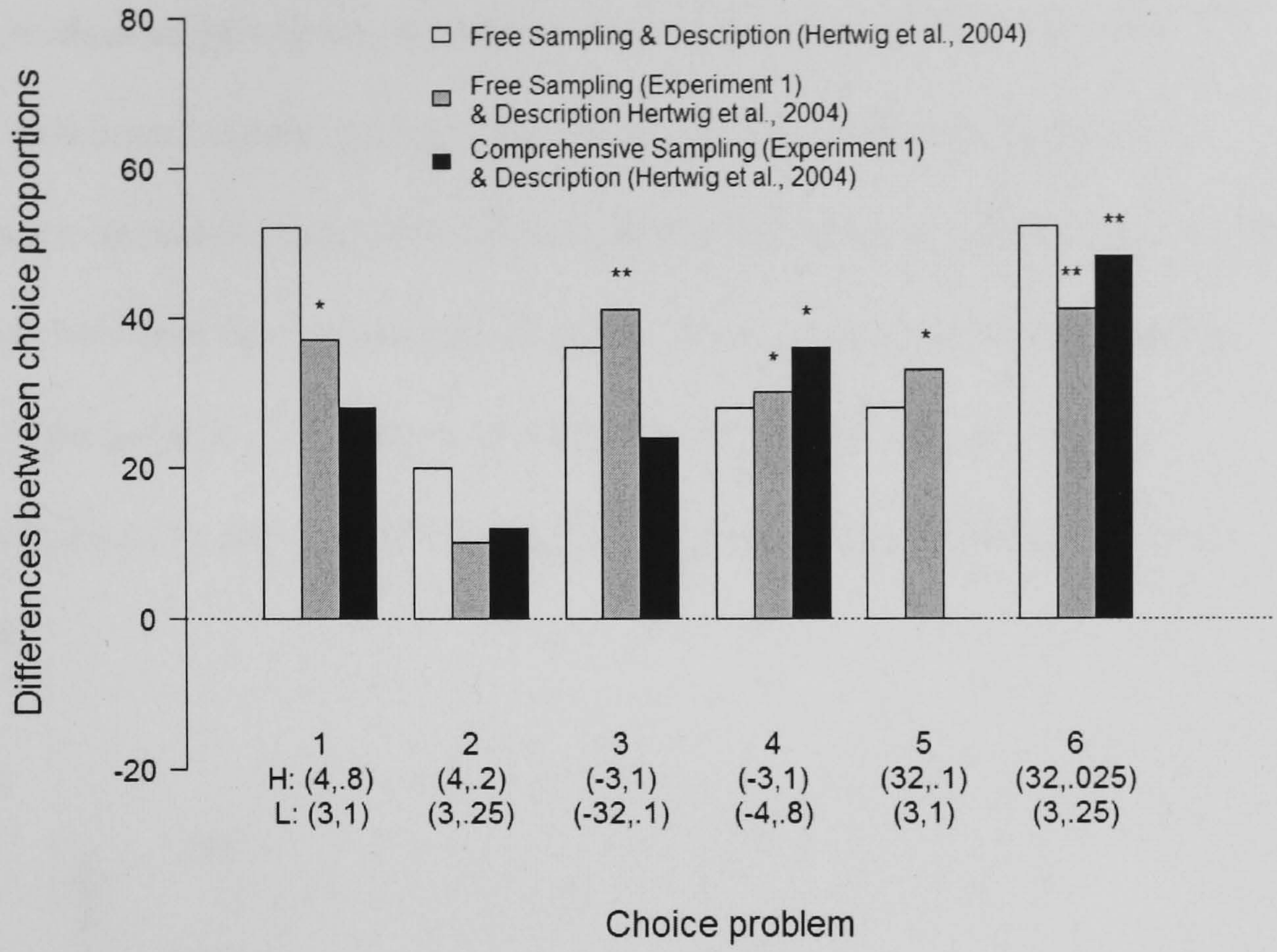


Figure 2.6. Differences in choice proportions plotted as differences between the descriptive and experiential choice proportions. Again, the values are transformed so that positive values represent deviations in the direction of underweighting. The white bars provide the original differences reported by Hertwig et al. (2004).

2.3.2.4 Recency weighting

Again recency weighting was examined by looking at the rate of correct predictions of the two halves of the sampled sequences on the basis of their expected values. There was no evidence for a higher predictive power of the more recently sampled outcomes. In fact, the opposite trend was observed with 65% correct predictions based on the outcomes from the first half and 58% correct predictions from the second half, but this difference was not significant ($t(24) = 1.26$, $p = .219$, two-tailed). Finally, I compared the proportions of correct predictions based on the expected values of the different quartiles of the sequence

by conducting a one-way repeated measures ANOVA with the different quartile splits as a within-subject factor. Mauchly's test indicated that the assumption of sphericity had been violated ($\chi^2(5) = 12.755, p = 0.026$); therefore degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = .74$). The results show that the proportions of correct predictions are not significantly affected by the position of the quartile ($F(2.204, 55.896) = .174, p = .861$), indicating that earlier samples predict choices as well as later samples (see also Figure 2.7).

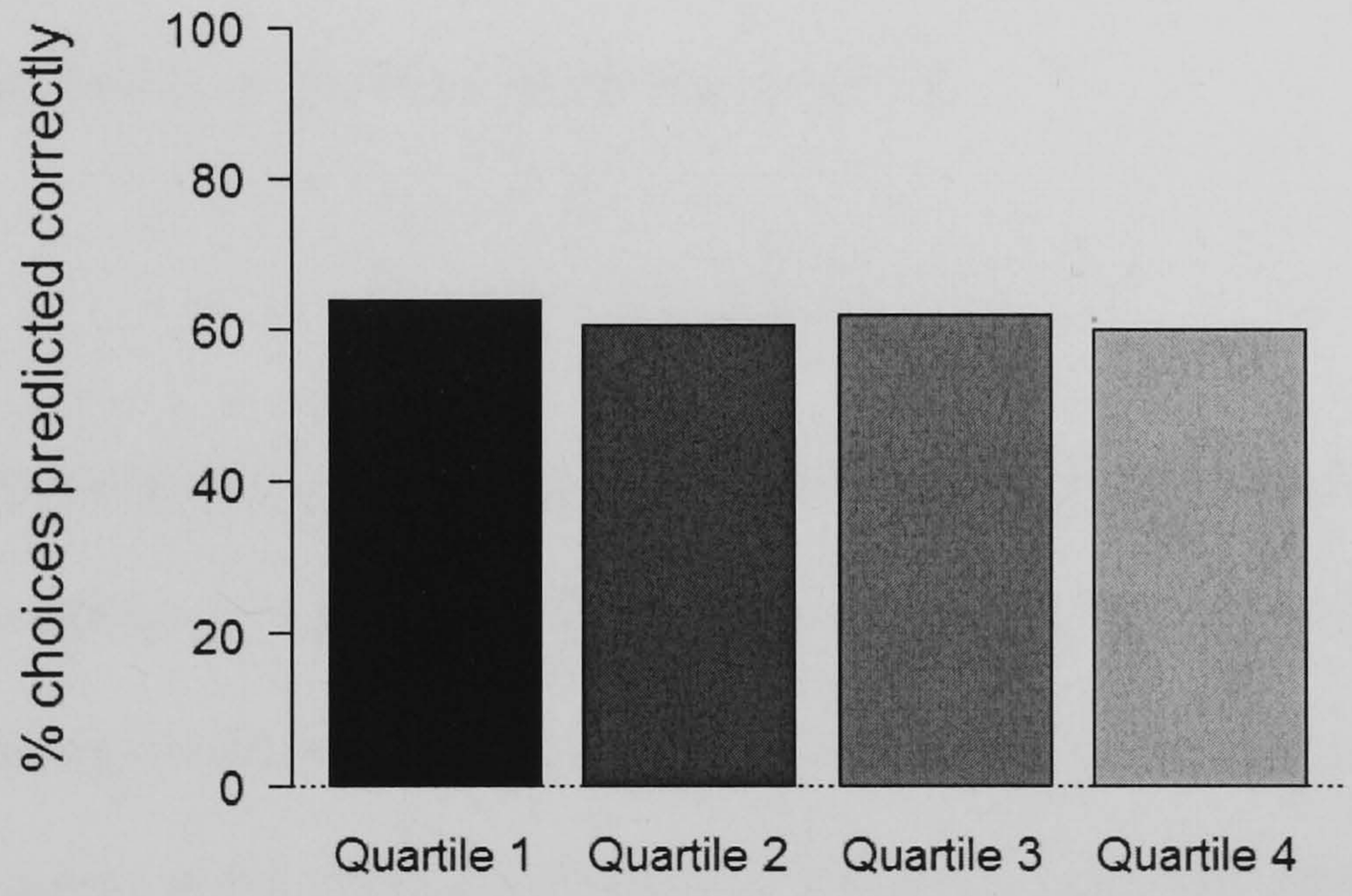


Figure 2.7. Mean percentages of correct predictions for the different quartiles. All quartiles provide similar levels of correct predictions.

2.3.2.5 Application of descriptive choice models

Under Comprehensive Sampling, expected value maximisation based on the actually experienced probabilities could account for only 65% of the choices correctly. The predictive power of PT with the median value- and weighting-function parameters reported by Tversky and Kahneman (1992) did not exceed chance level (51%). Again, the advantage of the EV model was significant ($t(24) = 2.93$, $p = .007$, two-sided). When comparing the predictive power between the two sampling conditions in Experiment 1 both models, EV ($t(49) = 2.07$, $p = .044$, two-sided) and PT ($t(49) = 2.96$, $p = .005$, two-sided), performed significantly better on the basis of the data from the Free-Sampling Condition.

2.4 Discussion

The main purpose of this first experiment was two-fold. Firstly, I aimed to test the stability of the choice pattern observed by Hertwig et al. (2004). The results from the Free-Sampling Condition can be interpreted as a successful replication of almost every aspect of the original experiment, including sampling behaviour, its impact on the representation of the available options and the distinct choice pattern. In terms of the information search prior to choice analogous sample sizes have been observed and a comparable pattern of underrepresentation of low probability events has been replicated. The maximising choice proportions found in the Free-Sampling Condition, also matched the data mentioned in earlier experiments and deviated in the same direction from their descriptive counterparts, providing additional support for less overweighting of small probabilities in DfXP. The same

applies to the predictive power of the EV model based on the participants experienced probabilities. However, no support was found for recency weighting. Early and recent samples predicted choices equally poorly.

Secondly, I included the Comprehensive-Sampling Condition to test whether the reversed choice behaviour can still be observed when sampling error is reduced as a consequence of increased sample size. Notably, due to the higher number of random samples drawn, the extent of sampling error in the Comprehensive-Sampling Condition was substantially reduced and there was also less underrepresentation of rare events. However, I still found significant differences in the direction of less overweighting of small probabilities in the choice behaviour, although to a lesser degree than under Free Sampling. This provides evidence for a DfXP choice pattern under conditions were undersampling is less apparent. Again no recency effect was observed and the extent with which descriptive choice models can account for the data is significantly reduced. The results presented here are confirmed by findings recently reported by Hau, Plescak, Kiefer and Hertwig (in press), who also found a remaining gap between descriptive and experiential choice proportions when using incentives to motivate participants to explore the options more thoroughly.

In general, this experiment confirms the robustness of the reversed choice pattern under DfXP and reduced overweighting of small probabilities within experiential choice. However, as we actually observe differences between choice proportions that cross the 50% line, implying actual reversals in participants' modal choices, this reduction of overweighting can be interpreted as apparent underweighting of small probabilities. Whether people actually underweight small probabilities, however, has to be inferred from the actual shape of the weighting

function and its parameters, which will be addressed in more detail in Chapter 5 and 6.

Moreover, the experiment indicates that phenomenon holds even when the samples provide representations with less sampling error in the form of underrepresentation. This extension seems to be contradictory to the predictions of the sampling error hypothesis by Fox and Hadar (2006). However, due to the uneven number of samples the experienced probabilities are not exactly matching the objective probabilities and residual sampling error remains. The descriptive choice problems and their experiential counterparts are still not structurally identical. It is therefore not possible to refute sampling error as one of the causes for the effect. The observed reduction of the effect in the second condition seems to indicate that sampling error is involved in causing the phenomenon but is not sufficient to explain it. To clarify the impact of deviations from objective probability it would therefore be necessary to alter the design in a way that allows the presentation of an outcome sequence without any sampling error, completely mirroring the descriptive choice problems.

Furthermore, this first experiment does not provide any support for the existence of recency weighting, which was originally assumed to be involved in causing the choice pattern. In the Free-Sampling Condition this seems to be the only aspect of the original study that could not be replicated. If recently-sampled information has more impact on choice, then the impact of outcomes experienced earlier in the sequence should be even smaller the higher the total sample size. However, even in the Comprehensive-Sampling Condition there is no significant difference between the predictive power of the first and second half of the outcome sequence or between the predictions of the different quartiles. This test was based

on the assumption though that people extract information regarding the outcomes and their probabilities in order to maximise expected value, as it was done in similar experiments. Thus, there might still be some underlying recency effect when predicting choice based on other properties of the experienced sequences. This point will be discussed further in Chapter 4 and 6.

Regarding the accountability of established choice models, two observations are important to mention. Firstly, the overall fit with actual choice is lower than is usually observed in the descriptive choice experiments. Secondly, and more interestingly, under both conditions the simpler EV model predicts choice significantly better than PT. This is the opposite of what is usually observed when gambles are presented as summary descriptions. It is important to point out though that this comparison is made with a PT model that is based on the parameters that have been established in descriptive choice tasks. One possibility is that the weighting function does not provide the same improvement for the PT model in the context of experiential choice that we normally find in descriptive choice tasks because the probability weighting function under DfXP differs from the one established under DfD. A better predictive power for PT might therefore be achieved with weighting function parameters estimated on the basis of experiential choice data. A more comprehensive model comparison will be provided in Chapter 6.

One potential problem with the Comprehensive-Sampling Condition is that it does not only differ in terms of the sample size but also in terms of the control over when to stop sampling. As participants in the second condition were forced to continue the sampling process until they reached the fixed number of samples it is not possible to investigate the potential influence of individual stopping rules. On

the other hand, the experimental results by Hau et al. (in press) mentioned above, seem to indicate that similar results can also be obtained when individuals draw larger samples without losing control over when to stop. Another problem could be the usage of an uneven number of samples which makes it more difficult to translate the experienced sequence into a description format. Participants might have extrapolated in different ways and might have used probabilities different from those deriving directly from the frequencies that they actually experienced to make their final decision.

This first experiment shows once more how difficult it is to identify the properties behind the reversed choice pattern in experiential choice tasks as the sampling process in the form it has been used until now is actually changing the structure of the underlying task. A comparison with descriptive choice is therefore difficult as both tasks are in practice no longer identical. The clarification of the effect of sampling error in particular demands a design that provides structurally similar tasks. In Chapter 3, I will therefore introduce a new experiential choice paradigm which will help to overcome experimental flaws that have been underlying previous designs. By using a novel sampling mechanism this new experiential task will resemble descriptive choice more closely. More importantly, by providing outcome sequences matching the objective probabilities, this approach will allow a more suitable investigation of the impact of sampling error. This will bring us back to the question of whether choice behaviour depends on whether the choice problem is presented in the form of a description or is experienced as a sample of outcomes.

CHAPTER 3

THE MATCHED SAMPLING DESIGN

3.1 Introduction

In the previous chapter, I demonstrated that the unusual choice pattern in decisions from experience can be replicated even under conditions where sample size is increased and sampling error is consequently reduced. Differences in choice proportions were still found but the gap was noticeably reduced. Therefore, it remains unclear whether the gap between decisions from experience (DfXP) and decisions from description (DfD) can be completely closed if participants are forced to experience perfectly representative samples. Yet, with the design used so far, sampling error can not be eliminated. In this chapter, I present an alternative design in which the frequencies people sample precisely match the underlying probabilities of the options. This eliminates sampling error completely and so allows a direct test of the sampling error hypothesis. If Hertwig et al. (2004) and Fox and Hadar (2006) are correct, then apparent probability underweighting should be eliminated, or indeed, reversed under such experimental conditions.

3.2 Matched Sampling in the Lab (Experiment 2)

The following laboratory experiment will introduce the Matched Sampling design. By sampling exhaustively and without replacement from an underlying distribution where the probabilities exactly match those in the descriptive choice, it is possible to compare experiential choice under conditions where sampling

error is eliminated with choice under Free Sampling (where sampling error remains) and descriptive choice.

3.2.1 Methods

3.2.1.1 Participants and Stimuli

75 participants were recruited from students at the University of Warwick through flyer advertisement and a subject panel of the Psychology department. The majority (88%) of the participants were first year undergraduate students from different disciplines; the rest consisted of postgraduate students and members of staff. The age within the sample ranged from 18 to 52 years with an average age of 21. The gender split was approximately 2:1 with 46 male and 29 female subjects. For the completion of six choice tasks in the laboratory the participants received £2. Performance dependent incentives were not provided. The six decision problems were the same used in the previous experiment.

3.2.1.2 Design and procedure

The experiment was implemented as a between-subject design with three different conditions; a Free-Sampling Condition, a Matched-Sampling Condition, and a Description Condition to which participants were assigned randomly. The two experiential conditions consisted of a sampling phase in which participants explored the two options represented by two buttons, 'A' and 'B', on a computer screen, followed by a final decision phase where they chose the option they would like to play once for real. The Free-Sampling Condition follows Hertwig et al.'s (2004) paradigm: Participants could stop the exploration of the buttons as soon as they felt confident enough to make a decision; and outcomes were drawn with replacement for each participant from the underlying distributions.

In the crucial Matched-Sampling Condition, however, I ensured that the experienced probabilities for a given option matched the objective probability of its underlying distribution. This was achieved by several alterations to the Free Sampling design. Firstly, participants had to sample exactly 40 outcomes from each option. Once this limit was reached the button became shaded and no further sample could be obtained. Only after both buttons had been sampled 40 times could the participant proceed to the decision phase. More importantly, the proportions of outcomes within this sequence of 40 events precisely matched the options' underlying probabilities, with the order of outcomes randomly generated for each participant (see also Figure 3.1). Finally, the sampling mechanism was altered. Instead of sampling with replacement, participants sampled exhaustively from 40 outcomes, with the frequency of each outcome exactly matching the probabilities from the Description Condition. Together, these properties completely eliminate sampling error from the experienced sequence. For example, in the first decision problem (3, 1.0; 4, .8), the second option, offering 4 points with a probability of .8 and nothing otherwise, can be represented as a sequence of forty outcomes with 32 '4's and eight '0's. When sampling all 40 outcomes from this sequence, participants will experience the exact objective probability (see also Figure 3.1).

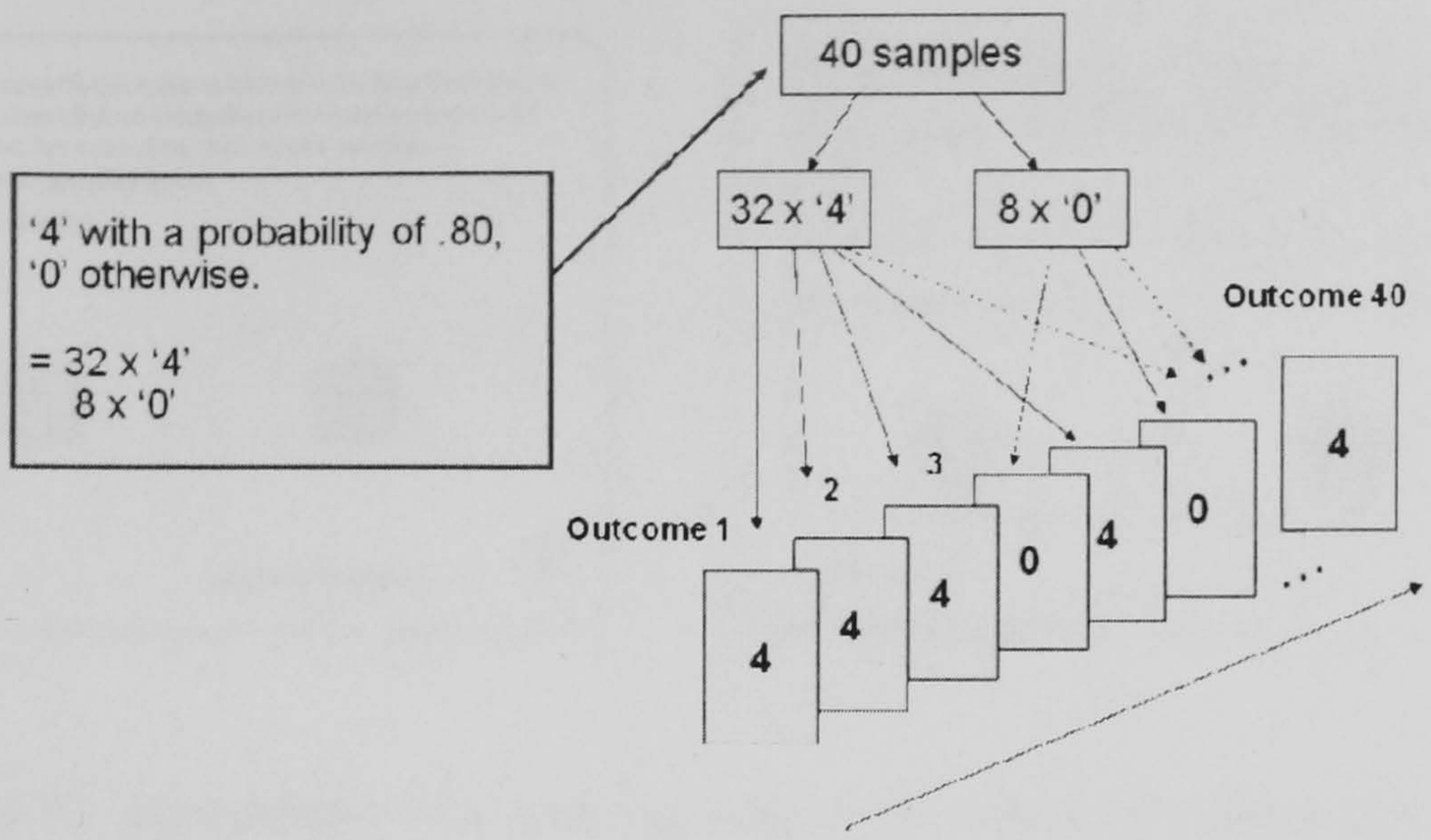


Figure 3.1. Scheme of the matching process under Matched Sampling to eliminate sampling error.

For both, the Free- and the Matched-Sampling Conditions, the order in which to sample from the two buttons during the sampling phase was arbitrary. The basic layout of the sampling experiments was the same as in the first experiment (see Figure 3.2).

The Description Condition involved the presentation of summaries of the same lotteries in the format:

20% chance to win 4 points;

80% chance to win 0 points.

The Description Condition provided an alternative set of descriptive choice data which will allow a within-experiment analysis of the data rather than a between-experiment comparison with the proportions reported in the original experiment by Hertwig et al. (2004).

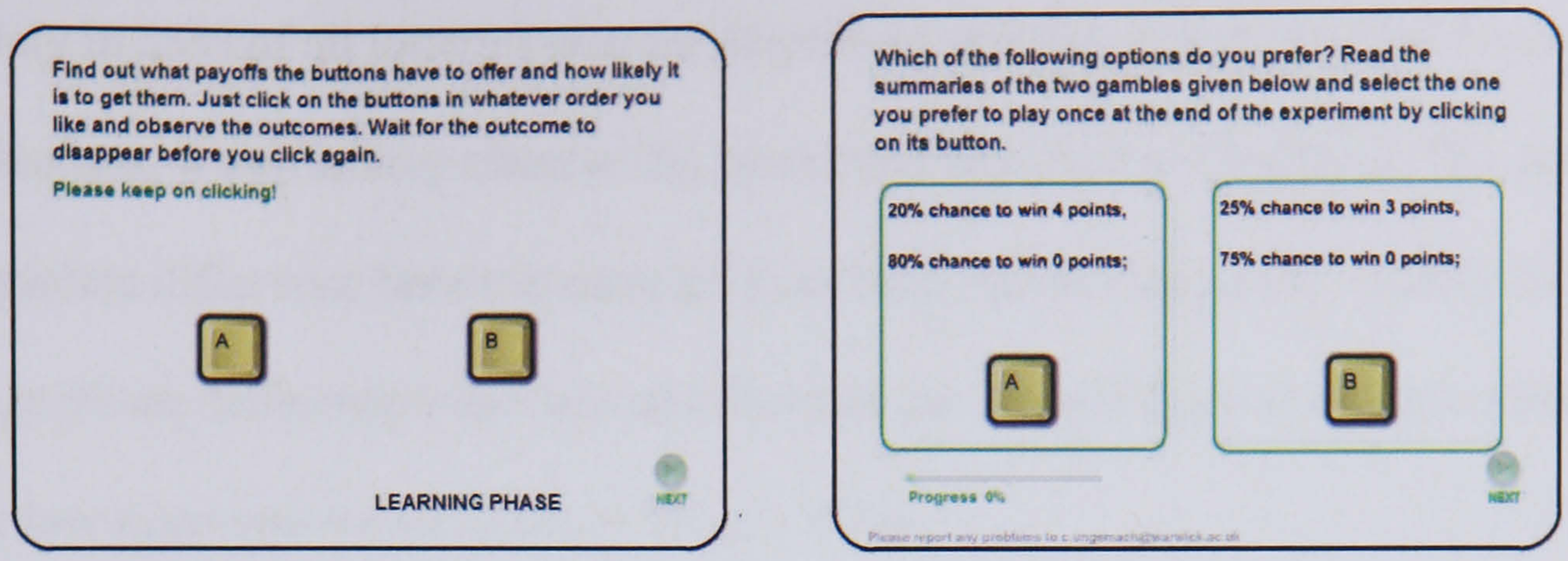


Figure 3.2. Screenshots of the sampling phase of the Matched-Sampling Condition (left) and the gamble description in the Description Condition (right).

All participants were instructed to maximise the number of points they accumulated within the six choice problems. At the end of the experiment, the selected lotteries were played randomly for each participant before they were informed of their points total.

3.2.2 Results

Similar to the presentation of results in Experiment 1, I will first provide a summary of the sampling behaviour within the two experiential conditions before I compare the observed choice behaviour across all three conditions. In the context of the Free-Sampling Condition particularly, the information search pattern and the resulting representation of the different outcomes is a prerequisite to understanding the choice pattern discussed later on.

3.2.2.1 Information search under Free Sampling

With a median number of 19 ($M = 24$) draws per choice problem, participants did sample slightly more data than in the previous experiments using Free Sampling. Across the different choice problems the number of draws turned out to be less stable than in earlier Free Sampling tasks with a median correlation of $r = .42$.

Only in 26% of all lotteries was the number of samples drawn from both options identical, which is very close to the percentage reported in Chapter 2. The median absolute difference between samples from both options was 2 ($M = 4.43$). No significant difference was observed between the sample sizes of options with high or low expected values ($t(24) = .57, p = .573$).

The median number of switches observed was 3 ($M = 6.27$). The actual distribution of switches is provided in Figure 3.3. When put in context of the possible number of switches the median switch ratio was .16 ($M = .38$) which indicates slightly more switching than in the previous experiments. Consistency was also found in the sampling strategies where a total of 29% of all sequences were completed only switching once. Alternating sampling was observed in 13% of the sequences, leaving 58 % for mixed strategies. The switching also proved to be relatively stable across the six problems with a median correlation of $r = .77$.

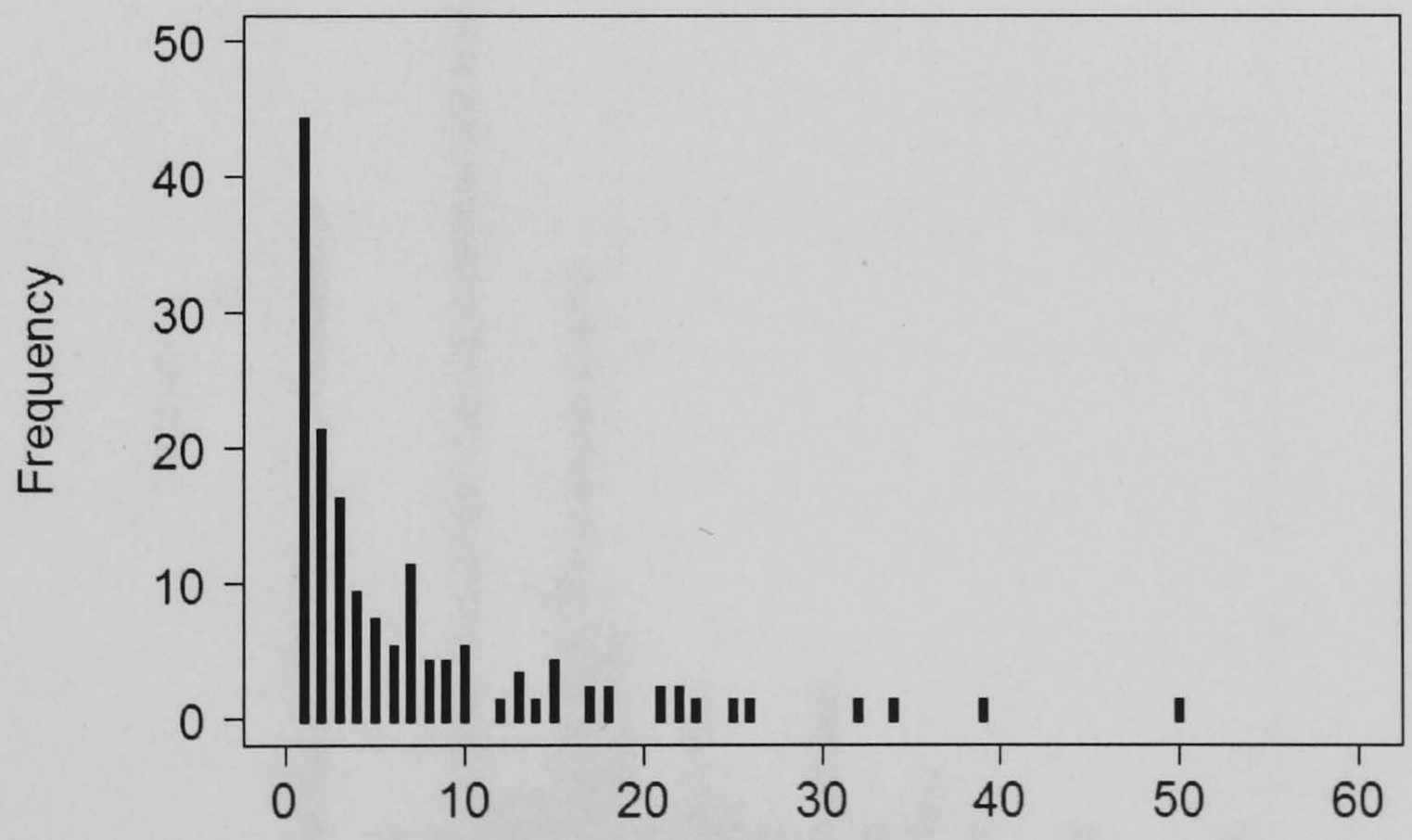


Figure 3.3. Histogram with the number of switches observed under Free Sampling

Inspection of the differences between the experienced probabilities and the objective probabilities of the rare events revealed a median absolute difference of 6% (SD = 10.83) and, across all problems, the rare event was encountered less frequently than expected in 50% of the sequences. This is still substantial but less than observed within the replication of the previous chapter (69%). The reduced positive skew in the difference between observed and objective probabilities could be due to the higher number of samples drawn. In 71% of the reported cases of underrepresentation (the same percentage found in the first experiment) the rare event was not encountered at all. This equals 35% of all the sequences involving rare events. Sequences with relative frequencies of rare events that actually matched the rare event's objective probability were only experienced in five percent of the cases (see also Figure 3.4).

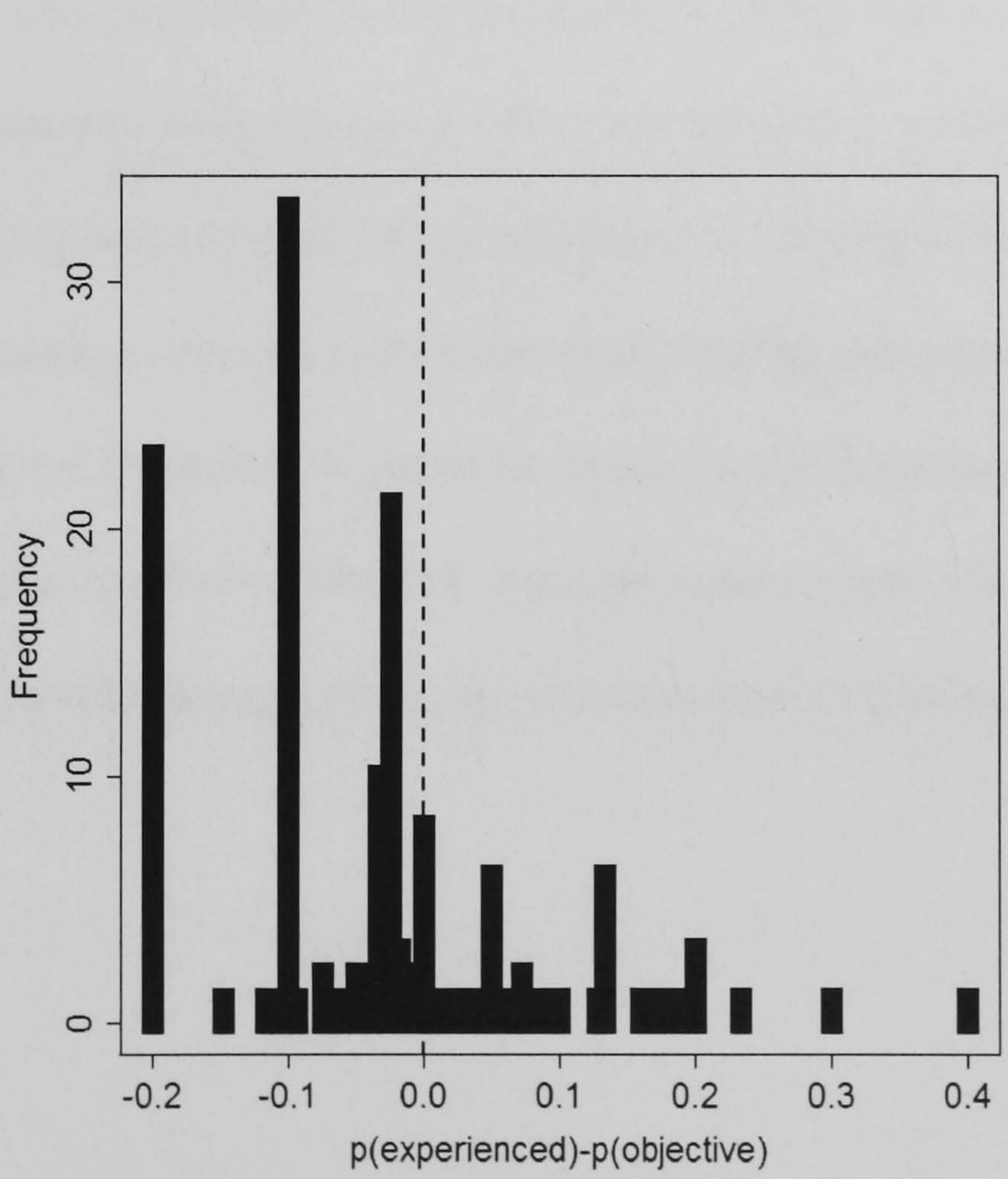


Figure 3.4. Histogram with the distribution of differences between experienced and objective probabilities in the Free-Sampling Condition.

In summary, the sampling process in the Free-Sampling Condition resembles all the properties typical for DfXP formats with undersampling resulting in extensive sampling error in the direction of underrepresentation of the rare outcomes.

3.2.2.2 Information search under Matched Sampling

In the Matched-Sampling Condition both options had to be explored 40 times. However, in this condition we could still observe switching between the two buttons during sampling (see Figure 3.5 for a histogram depicting the distribution of switches). This happened less often than under Free Sampling with a median number of switches of 1 ($M = 4.79$). Given a high number of 79 potential switches this led to very small switch ratios with a median of only .01 ($M = .06$). In terms of the sampling strategy used, the picture was quite different to the observations under free sampling. The majority of the sequences (75%) were explored with only one alternation between the sets of forty outcomes from each button. Constant switching was found to be used by only 1%, leaving 24% with mixed strategies. The median correlation between the switching ratios across the six problems was only .21 though. A potential reason for the decreased swapping could be the higher number of samples: it simply requires less effort to stay on one button and sample through all the 40 outcomes than switching between buttons.

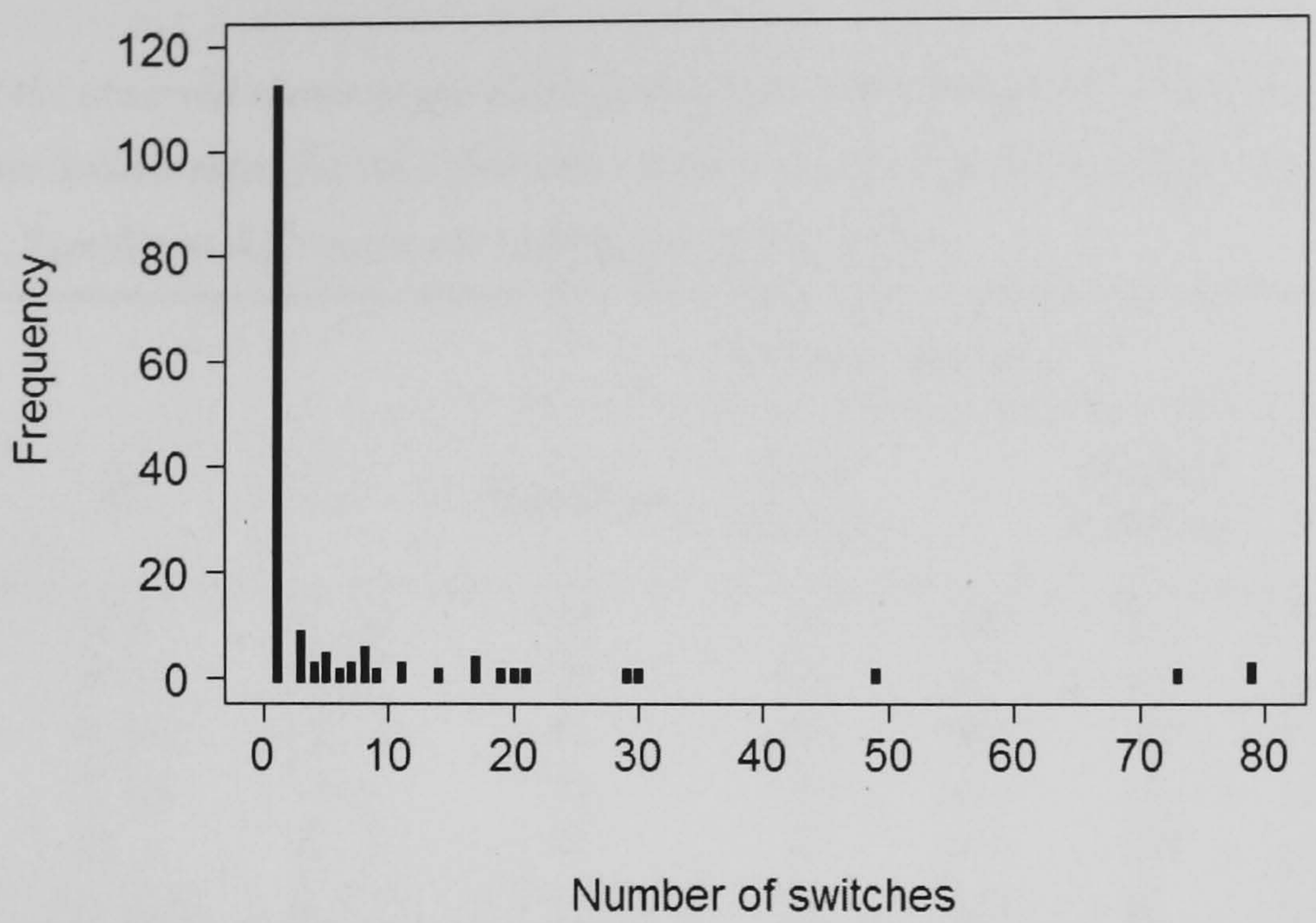


Figure 3.5. Histogram with the number of switches observed under Matched Sampling

3.2.2.3 Choice behaviour

The proportions within the Description Condition were remarkably similar to the descriptive choice proportions reported by Hertwig et al. (2004) ($r(4) = .89$, $p = .02$, with a mean absolute difference of only 4.67%. The proportions of participants choosing the H option across the six choice problems within the three conditions are summarised in Table 3.1, which shows that most of the proportions differ across the 50%-line (i.e., with the description proportion and the experiential proportion either side of .5), indicating actual reversals of preferences. This applies to five out of six problems under the Free-Sampling Condition and two out of six problems in the Matched-Sampling Condition.

TABLE 3.1

Summary of the observed choice proportions for the three experimental conditions including the p-values (Fisher's exact tests) for the differences between the experiential and descriptive choice proportions. Significant differences are highlighted with asterisks.

			Percentage choosing H				
Decision Problem	H	L	Description	Free Sampling	*p*	Matched Sampling	*p*
1	4, .8	3, 1.0	36	64	.089	48	.567
2	4, .2	3, .25	72	56	.377	60	.551
3	-3, 1.0	-32, .1	64	16*	.001	28*	.022
4	-3, 1.0	-4, .8	36	68*	.046	32	1.00
5	32, .1	3, 1.0	48	8*	.003	16*	.031
6	32, .025	3, .25	52	28	.148	28	.148

A further illustration is provided in Figure 3.6 which shows the differences between proportions converted so that positive bars indicate deviations in the direction of less overweighting of small probabilities in DfXP. All the bars, except one, have the same direction consistent with the original findings. The only exception is choice problem 4 in the Matched-Sampling Condition where the percentage of H choices is more similar to the one found in DfD.

However, when comparing the average differences in proportions in the direction of less overweighting of small probabilities across all six choice problems (the mean bars in Figure 3.6), there were significant differences between choice proportions in Description Condition and the Free-Sampling Condition ($t(48) = 6.03, p < .0001$) and between the Description Condition and the Matched-Sampling Condition ($t(48) = 3.43, p = .001$), independent of sampling method. Furthermore, a comparison between the two experiential conditions shows that in the Matched-Sampling Condition, where sample size is controlled and sample frequencies precisely match the underlying probabilities, the apparent underweighting is reduced in comparison to the Free-Sampling Condition ($t(48) =$

-2.61, $p = .012$, two-sided). Crucially, though, it is not eliminated. Thus, sampling error seems to explain some of the differences, but is not sufficient to account for the whole phenomenon.

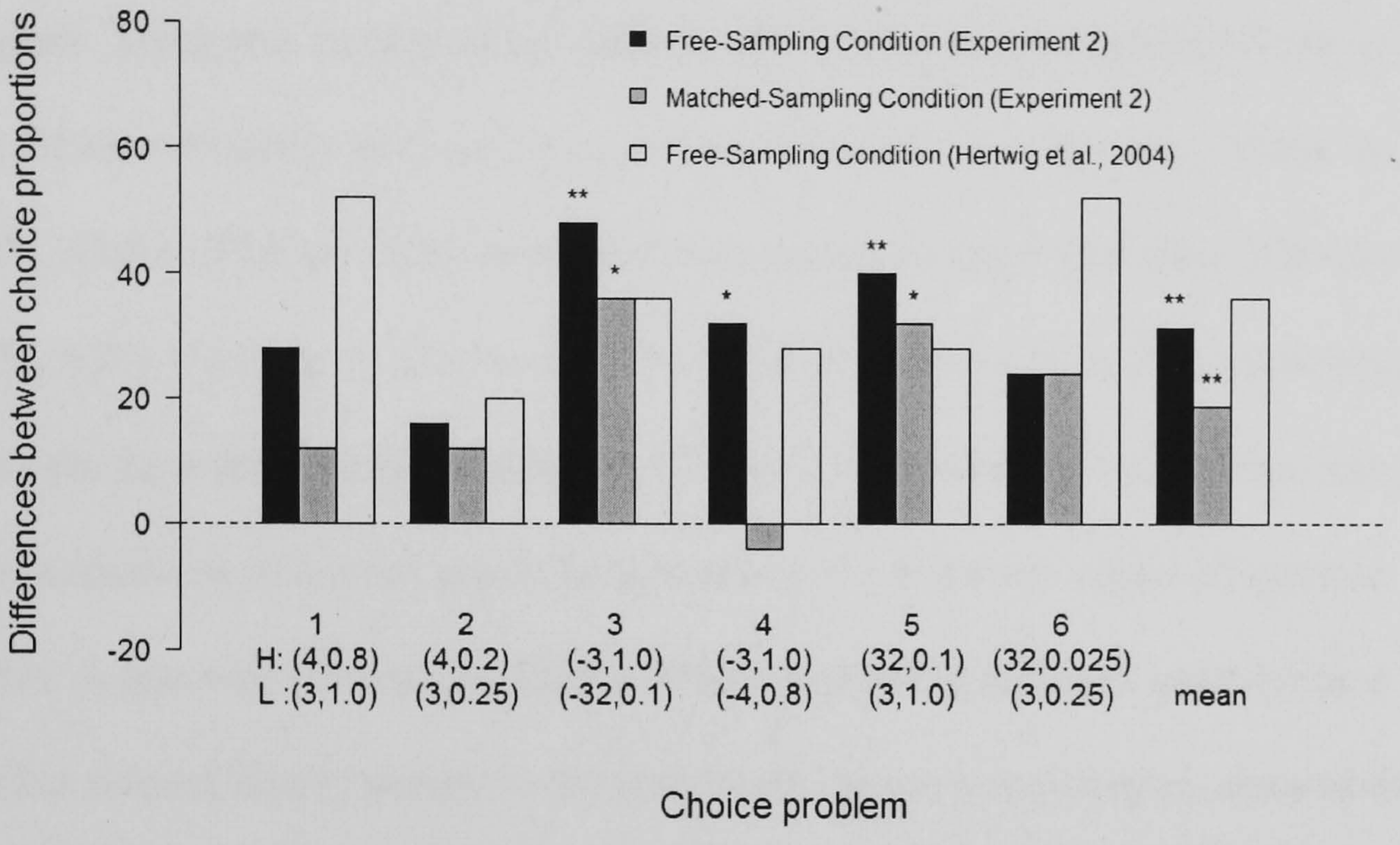


Figure 3.6. Differences in choice proportions between the experiential and descriptive choice tasks across the different decision problems used. Positive bars indicate less overweighting under experiential choice. The results reported by Hertwig et al. (2004) have been added for comparison. Significant differences are marked by asterisks (* $p < .05$, ** $p < .01$).

3.2.2.4 Recency weighting

One potential explanation why the effect is maintained with matched, equal samples for each option is that, even when presented with a large sample, it is possible that people can only remember a small sample of the most recent items. Thus although the matched sample accurately represents the objective probabilities, if only a part of this sample can be held in memory, then this sample will typically underrepresent rare events. If people only remember the most recent

samples, then more recent outcomes should predict actual choices more accurately then outcomes sampled earlier. Following the analysis conducted in Chapter 1, I split the outcome sequences for each button into two halves and predicted final choice separately for both buttons and both halves on the basis of the expected value of the outcomes included in sequence splits, assuming people to choose the highest. There was no significant difference between the percentage of correct predictions of earlier and later samples, neither under Free Sampling (69% vs. 65%, $t(24) = 0.54$, p = .596, two-sided) nor under Matched Sampling (48% vs. 42%, $t(24) = 1.12$, p = .272, two-sided). With 40 samples from each participant, the data from the Matched- Sampling Condition also allowed for a comparison of the proportions of correct predictions based on the expected values of quartile splits. A one-way repeated measures ANOVA with the different quartiles as a within-subject factor, similar to the analysis in the previous chapter, showed that the proportions of correct predictions are not significantly affected by the position of the quartile ($F(3, 72) = 1.098$, $p = .356$), indicating that the four sequences predicted choices equally well (see Figure 3.7).

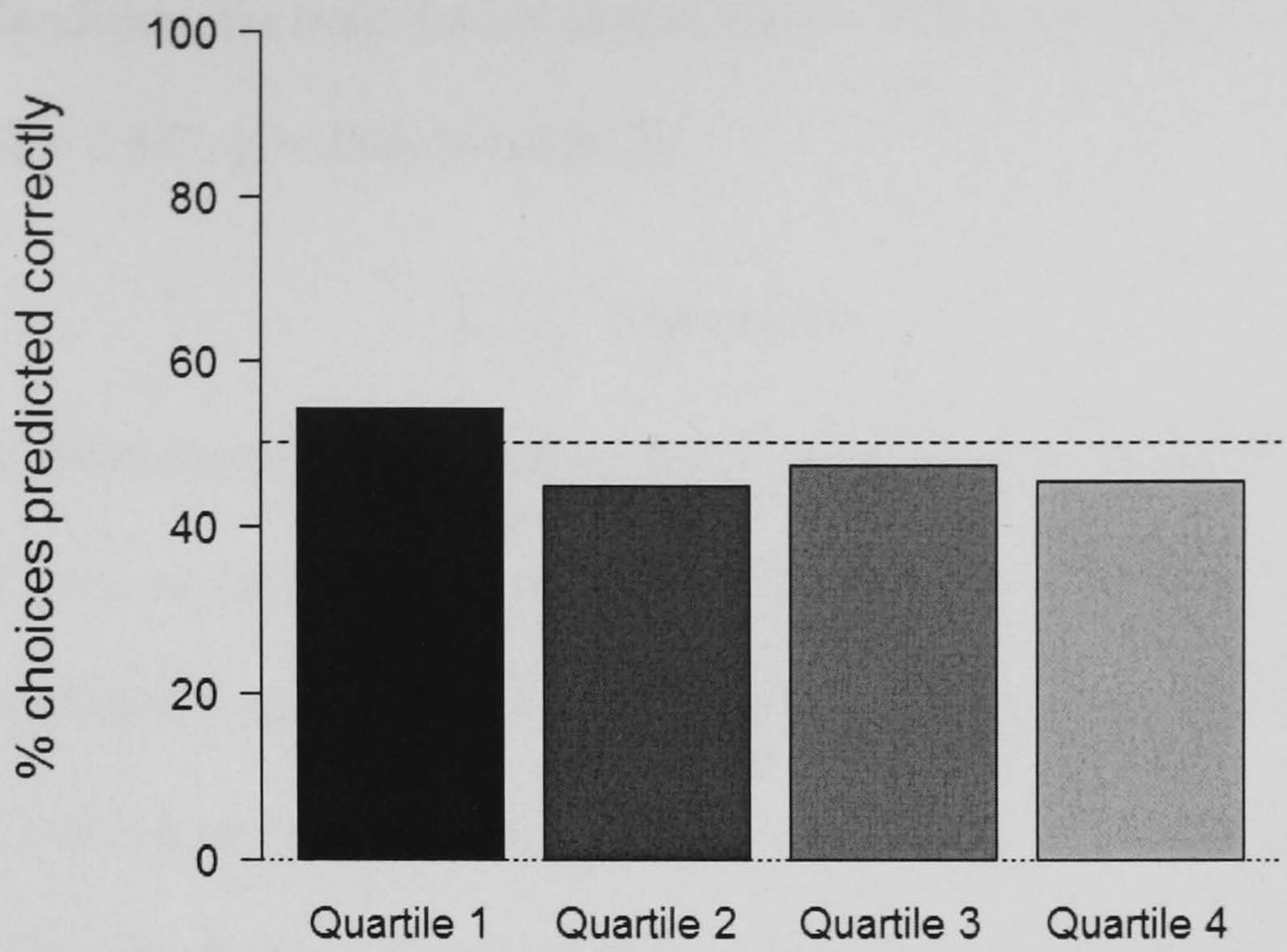


Figure 3.7. Mean percentages of correct predictions for the different quartiles of the Matched-Sampling Condition in Experiment 2.

3.2.2.5 Application of descriptive choice models

To investigate the extent to which established models can account for the choice behaviour found in this experiment, I again calculated the rate of correct predictions that can be obtained when using either EV maximisation or the PT model. In the Free-Sampling Condition the EV model based on the experienced probabilities predicted 72% of the choices correctly. For the PT model with parameters taken from Tversky and Kahneman (1992) this rate of correct predictions dropped to 60%. In the Matched-Sampling Condition both models performed at below chance with 39% for the EV model and 43% for PT. The difference between the predictive power of the EV and PT model was significant under Free-Sampling $t(24) = 3.304$, $p = .003$, two-sided) but not in the Matched-Sampling Condition $t(24) = -0.824$, $p = .418$, two-sided). When comparing between conditions the higher predictive power was always found in the Free-

Sampling Condition, for both the EV model $t(48) = 5.216$, $p < .001$, two-sided) and PT $t(48) = 2.887$, $p = .006$, two-sided).

3.2.3 Discussion

In summary, Experiment 2 demonstrates the robustness of the underweighting under DfXP even when there is no sampling error involved. The Free-Sampling Condition provided evidence that the original result is stable and can be replicated, both in terms of the sampling behaviour and the resulting choice behaviour. The switch from Free-Sampling to Matched-Sampling reduced the amount of underweighting, but did not eliminate underweighting: Reducing sampling error seems to have a moderating effect but is not sufficient to explain the choice phenomenon. Similar to the findings in the previous chapter, the effect was found without any indication of recency weighting. Instead, the percentages of correct predictions were found to be similar for the different sequence splits.

Furthermore, there was a significant difference between the predictive power of the expected value and prospect theory models. The superiority of the simpler EV model indicates that the addition of a weighting function with parameters estimated under descriptive choice, describing overweighting of small probabilities, is less suited to describe the probability weighting under DfXP than the linear weighting that is incorporated in the simpler EV model. The difference in model fit between the two experiential conditions is slightly puzzling. One possible explanation could be the use of different strategies when exploring the options in the two conditions. This issue and the question of whether the performance of the PT model can be increased with a different set of parameters will be picked up again in Chapter 6.

Taken together, findings from Experiment 2 suggest that all of the explanations put forward by Hertwig et al. (2004, 2006) to explain the original results, the reliance on small samples and recency weighting, are not sufficient. This also applies to the sampling error Hypothesis by Fox and Hadar (2006). Their deconstruction of decisions from experience described earlier provides one last alternative explanation though: People could systematically misjudge the probabilities from their sample. However, this possibility cannot be ruled out by the previous experiment, because people's judgements of the probabilities that they experienced were not collected. The assessment of judgement error under Matched Sampling will therefore be the focus of the following experiment.

3.3 Matched Sampling with Frequency Estimations (Experiment 3)

In order to investigate the potential impact of judgement error I extended the Matched Sampling design with a judgement task that captured the differences between the experienced probability of an outcome and its subjective probability. Due to the nature of the sampling task, providing counts for the occurrences of various outcomes, a frequency estimation task was selected to assess the participant's representation of the experienced variability of the outcomes. As I pointed out in the introduction on probability and frequency judgements, the results usually reported in similar tasks indicate that people are quite accurate with a slight tendency to overestimate low-frequency events. The only way in which judgement error could account for the underweighting of rare events, though, would be the opposite pattern, a systematic underestimation of the occurrence of rare events. Fox and Hadar (2006) failed to find such a pattern in the context of probability judgements under Free Sampling. Yet, if judgement error is

responsible for the underweighting of rare events under Matched Sampling, one should expect systematic underestimation of low frequencies

3.3.1 Methods

3.3.1.1 Participants

The design was implemented in the form of a Web-based experiment which was completed by a total of 197 participants, consisting of 94 men and 103 women, aged between 13 and 63 years with a mean of 28 years. The participants were recruited through different portals advertising psychological experiments on the Internet provided by the University of Warwick, the Hanover College and the University of Central Lancashire.

3.3.1.2 Design and procedure

For this experiment only a Matched-Sampling Condition was employed using a design similar to the Matched-Sampling Condition in Experiment 2, with 40 outcomes per button matching the underlying probabilities, sampling without replacement and arbitrary sampling order. Only after both buttons had been sampled 40 times could the participants proceed to the decision phase to select the option they would like to play once for real. The judgement task was added after the decision phase to avoid any impact on the final choice. Estimating the frequency beforehand could change the representation of the problem, making it more similar to a descriptive choice task. Participants estimated the number of times they had seen the rare event for both of the options. For the option offering 4 points with a probability of 80%, for example, they had to estimate the number of times they encountered the outcome '0' within the 40 outcomes sampled from this option (which would have actually been eight times). Also, to prevent

participants from using a counting strategy in subsequent problems, each participant only received one of the six choice problems previously used. Both the choice problem and the assignment to the two buttons were randomised. Again, subjects were instructed to maximise their score which was determined by a random draw from the underlying distribution of the chosen option. The experiment ended with the presentation of feedback on the number of points received from the chosen lottery, the participant's frequency estimates for the rare events, and their true frequencies.

Instead of running the experiment in the laboratory from a server it was made accessible through the World Wide Web. This allowed participants to run the experiment on their own computers. By using the same Adobe Flash technology the layout of the experiment was identical to the one in the previous laboratory experiment. Only the instructions had to be changed slightly to ensure the self-contained program execution of the Web-based format.

3.3.2 Results

3.3.2.1 Information search

Due to the design, there was no difference between the objective and experienced probabilities as the latter were matched to the former. The number of samples drawn from each button was fixed to 40 for all participants. The order in which samples were drawn from the two options though did differ between participants. In terms of the exploration strategies used the majority of the participants (90%) seemed to switch several times between the available options. Nine percent of the participants switched only once which means that all 40 samples from one button were sampled exclusively before starting the exploration of the remaining button. Continuous alternation between the two buttons was only observed in 1% of the participants. The median number of switches was 16 ($M = 18.74$) and the median switch ratio was 0.41 ($M = 0.48$) which indicates that there was more swapping between options than in the previous Matched Sampling Condition. The reasons for this could be the fact that there was only one choice problem which makes the task less tedious. A histogram with the complete distribution of switches is provided in Figure 3.8.

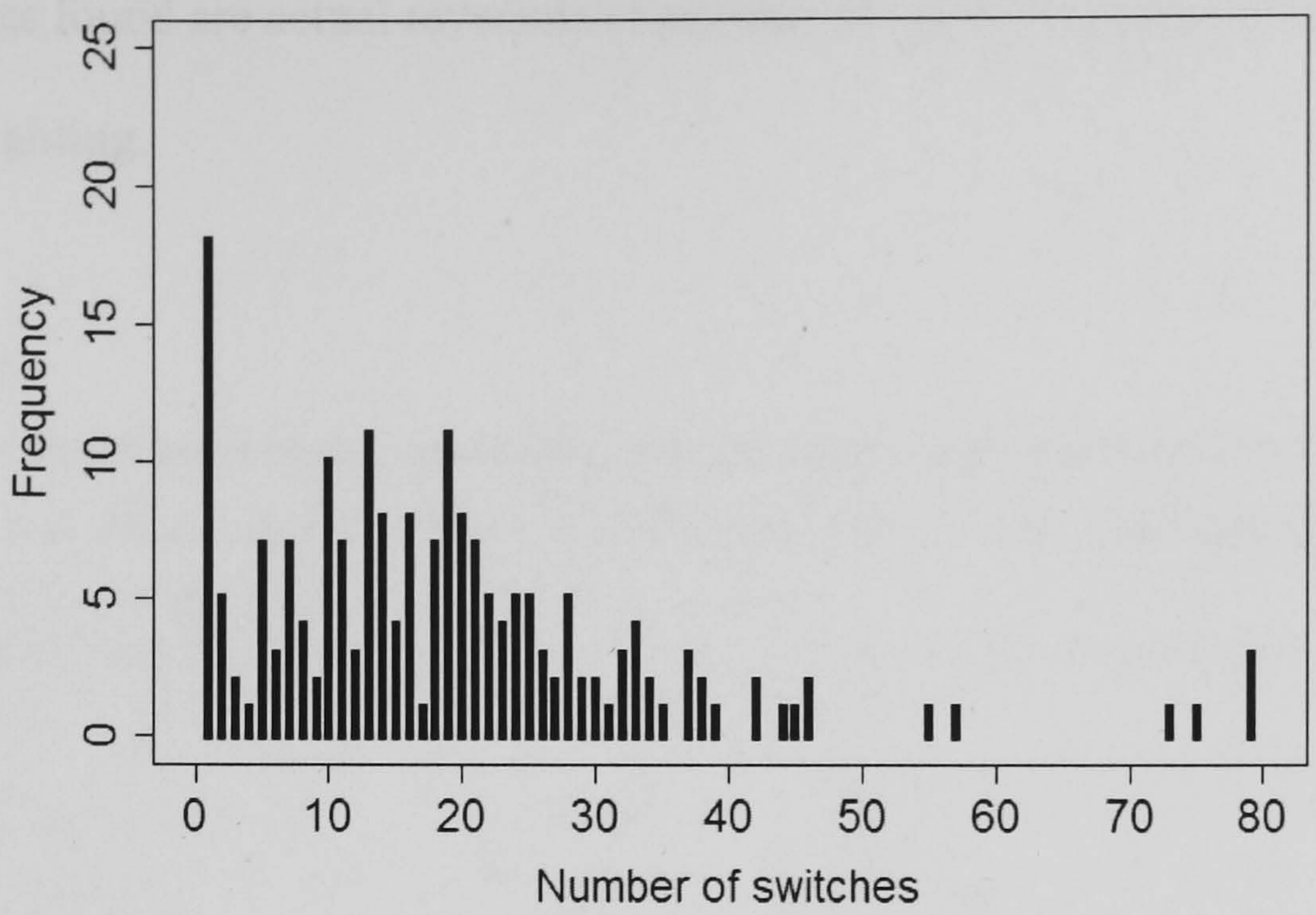


Figure 3.8. Histogram with the number of switches observed in Experiment 3.

3.3.2.2 Choice proportions

The observed proportions of choices in the direction of overweighting of small probabilities in the Matched-Sampling Condition from Experiment 3 were compared with the proportions in the Description Condition from the previous experiment. Although sampling error was eliminated, there were still a significantly smaller proportions of choices in the direction of overweighting under Matched Sampling, $t(220) = 4.22$, $p < .0001$. However, with a mean absolute difference of 22.67 percentage points between the two sets of proportions, the differences were less extreme again than the ones observed under Free Sampling (Experiment 1, 2 and Hertwig et al.'s data).

Significant differences were also observed within the proportions of H choices in the individual choice problems. Both sets of proportions are shown in Table 3.2, together with the *p*-values of the Fisher exact tests, and the number of participants behind each cell in Experiment 3. The fact that five out of the six pairs of proportions lie on opposite sides of the 50% line indicates that the

differences found are actual reversals in preferences from overweighting to underweighting.

TABLE 3.2

Summary of the observed choice proportions under Matched Sampling and descriptive choice (Experiment 2). The p-values have been calculated using Fisher's exact tests. Significant differences between the two are highlighted with asterisks.

Decision Problem	H	L	Percentage choosing H			
			Description (Experiment 2)	Matched Sampling (Experiments 3)	*p*	n
1	4, .8	3, 1.0	36	68 *	.030	31
2	4, .2	3, .25	72	39 *	.017	31
3	-3, 1.0	-32, .1	64	42	.116	31
4	-3, 1.0	-4, .8	36	55	.198	38
5	32, .1	3, 1.0	48	45	1.00	31
6	32, .025	3, .25	52	26	.056	35

This point is also illustrated in Figure 3.9 which depicts the converted differences between experiential and descriptive choice proportions. All the proportions consistently deviate from the ones in descriptive choice tasks in the direction of less overweighting of small probabilities.

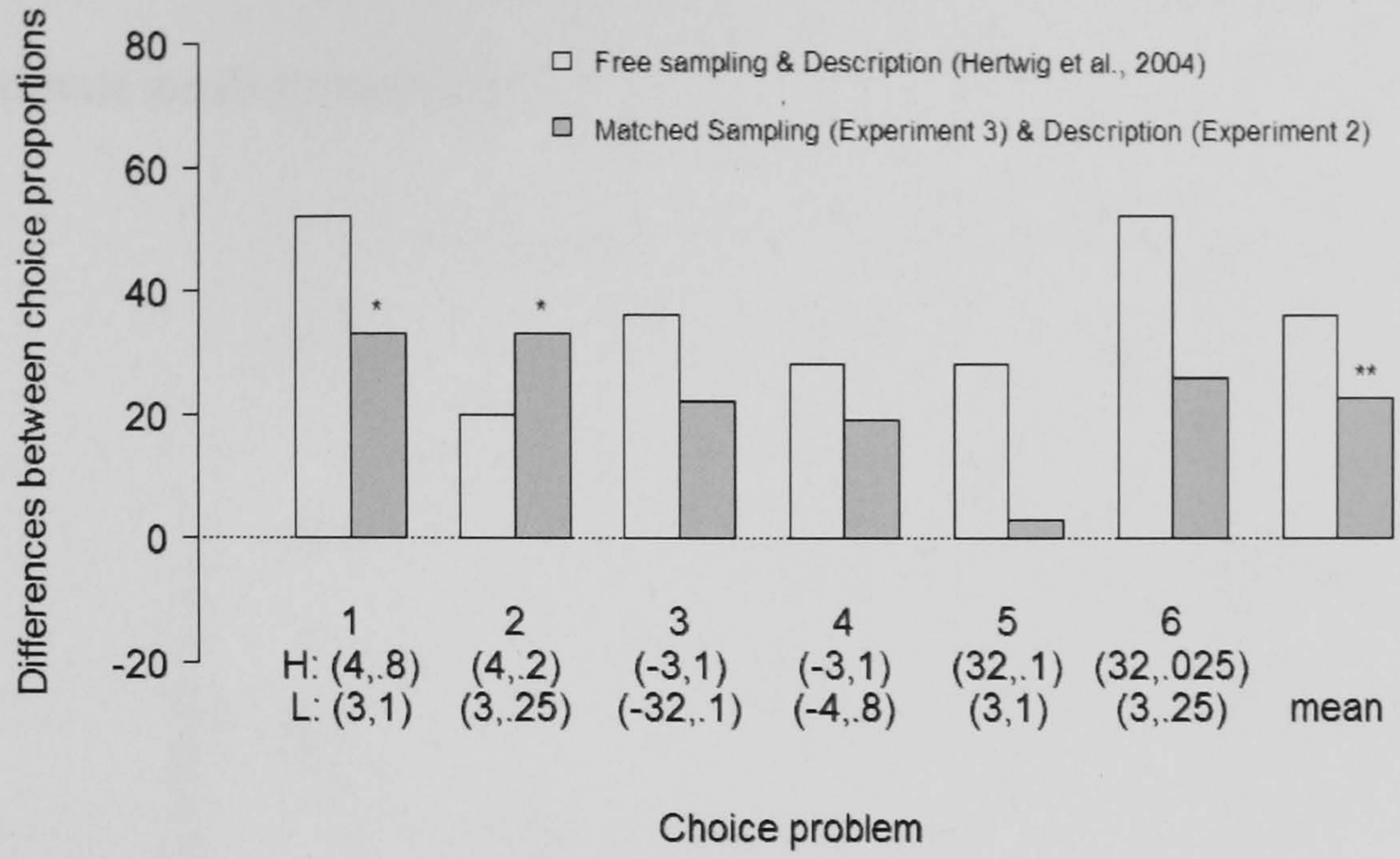


Figure 3.9. Differences in choices proportions between the Matched-Sampling Condition in Experiment 3 and the Description Condition in Experiment 2 across the six decision problems. Positive bars indicate choices in the direction of less overweighting. Significant differences are marked by asterisks (* $p < .05$, ** $p < .01$). The mean bars on the right correspond to the t-test results provided in the text. The results for the data reported by Hertwig et al. (2004) have been added for comparison.

3.3.2.3 Frequency judgements

In order to examine the frequency data provided by the participants, the mean absolute differences between the actual experienced frequencies of the rare events and their estimates were calculated and extreme values (more than 3 standard deviations away from the mean) were reiteratively removed. Out of the 384 judgements 22 had to be excluded. The majority of these extreme outliers did actually match with the frequency of the common event and might have been the result of misreading the instructions. The judgements observed here were well-calibrated with a high correlation between judged and actual frequencies, $r(370) = .98$, $p < .0001$. The mean absolute difference was 1.57 ($SD = 2.37$). Figure 3.10 shows a scatter plot of the estimated frequencies across the five frequencies

together with their estimation means. There was no obvious tendency to underestimate small frequencies.

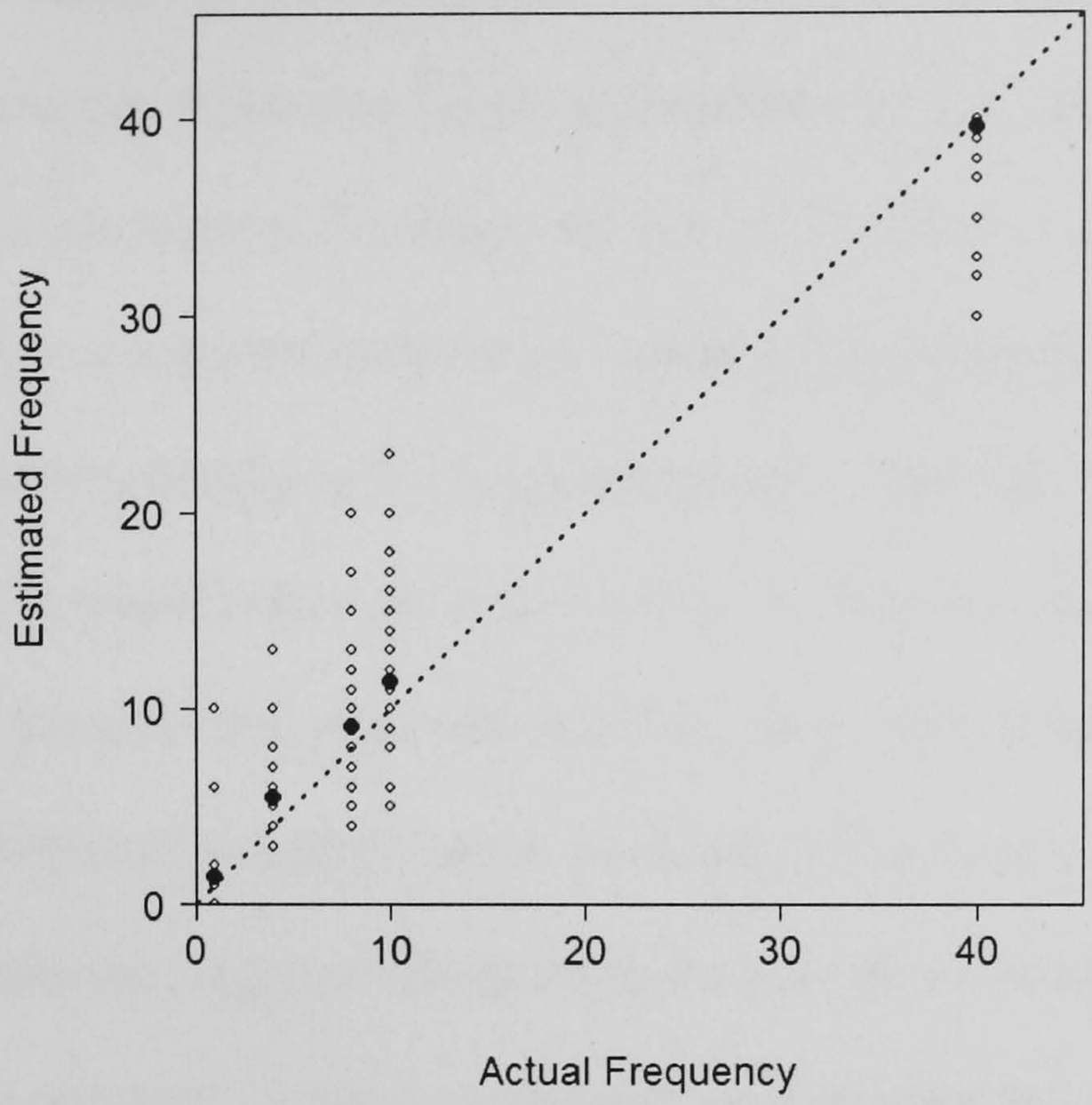


Figure 3.10. Deviations of the frequency judgements for the rare events in Experiment 3 plotted against the actually experienced frequencies. Due to the overlap of the probabilities in the six choice problems used there were only 5 different rare event frequencies. The dotted line indicates perfect calibration. The black dots are the mean estimates. The white dots indicate the observed estimation errors. One white dot may represent several data points from different participants.

Instead, an examination of the means of the estimation errors showed small deviations in the direction of overestimation of low frequencies and underestimation of high frequencies (from low to high frequency: $t(34) = 1.46$, $p = 0.154$; $t(61) = 5.16$, $p < 0.001$; $t(90) = 3.18$, $p = 0.002$; $t(64) = 2.71$, $p = 0.009$; $t(118) = 2.89$, $p = 0.005$, all two-sided).

3.3.2.4 Recency Weighting

Recency weighting in the form of a higher rate of correct predictions of the last half of the sampling sequences was again not found, 51% and 47% respectively,

χ^2 McNemar (1) = 0.006, p = 0.938. As in Experiment 2, I also compared the percentages of correct predictions on the basis of the expected values of the four quartiles of each option (from 1 to 4: 50%, 51%, 53%, and 50%). Again, no evidence for a recency effect was found, as there were no significant differences between the four percentages, Cochran's Q (3, 197) = .404, p = .909.

In addition, a separate analysis was conducted focusing on the data from decision problem 6, which provides a special property. The high H option in this problem offers 32 points with a probability of 2.5 %. Matched onto the sequence of 40 trials this payoff is therefore only observed once. Testing whether maximisation within this problem can be predicted by the location of the rare but high payoff within the sequence also provides an assessment of recency weighting. If an outcome's impact on choice is an increasing function of its position within the experienced sequence then recency weighting of such a form should be picked up by this analysis. The maximisation rate of participants who have seen the outcome in the first or second half (22% vs. 29%) did not differ significantly though, Fisher's Exact p = .711. Similarly, experiencing it in the first 7 of last 7 samples did not have a significant impact on maximisation either (20% vs. 44%, Fisher's Exact p = .301).

3.3.2.5 Application of descriptive choice models

On the basis of the experienced probabilities, the EV and PT model (using the same set of parameters as used to model the data in Experiment 1 and 2) did not predict choices better than chance, accounting for 46% and 38% of choices respectively. With the availability of the frequency judgements, it was also possible to test the predictions of the two-stage model (Fox & Tversky, 1998; Tversky & Fox, 1995). Recall, from Section 1.3, that the two-stage model

assumes a transformation similar to PT, but on the basis of judged probabilities, in this case generated from the frequency estimates. Following Fox and Hadar (2006), I used the PT parameters reported by Tversky and Kahneman (1992). By using the judged probabilities to predict participant's choices, this model also provides an indirect assessment of the involvement of judgement error in the observed choice pattern. If the performance of the model is substantially enhanced when using the subjective probabilities instead of the objective ones, this would be an indication for the mediation of the effect through judgement error. However, the actual model fits remained below chance performance of 50% (see Figure 3.11).

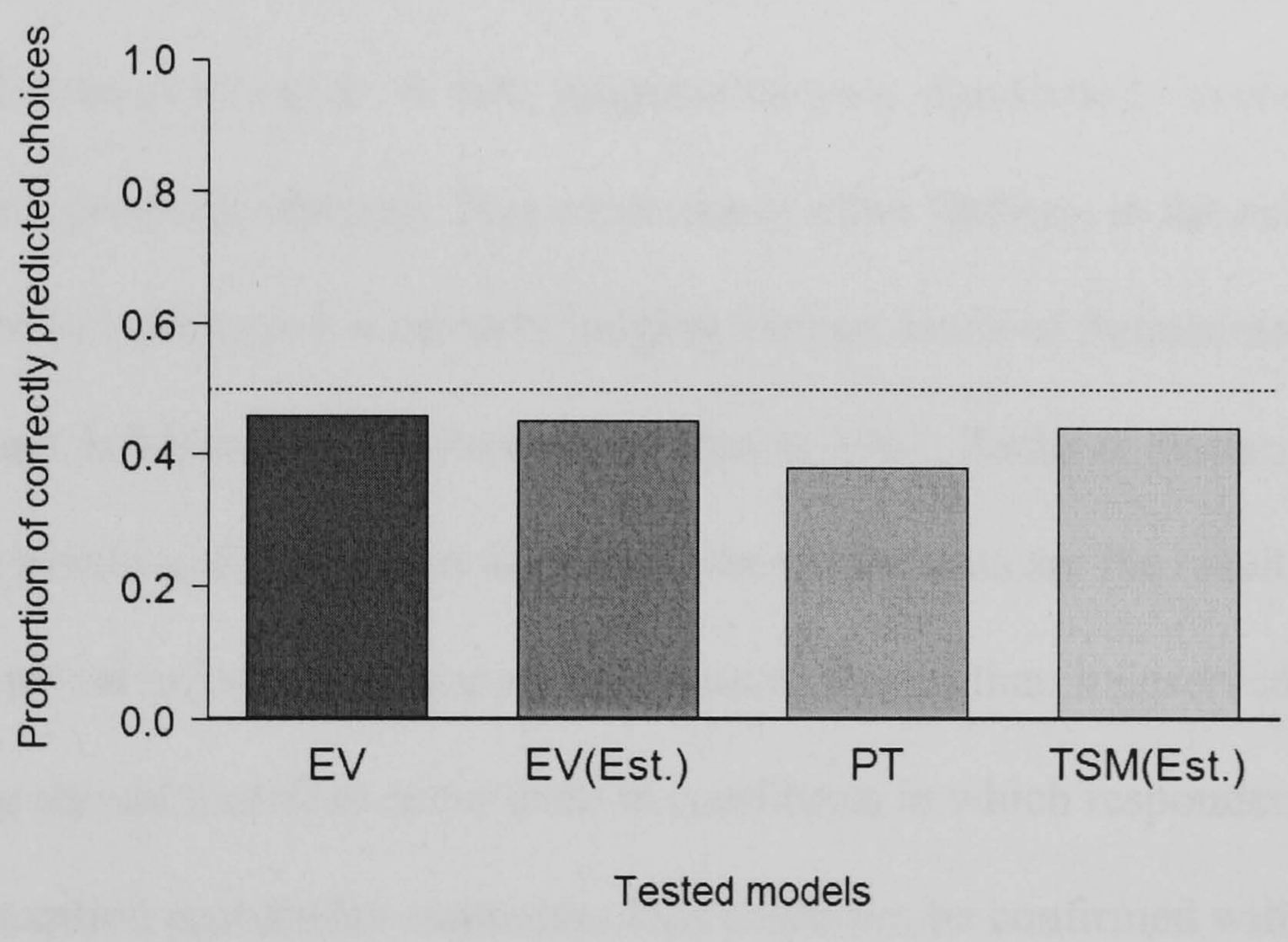


Figure 3.11. Percentage of correct predictions for the different choice models: expected value, expected value based on judged probabilities, prospect theory and the two-stage model (also based on judged probabilities). The latter two the fits were calculated on the basis of the Tversky and Kahneman (1992) parameters.

The fit of the EV model on the basis of the judged probabilities was very similar with 45% correct predictions (χ^2 McNemar (1) = 0, p = 1.00). The two-stage

model predicted 44% of the observed choices correctly, which is a small but significant improvement compared to the performance of the PT model presented above (χ^2 McNemar (1) = 5.263, p = .022). However, overall, the best fit is provided by the EV model based on the objective probabilities.

3.3.3 Discussion

In summary, the results of this second Matched-Sampling Experiment provide further support for the existence of deviations in the direction of underweighting of small probabilities in decision from experience under conditions where statistical sampling error is eliminated. Moreover, in the light of well-adjusted frequency estimations (with a slight tendency to overweight small frequencies), judgement error in the form of underestimation of small frequencies can also be excluded as an explanation. In fact, judgements were significantly overestimated, in line with previous research. This conforms to other findings in the people are quite good in storing and accurately judging various kinds of frequencies (Gigerenzer & Murray, 1987; Peterson & Beach, 1967; Zacks & Hasher, 2002).

According to Hertwig et al. (2006), recency effects are the result of giving the most recent information proportionally more weight than it deserved. Recency weighting should therefore occur even in conditions in which respondents have accurate explicit probability estimates. This could not be confirmed within this experiment as there was again no evidence for recency weighting under Matched Sampling within the various tests of recency weighting employed.

With regard to the application of established choice models, I again found that rates of correct predictions that did not exceed chance level. The results of the comparison of models based on the actually experienced frequencies within the sample and models using the judged frequencies as input are also interesting.

Even the usage of subjective probability measures did not help any of the models to lift the predictive power above chance level. Given that the judgements were more or less unbiased this is not surprising as their inclusion does not add any information that could help increase the performance of the models. This can therefore be seen as additional evidence against the judgement error hypothesis. Further claims regarding the validity of the two-stage model cannot be made at this point, but will be discussed following a more detailed model comparison in Chapter 6.

Furthermore, as Experiment 3 was conducted over the Web, this study also shows that the effect is robust enough to be replicated in a more general population with demographics that go beyond that of undergraduate students. At the same time, the wider range of risk attitudes within the Web-sample might have been the reason for the reduction of the effect. A potential issue could be the comparison of the Matched Sampling data, which was collected in a Web-based environment, with the description data from Experiment 2, which was collected in a laboratory setting. It could thus be argued that the differences in choice proportions are due to differences in the demographic characteristics of the two samples. Nevertheless, as we have already replicated the effect within one sample of subjects (undergraduate students in Experiment 2) this seems less problematic but I will return to this issue in Chapter 4.

The use of frequency judgements instead of probability judgements could also be a source of potential criticism. However, I have justified this on the basis of findings from the judgement literature which seems to indicate that the accuracy of both formats is similar and that one can be easily transformed into the other. Einhorn and Hogarth (1978) mention that such a transformation requires the

incorporation of both occurrences and non-occurrences of an event. In the context of dichotomous outcomes within sequences this should not constitute a problem as non-occurrence can be derived without difficulty by subtracting the number of occurrences from the total number of samples. Nevertheless, it remains unclear whether participants actually translate estimated frequencies directly into probabilities. Even if participants do, there is still the possibility of a potential bias during the translation process. As a consequence, even in the case of well adjusted frequency judgements, participants' subjective probabilities could still be biased and therefore impact the decision making process in a systematic way. This extreme case could also explain the poor performance of the two-stage model on the data based on frequency estimations presented above. Under such circumstances the frequency estimations method might not be able to assess the actual judgement bias that is linked to the choice behaviour on the appropriate level. Another issue is the focus on frequencies of the rare events only, which, in the context of the six choice problems used, leaves us with only a small range of five probabilities to assess judgement error. An alternative method would be to ask for estimates of both the rare and the common event. Fox and Hadar (2006) have essentially done this by asking participants to provide probability judgements for all the encountered outcomes. However, this comes with the downside of forcing subjects into a representation of the events and their probabilities which conforms with probability theory (e.g. the fact that probabilities must add up to one), and might not be an accurate description of the people's subjective probabilities either. Asking for all possible outcomes could undermine the existence of super- or subadditivity as it has already been found in the domain of uncertainty (e.g. Fox & Tversky, 1998; Tversky & Fox, 1995;

Tversky & Koehler, 1994). All these issues will be addressed in the last experiment of this chapter which will look into probability judgements in the context of the Matched Sampling paradigm.

3.4 Replication with probability judgements (Experiment 4)

In the previous experiment, I demonstrated that underweighting of small probabilities can be observed under Matched Sampling without any evidence for biased frequency judgements. This experiment will investigate whether this can be generalised to probability judgements. In addition, I will test whether the representation of uncertainty information attached to an outcome can be altered by the order of the judgement task and whether this has an indirect effect on choice. In the last experiment, participants were only asked to provide judgements after they had made their decision to explicitly avoid a potential switch in representation from experiential to descriptive choice. Fox and Hadar (2006) seem to have collected data with probability judgements before and after choice but do not report any analysis on their equivalence. The following experiment will investigate whether recalling the experienced probabilities before choice facilitates a representation of the problem more similar to a descriptive choice task and whether this is also reflected in the observed choice pattern.

3.4.1 Methods

3.4.1.1 Participants

Like in the Matched Sampling design of the last section, the experiment was implemented as a Web-based experiment using Adobe Flash. This time the advertisement of the experiment and payment of the participants was organised

through the "ipoints" reward scheme (www.ipoints.co.uk). All ipoints collected in this scheme can be exchanged for CDs, flights and other goods. In this way ipoints is able to maintain a fairly large database with a good spread across a wide range of demographic variables. For this particular experiment, participants received a reward of 50 ipoints (worth £ .5) for taking part. Datasets from 360 participants were collected. Their age ranged from 13 to 80 years with an average age of 42 years. Around two thirds (245) were female.

3.4.1.2 Design and procedure

The main structure of the experiment was the same as in the previous Matched-Sampling Condition with a learning phase, a choice task and a judgement task. Again, the same six gambles were used to compare the results with previous findings. Two independent variables – 'order of the judgement task' (before or after choice) and 'type of event' (common or rare event) – were systematically varied in the form of a 2x2 between-subjects design with four groups of 90 subjects each (see Figure 3.12). The second variable was mainly introduced to collect judgements for a wider range of probability values providing a more accurate assessment of potential judgement bias. The allocation to one of the four experimental conditions was randomised. During the first stage the participants had to sample 40 outcomes from the two options in whatever order they liked. In the decision phase they then had to choose the option they preferred to play once at the end of the experiment.

		Order of the judgement task	
		before choice	after choice
Type of event'	common	common events before choice	common events after choice
	Rare	rare events before choice	rare events after choice

Figure 3.12. 2x2 design with the factors 'order of the judgement task' and 'type of event'

One of the differences to the previous design was that, depending on the experimental condition, the judgement task had to be completed before or after the decision phase. In addition, instead of frequency estimations participants were asked to provide probability judgements for the rare or common events (again depending on the allocated condition), separately for both options previously sampled. The exact wording of the question was the following: "How likely do you think it is to receive the outcome x when pressing button y? Please provide a probability (0-100) in the window below the button and confirm by pressing the 'NEXT' button". For the risky option in choice problem 1, for example, the probability of the rare event ('0') would have been 20% and the probability of the common event '4' would have been 80%. As in the previous experiment, each participant only received one of the six choice problems to prevent counting in subsequent problems. Both the choice problem and the assignment of the two options to the buttons 'A' and 'B' were randomised. All participants were instructed to maximise the number of points they received, which would be determined by a random draw from the chosen option. The feedback presented at the end of the experiment contained the outcome of the chosen lottery, their

estimated probabilities (rare or common) for the two buttons, and the actual probabilities. Finally, participants also had the opportunity to provide comments and feedback regarding the experiment in a text box. Overall, the completion of the whole experiment took only between 3 to 5 minutes.

3.4.2 Results

3.4.2.1 Information search

Switching between options was even more prominent than in the previous design using only one choice problem. Only a very small proportion of participants did not switch more than once while exploring the options (4%). However, continuous alternation between options was also rare with less than 1% of the cases. The median number of switches was overall higher with 22 ($M = 25.1$). The median switch ratio was 0.28 ($M = .32$). The overall distribution of switches is shown in Figure 3.13.

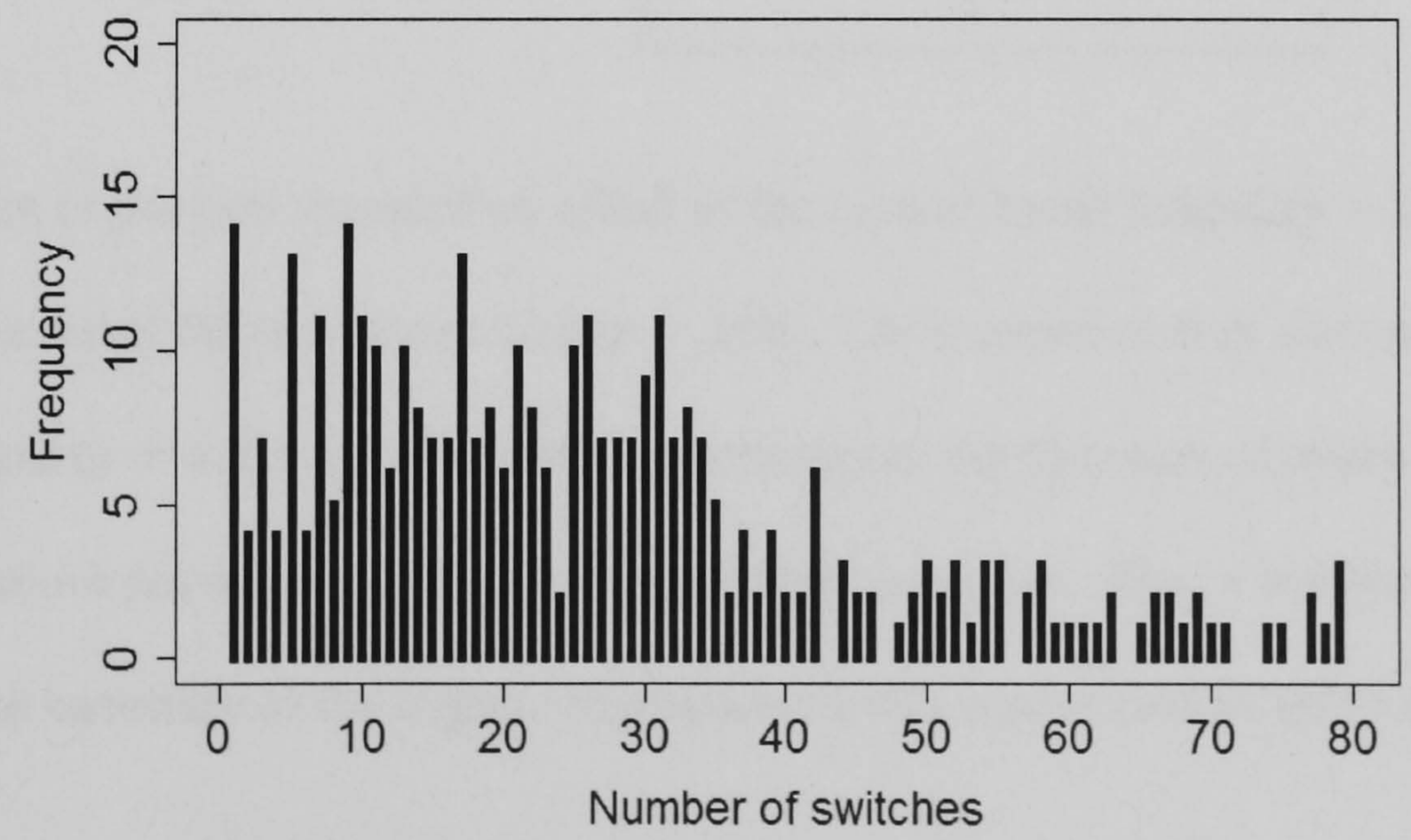


Figure 3.13. Histogram with the number of switches observed in Experiment 4.

3.4.2.2 Effect of the experimental variables

In order to assess whether the experimental variables affected the decision-experience gap observed in previous experiments, I calculated the mean proportions of choices in the direction of overweighting of small probabilities across all six choice problems for the four experimental groups. A first inspection of these proportions in Table 3.3 shows that proportions are quite similar within all of the groups and that overweighting is observed in less than 50% of the choices.

TABLE 3.3

Mean proportions of choices in the direction of overweighting across the four experimental groups

		Order of the judgement task	
		before choice	after choice
Type of event'	Common	.35	.43
	Rare	.41	.46

A logistic regression revealed no effect of the type of event judged (p = .653), or for the order of the judgement task (p = .286). The interaction was also not significant (p = .815). The proportion of choices in the direction of overweighting did therefore not depend upon which event was judged or when it was judged. A complete summary of the logistic regression results is provided in Table 3.4.

TABLE 3.4

Logistic regression results for the effects of order of judgement task and type of event

	Estimate	Std. Error	*Z*	*p*
Model 1				
(Intercept)	-.268	.211	-1.261	.207
Order of the judgement task	-.326	.301	-1.066	.286
Common or rare events	.135	.298	.449	.653
Order of the judgement task × Common and rare events	.101	.429	.234	.815

Similar results were also obtained when comparing the proportions of H choices for participants judging the probabilities before choice and after choice within the individual choice problems. No significant difference was found in any of the six problems (see Table 3.5).

TABLE 3.5

Percentages of H choices for the two judgement order conditions across the six choice problems

	Percentage choosing H		
Decision Problem	Before choice	After choice	*p*-value (2-Tail) Fisher's exact test
1	47	60	.437
2	50	63	.435
3	17	33	.233
4	50	53	1.00
5	20	47	.054
6	40	40	1.00

As there were no significant differences in terms of the inherent overweighting of small probabilities the data of the four experimental groups was combined for the subsequent analyses.

3.4.2.3 Comparison with DfD choice proportions

In order to test whether the actual decision-experience gap could be replicated for the Matched Sampling data in Experiment 4, the obtained choice proportions were compared with the descriptive choice proportions from Experiment 2. The mean absolute difference between the proportions within the two formats was 19, which is again smaller then the differences observed under Free Sampling in the earlier experiments. However, the comparisons of the overall mean choice proportions in the direction of overweighting reveal that there are actual differences between the two formats in the direction of less overweighting under DfXP, $t(383) = 3.934$, $p < 0.001$). This is also partly evident within the six choice problems (see Table 3.6).

TABLE 3.6

Percentages of H choices for the combined Matched-Sampling Conditions from Experiment 4 and the descriptive choice proportions reported in Experiment 2.

	Percentage choosing H		
Decision Problem	Matched Sampling (Experiment 4 combined)	Descriptive Choice (Experiment 2)	*p*-value (2-Tail) Fisher's exact test
1	53	36	.161
2	47	72	.227
3	20	64	.001
4	27	36	.236
5	20	48	.227
6	40	52	.344

As the proportions in four out of the six choice problems lie on opposite sides of the 50% line this can be interpreted as a transition from overweighting to underweighting. Significant differences though were only found in choice problem 3. Figure 3.14 demonstrates this once more, showing the net differences between of the converted choice proportions of the DfD data (Experiment 2) and

the combined Matched Sampling data. Across all six problems and their mean, the bars point consistently in the direction of less overweighting of small probabilities under DfXP. Again the differences are smaller than the ones observed under Free Sampling, replicating the finding from Experiment 2 and 3, that Matched Sampling is attenuating the effect but not eliminating it.

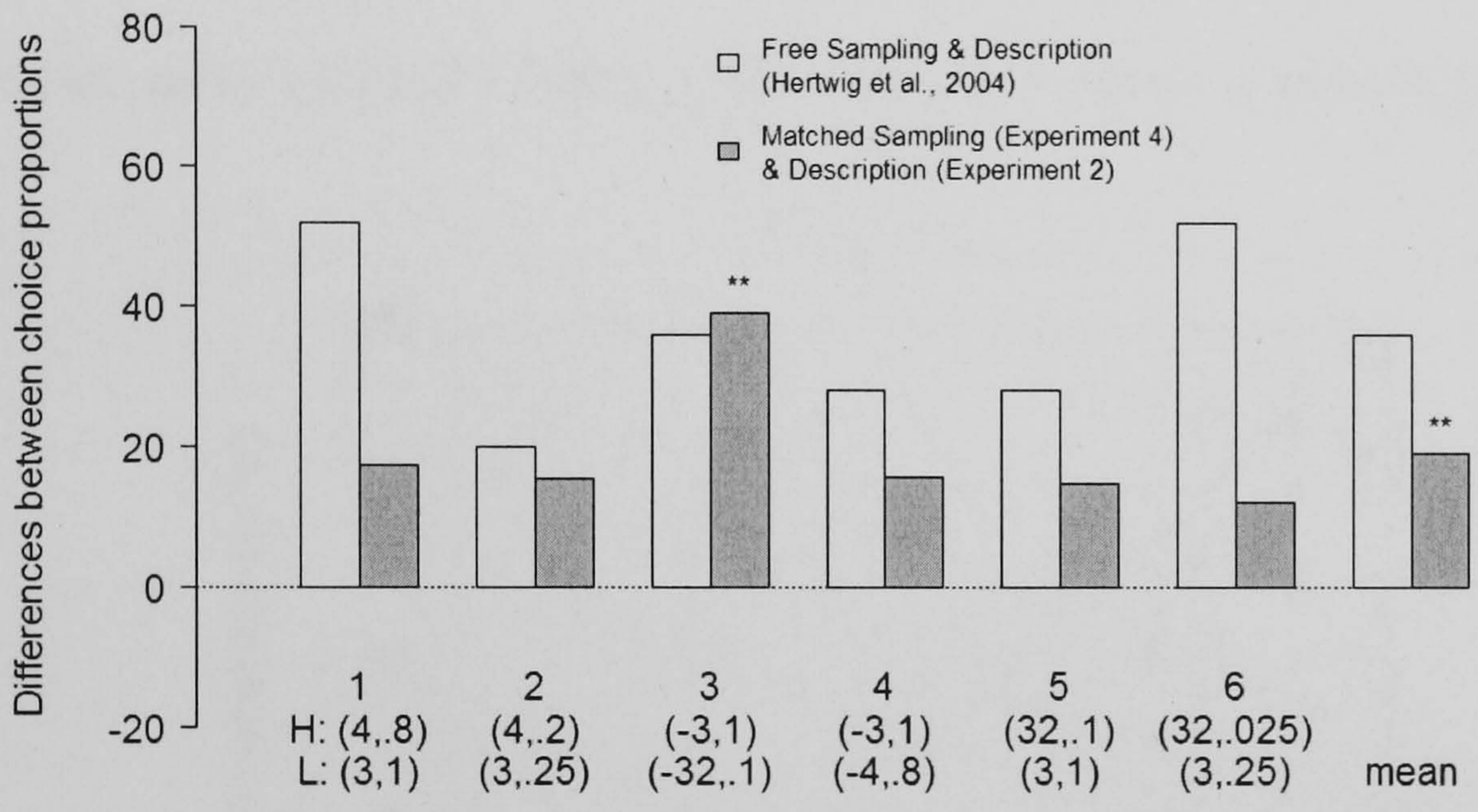


Figure 3.14. Differences in choice proportions between the combined Matched Sampling data from Experiment 4 and the DfD data from Experiment 2 (grey bars). Positive bars indicate differences in the direction of less overweighting. The differences for the proportions reported by Hertwig et al (2004) are included for comparison. The two asterisks for the mean relates to the t-test results presented above.

3.4.2.4 Probability judgements

Before analysing the data, extreme outliers (more than 3 standard deviations away from the mean) were successively removed, which applied to 81 out of the 720 judgements. Across the remaining data, probability judgements were again well adjusted with a correlation between the provided judgements and the actual probabilities of $r = .92$, $p < .0001$ and a mean absolute difference of 10.12 (SD =

12.07). From the scatter plot in Figure 3.15, it is evident that there is again no systematic underestimation of small probabilities. Rather, the mean estimation errors implied significant deviations in the direction of overestimation for low probabilities (from 0 to 25%: $t(93) = 5.09$, $p < 0.001$; $t(20) = 2.45$, $p = 0.023$; $t(45) = 2.37$, $p = 0.022$; $t(71) = 2.14$, $p = 0.035$; $t(45) = 1.55$, $p = 0.128$, all two-sided) and significant deviations in the direction of underestimation of high probabilities (from 75 to 100%: $t(59) = 3.43$, $p = 0.001$; $t(89) = 7.61$, $p < 0.001$; $t(59) = 4.77$, $p < 0.001$; $t(29) = 2.28$, $p = 0.030$; $t(119) = 6.71$, $p < 0.001$, all two-sided).

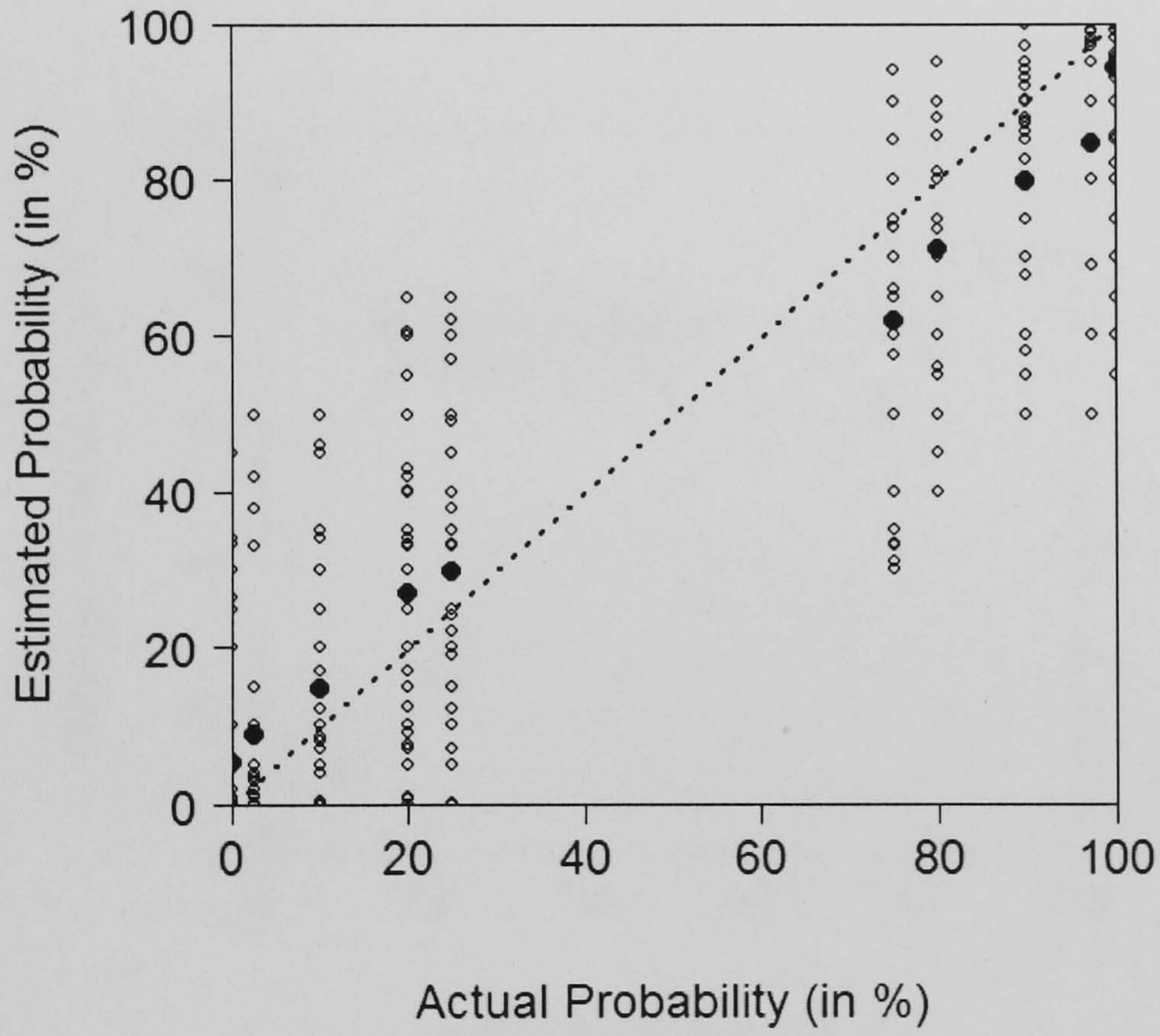


Figure 3.15. The deviations of the probability judgements for both rare and common events in Experiment 4 plotted against the actually experienced probabilities. The dotted line indicates perfect calibration. The black dots are the mean estimates. The white dots indicate the observed estimation errors. One white dot may represent several data points from different participants.

When analysing the probability judgements separately for the groups of participants providing their judgements either before or after choice task, the

Before-Choice Condition was found to provide more accurate judgements with a smaller mean absolute difference ($M_{\text{before}} = 8.54$, $SD = 10.55$, $M_{after} = 11.79$, $SD = 13.32$; $t(637) = -1.50$, $p = .13$). However, a stronger relationship between judged and actual probabilities, between the Before-Choice Condition ($r(329) = .94$, $p < .0001$) and the After-Choice Condition ($r(310) = .90$, $p < .001$) was not confirmed ($z = 3.99$, $p > .999$). Figure 3.16 gives the distribution of the means for the different probability values, separately for the two groups. It can be seen that the difference between the two conditions stems mainly from slightly higher overestimation within the small probabilities in the After-Choice Condition.

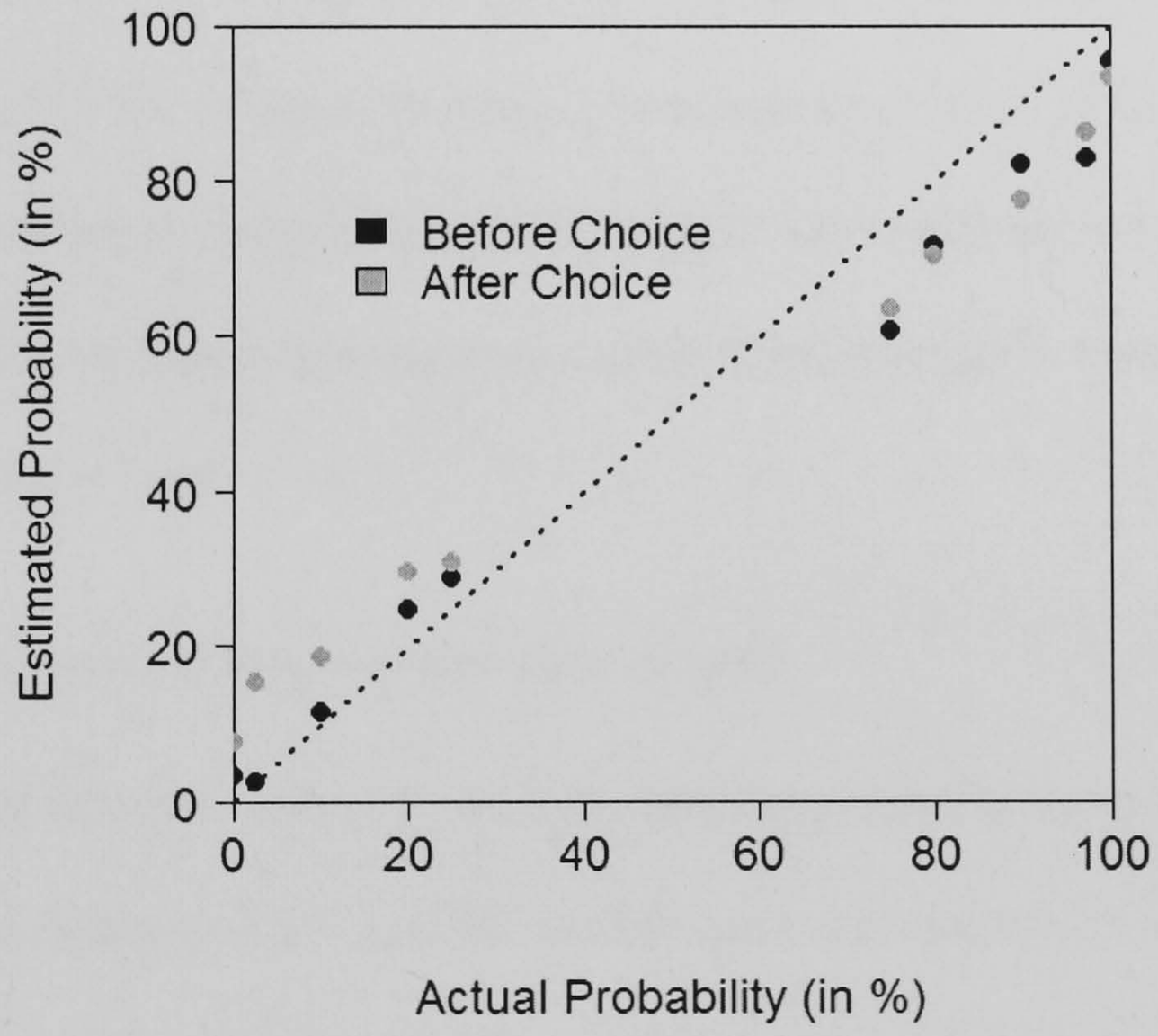


Figure 3.16. Mean deviations of the probability judgements across the actually experienced probabilities, separately for the Before-Choice and After-Choice Conditions.

A significant difference between the mean absolute differences of the participants judging rare or common events was not found ($t(637) = -1.50$, $p = .13$).

3.4.2.5 Recency Weighting

Consistent with the results from previous experiments there was no indication for recency weighting, neither between the rate of correct predictions from the first and second half of the outcome sequences (50% and 47% respectively, χ^2 McNemar (1) = 0.12, p = 0.731, nor between the rates of correct predictions from the four quartiles (from 1 to 4: 49%, 48%, 49%, and 48%), Cochran's Q (3, 360) = .199, p = .978.

Again, the position of the rare but high outcome of 32 points within the sampled sequences of Problem 6 did not seem to have an impact on the rate of maximising choices. This applies to predictions based on the first or second half of the sequence (37% versus 43% correct predictions, Fisher's Exact p = .793) and the more extreme comparison of participants encountering the rare event either in the first or last seven outcomes of the sequence (29% versus 40%, Fisher's Exact p = 1.00).

3.4.2.6 Application of descriptive choice models

On the basis of the data with two valid probability judgements per participant (N = 294), the predictions of EV and PT models were calculated. EV maximisation based on the experienced probabilities predicted 43% of the choices correctly. By using the estimated probabilities this could be increased by 7% (χ^2 McNemar (1) = 3.375, p = .066). PT based on the experienced probabilities (using the same set of parameters as in previous analyses) could account for 43% of the choices correctly. Again using the judged probabilities instead (two-stage model) could increase the percentage of correct predictions slightly (48%), but not significantly so (χ^2 McNemar (1) = 1.750, p = .186). The differences between the EV and PT

models were also not significant, neither for the models based on experienced probabilities (χ^2 McNemar (1) = .006, p = .94), nor for the models based on judged probabilities (χ^2 McNemar (1) = 0.352, p = .553). Table 3.7 provides a summary of the rate of correct predictions across all the different subgroups of the two experimental variables.

TABLE 3.7

Rate of correct predictions across the subgroups of the experimental variables in Experiment 4

	EV (experienced probabilities)	EV (estimated probabilities)	PT (experienced probabilities)	Two-stage (estimated probabilities)
All	.43 (128/294)	.50 (146/294)	.43 (127/294)	.48 (141/294)
Before	.39 (60/155)	.48 (75/155)	.42 (65/155)	.49 (76/155)
After	.49 (68/139)	.51 (71/139)	.45 (62/139)	.47 (65/139)
Rare	.41 (63/154)	.47 (73/154)	.45 (70/154)	.49 (76/154)
Common	.46 (65/140)	.52 (73/140)	.41 (57/140)	.46 (65/140)
Before & Rare	.35 (29/83)	.48 (40/83)	.47 (39/83)	.52 (43/83)
Before & Common	.43 (31/72)	.49 (35/72)	.36 (26/72)	.46 (33/72)
After & Rare	.48 (34/71)	.46 (33/71)	.44 (31/71)	.46 (33/71)
After & Common	.50 (34/68)	.56 (38/68)	.46 (31/68)	.47 (32/68)

In order to assess the impact of the two independent variables used in this experiment on the rate of correctly predicted choices four logistic regressions were conducted with the prediction of the four models presented above as a dependent variable and the two experimental factors as explanatory variables. A significant coefficient was not found in any of the four analyses. It can therefore be concluded that the proportion of correctly predicted choices does not depend on whether the probabilities were judged before or after choice or whether rare or common events had to be judged during the estimation phase.

3.4.3 Discussion

Experiment 4 has shown once more that the apparent underweighting in DfXP can still be observed under Matched Sampling, in which sampling error is eliminated.

As in the previous experiments, the differences between the proportions in descriptive and experiential choice are less extreme which seems to indicate that sampling error has only a moderating effect. The analysis of the probability judgements has confirmed the findings from the preceding experiment, providing clear evidence against a systematic underestimation of small probabilities that would have been able to account for the apparent underweighting of small probabilities. Instead, the judgements provided by the participants were again well calibrated with deviations in the opposite direction. The overall accuracy though seems to have been lower in probability judgements than in frequency judgements. The reasons for this remain unclear but it could indicate that frequency counts are not directly transformed into probability judgements, although both are following a similar pattern.

The two experimental variables employed in the last experiment did not have an impact on the extent of the apparent underweighting of small probabilities under Matched Sampling. This is not necessarily a surprise for the 'type of event' to be judged which only had the purpose to provide an extended range of probabilities. More interesting, is the observation that it does not seem to make a difference whether probability judgements are made before or after choice. The transformation of observed frequencies into probability values does not seem to result in a representation similar to descriptive choice as the resulting choice behaviour still resembles an experiential choice pattern. This is even more surprising as we know that the judged probabilities of rare events are slightly overestimated.

In terms of recency weighting, the results are again in line with the general observation that the position of the rare events within the sample does not explain

the choice pattern either. This does not necessarily allow the conclusion that the whole sequence is actually used for the decision process but it seems to show that choices are not made on the basis of average payoffs derived from parts of the whole sequence. The results from the application of descriptive choice models have been very consistent again across all the experiments presented here supporting the conclusion that neither EV nor the established form of PT provide predictions better than chance. Further, whether the models are based on the actually experienced probabilities or the subjective probabilities in the form of frequency or probability judgements does not seem to make a real difference. Given that the judgements were either very well calibrated or deviating in the direction of overestimation of small probabilities, it is not surprising that the two-stage model does not provide a significant increase in performance. This can be seen as additional evidence against the involvement of judgement bias in decisions from experience. Taken together, the model applications seem to indicate that at least with the parameterisation established under descriptive choice PT does not seem to describe choice behaviour in decisions from experience appropriately. Whether this can be generalised across a wider range of parameter values, and whether we have to assume different underlying processes within DfXP, will be investigated in Chapter 6.

Moreover, the results presented here are also interesting in the context of a recent explanation that has been put forward to reconcile the differences in weighting of small probabilities between decisions from experience and decisions from description. According to Erev, Glozman and Hertwig (2008), one of the factors that determines the psychological impact of rare events is explicit presentation. In the experiential sampling tasks, for example, where people have

to rely on memory, the rare event might be neglected and therefore underweighted; whereas the explicit presentation of the event in a gamble description will have the opposite effect. However, the findings presented seem to indicate that this is not the appropriate explanation here. The group of participants who had to provide probability judgements for the rare events before choice, who hence had to explicitly recall the occurrences of the rare event and who on average tended to overestimate the rare event's actual frequency, still made choices as if they underweighted small probabilities.

In summary, with the introduction of the Matched Sampling design and its combination with frequency and probability judgements, this chapter has contributed a range of new and important results suggesting that the phenomenon of underweighting in decisions from experience cannot be explained away by sampling error or judgement bias. In both Chapter 1 and 2, I have already demonstrated that the effect can be repeatedly replicated without the coexistence of any form of recency weighting. Taken together, the results from the first four experiments have therefore provided evidence that systematically rebuts all of the existing explanations put forward in the current literature (Fox & Hadar, 2006; Hertwig et al., 2004) to account for the differences between descriptive and experiential choice. The results from the Matched Sampling experiment will be of particular help in concluding the sampling error debate which has been dominating the recent discussion on decision from experience. The end of this chapter can therefore also be seen as the end of the first section of the thesis which has dealt with the systematic examination of existing explanations of decisions from experience phenomenon. With no obvious lines of further investigation, the second part of the thesis will explore potential alternative mechanisms which have

not yet been considered. The following chapter will begin this process by drawing from observations that have been made in this first set of experiments and looking into the impact of sampling order in decisions from experience.

CHAPTER 4

EFFECTS OF SAMPLING ORDER

4.1 Introduction

In Chapter 3, I have shown that neither of the original explanations from the literature can fully explain the underweighting of small probabilities in decisions from experience (DfXP). I therefore devote the second part of this thesis to the exploration of alternative causes for the phenomenon. Before presenting a further series of experiments, I briefly want to address the problem of equivalence of the underlying cognitive tasks in descriptive and experiential choice. From the start, results in experiential choice tasks have mostly been investigated in terms of EV and PT frameworks. This was an obvious starting point as it provided sophisticated models which had already been established in descriptive choice problems.

An alternative motivation from the beginning of the decision from experience work, has been to look at decision making processes that appear to connect to the reinforcement learning research tradition in animals (e.g., Erev & Haruvy, in preparation), which is based on completely different assumptions about the underlying mechanism. One of the few attempts that have been made in the context of DfXP to apply such a statistical learning model as an alternative to account for the findings was provided by Hertwig et al. (2006), who tried to model the underlying processes in terms of a variant of the fractional adjustment model (March, 1996). This will be explored in more detail in Chapter 6.

In terms of the PT framework, the simplest way of applying it to DfXP was to divide the DfXP task into two distinct stages. In the first stage, the

outcomes and their probabilities are inferred from the samples drawn, either by actually calculating the average payoff or through estimation. In the second stage, subjective probabilities are then multiplicatively combined with the actual outcomes in a way identical to EV and PT models. It is therefore not surprising that the two-stage model has been applied to account for the DfXP, as it very closely resembles such a process based on very similar assumptions. In terms of the differences between the choice behaviour in decisions from experience, the causes have mainly been attributed to the first stage. In particular, the equivalence in terms of the summary statistics derived from the sequences as a whole has been the focus of the investigation. As pointed out by Fox and Hadar (2006), the involvement of sampling error in the original DfXP design did not allow the conclusion that the same choice problem can lead to different choice behaviour when presented in a series of outcomes instead of a gamble description. In the last chapter, I have shown that this can be remedied and that sampling error can be eliminated. The Matched Sampling paradigm has made it possible to present an experiential format that is at least equivalent to a gamble description in terms of the summary statistics regarding the enclosed outcomes and their probabilities as they can be extracted from the sequence.

This does not necessarily mean, however, that the information within the two tasks are combined in the same way in order to be utilised for the decision making process. In the context of a sequential sampling process other strategies seem equally plausible and should be considered. Even if across the whole sequence the same information is presented, the aggregation of this information over time might trigger different cognitive processes and could result in a different representation of the uncertainty attached to the outcomes. This shows that there are limitations in terms of the equivalence between the two tasks.

After the rebuttal of sampling error and judgement error as the main causes for the underweighting, it is not clear what other hypotheses within the two-stage framework might yield more appropriate predictions. One option is to explore a potential difference in the actual weighing and value transformations in experiential choice and in descriptive choice, which will be the focus of Chapter 5 and 6.

The claim of this chapter is that by imposing the two-stage framework onto the experiential choice task we may have missed out on an opportunity to discover alternative properties and processes that are part of the sampling task and which could provide a wider set of explanations for the differences in choice behaviour. I therefore want to take a step back and re-examine the sampling process by exploring some of its properties that have been observed in the experiments of the previous chapters. In particular, the repeated switching between options during the exploration phase has been shown to be a consistent finding in the information search within decisions from experience. This implies that participants form clusters of samples from the buttons which could facilitate repeated relative comparisons throughout the sampling process whenever they switch sampling from one option to the other. By assuming such additional cognitive processes preceding the final choice tasks we can see how, all of a sudden, different evaluation strategies have to be taken into account. Consequently, the same lottery can be processed in numerous ways depending on how it is partitioned into different sub-samples.

One variable that determines this partitioning of the outcome sequence of an option is sampling order. Preventing participants from switching between options and forcing them to explore both options separately, for example, could facilitate an evaluation of two sequences as a whole and would provide representations of the options' outcomes and probabilities that map more naturally onto a two stage decision from experience model. Despite the differences in obtaining the

information during the first stage, the representations of the information that enters the second stage would then be more similar between experiential and descriptive choice tasks. As a result, choice behaviour might also be more comparable to descriptive choice. Controlling sampling order would therefore induce a closer equivalence between the descriptive and experiential decision problems, and could potentially reduce or eliminate the description-experience gap.

By investigating decisions from experience under conditions with systematic manipulations of sampling order, the experiments presented in this chapter will provide an initial experimental test of the impact of different representations induced by properties of the sampling sequence and the involvement of relative evaluation of the two options in DfXP. A more theoretical analysis investigating additional properties of the experienced samples will be presented in Chapter 6.

4.2 Fixed Sampling with Probability Judgements (Experiment 5)

The first sampling order experiment takes up the example provided above and examines whether the underweighting of small probabilities can still be observed when the two options have to be explored separately. By fixing sampling order experimentally and preventing switching between options, the experiment provides an experiential choice task that resembles descriptive choice more closely, enabling the evaluation of the options as a whole. If the partitioning of the options into sub-samples and the opportunity for additional comparisons between these sub-samples has been involved in provoking the pattern of underweighting of small probabilities under unrestricted sampling in the experiments previously

reported, then we would expect the underweighting to be eliminated or reversed under conditions involving restricted sampling order.

4.2.1 Method

4.2.1.1 Participants

200 participants, 65 male and 135 female, aged between 15 and 60 with a mean of 25 years completed the experiment which was again conducted over the Web. Recruitment was once more organised through various internet portals of the University of Warwick, the Hanover College and the University of Central Lancashire advertising psychological experiments.

4.2.1.2 Design and procedure

The design was identical to the Matched Sampling design in Experiment 3 with a frequency judgement task at the end of the experiment in which the participants were asked to provide estimates for the frequencies of the rare events in both sequences sampled. The same six choice problems were used and every participant was presented with only one problem which was randomly selected. 40 samples had to be drawn without replacement from each button. The only difference was the order in which the options could be explored. Instead of allowing exploration in whatever order preferred, participants had to sample all 40 outcomes from option 'A' before they could start sampling from option 'B' (40-40 sampling order). Both the gamble and the allocation of its options to the two buttons were randomly selected. The experiment ended with the presentation of feedback on the subjects' obtained score, their frequency estimates, and the actual frequencies.

4.2.2 Results

As both the number of samples (Matched Sampling design) and the actual order in which to sample from the two options was fixed and identical for all participants, there were no differences in the actual information search pattern between participants. The result section presented here will therefore start with the analysis of the observed choice behaviour without further delay.

4.2.2.1 Choice proportions

I first tested whether there were any deviations in the direction of overweighting of small probabilities observable under fixed sampling order by comparing the choice proportions obtained in this experiment with the descriptive choice proportions from Experiment 2 reported in the previous chapter. The mean absolute difference between the proportions was 15%. The overall mean of choices in the direction of overweighting of small probabilities was significantly higher under descriptive choice (61%) than under 40-40 sampling (49%), $t(273) = 2.255$, $p = .027$, two-sided. The proportions of H choices within the different choice problems are provided in Table 4.1., together with the actual number of participants in each choice problem. The p-values of Fisher's exact test for the comparisons of the differences between the two conditions across the six decision problems are presented in the column next to the choice proportions. Reversals in terms of preferences across the 50% line are still observed for four out of six choice problems (3, 4, 5 and 6). Significant differences though were only found in decision problem 4.

In a second analysis, I tested whether there were any differences between the Matched-Sampling Condition with fixed order (40-40) and the Matched-

Sampling Condition with free sampling order from Experiment 3. The mean absolute difference between the proportions of the two Matched-Sampling Conditions was relatively high though with 14.33%, which is also very similar to the value reported for the comparison of fixed sampling and descriptive choice. The overall proportions of choices in the direction of overweighting was higher under Matched Sampling with fixed sampling order (49%) than under Matched Sampling with free sampling order (38%), though the difference was not significant with Fisher's exact $p = .83$. Within the different choice problems only one significant difference was found for choice problem 1 (see last two columns in Table 4.1) for which the proportion under fixed sampling was close to the proportion found under descriptive choice.

TABLE 4.1

Summary of the observed proportions of H choices for the experimental conditions including the p-values (Fisher's exact tests) for the differences between the experiential and descriptive choice proportions. Significant differences are highlighted with asterisks.

			Percentage choosing H					
Decision Problem	H	L	Fixed Sampling Order	n	Description (Exp.2)	*p*	Matched Sampling (Exp.3)	*p*
1	4, .8	3, 1.0	38	29	36	1.00	69*	.037
2	4, .2	3, .25	55	33	72	.274	39	.222
3	-3, 1.0	-32, .1	48	31	64	.288	42	.799
4	-3, 1.0	-4, .8	66	36	33*	.040	55	.486
5	32, .1	3, 1.0	59	32	48	.432	45	.317
6	32, .025	3, .25	36	39	52	.300	26	.452
mean	all problems		49	200	61*	.027	38	.832

The differences between the different conditions across the six decision problems are again presented in Figure 4.1. The grey bars show that the differences between description and fixed sampling are still pointing in the direction of less overweighting of small probabilities with the exception of

decision problem 5. However, the differences are also no longer as large as under Matched Sampling with free sampling order.

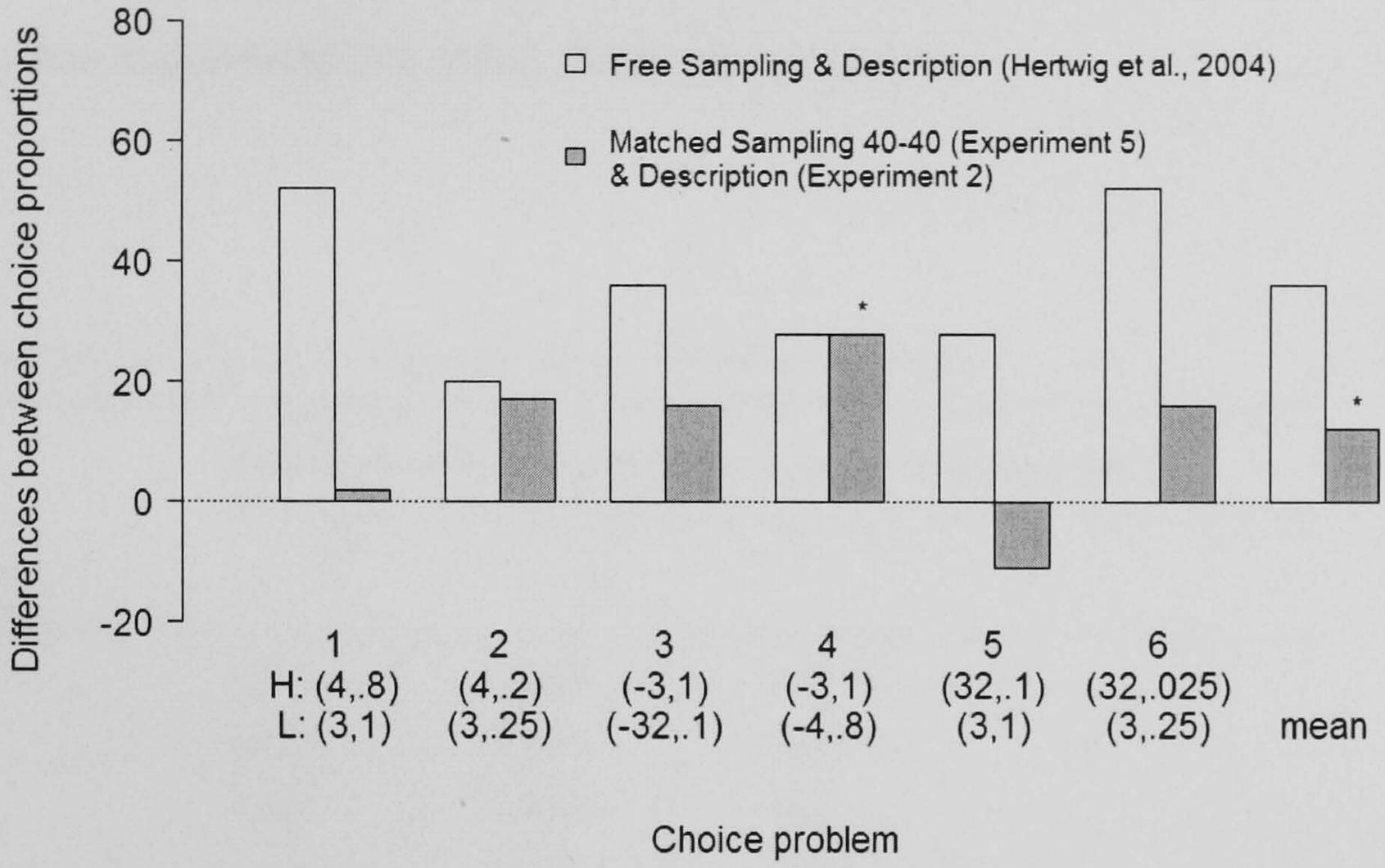


Figure 4.1. Differences in proportions of maximising choices between Matched Sampling with fixed 40-40 sampling order (Experiment 5) and descriptive choice (data from Experiment 2) across the different decision problems. For comparison the proportions reported by Hertwig et al. (2004) have been added (* $p < .05$, ** $p < .01$). The asterisk for the mean refers to the t-test results provided above.

Given all these results, it seems like the proportions from the 40-40 sampling condition lie somewhere between descriptive choice and Matched Sampling with free sampling order. To test whether the proportions from the different conditions follow a qualitative ordering an additional analysis was conducted. The details of this analysis are described in Barlow (1972) and Fleiss, Levin and Paik (2003). For this test, I assumed that the three proportions of maximising choices are ordered in terms of the extent to which they exhibit underweighting of small probabilities. The following ordering was predicted:

H_1: $p_{xp\_free} > p_{xp\_fixed} > p_{descr}$.

This was tested against the null hypothesis that all proportions are equal:

$$H_0: p_{xp\_free} = p_{xp\_fixed} = p_{descr.}$$

The $\bar{\chi}^2$ statistic for the test is shown in Table 4.2. In four out of the six choice problems a significant ordering could be found in the predicted direction, indicating less underweighting under fixed sampling order.

TABLE 4.2

Statistics and p-values for the tests of equality of ordered proportions

	Descriptive choice proportions from previous experiment		
Choice problem	$\bar{\chi}^2$	*C*	*p*
1	8.069	0.489	<.01 **
2	6.082	0.457	<.05 *
3	2.797	0.472	>.05
4	4.106	0.459	<.05 *
5	0.678	0.465	>.1
6	4.268	0.430	<.05 *

The p-values are looked up from the table provided by Barlow (1972)

In addition, a logistic regression was conducted using the data from (a) the 40-40 Sampling Condition, (b) the equivalent Matched-Sampling Condition with free sampling order, and (c) the Description Condition from Experiment 2. The categorical explanatory variables were presentation format ('experienced' vs 'description') and exploration mode ('sub-samples' vs. 'whole sequence'), and choice in the direction of underweighting was the dichotomous dependent variable. The exploration mode 'whole sequence' included Matched Sampling with 40-40 sampling order and descriptive choice, as they were assumed to have similar representations. The mode 'sub-samples' included the Matched Sampling data with free sampling order. Both factors' effects on the likelihood of making choices in the direction of underweighting was statistically significant (see Table

4.3), with no significant interaction. The significant slopes in the table indicate that proportionately more choices in the direction of underweighting can be observed under conditions where information is experienced and under conditions where information can be sampled freely.

TABLE 4.3

Logistic regression results for the variables of presentation format and exploration mode

	Estimate	Std. Error	*z*	*p*
(Intercept)	-.433	.167	-2.592	.010
Presentation format ('experienced' vs. 'description')	.427	.204	2.092	.036
Exploration mode ('sub-samples' vs 'whole')	.493	.219	2.253	.024

In summary, the proportions found under fixed exploration do still reflect underweighting of small probabilities although to a lesser extent than observed in the earlier DfXP experiments. The intermediate position of the proportions exhibited under Matched Sampling with fixed sampling order seems to be a result of the separate contribution of the two variables presentation format and exploration mode.

4.2.2.2 Frequency judgements

The provided frequency estimations were again analysed after reiteratively removing extreme outliers that deviated by more than 3 standard deviations from the mean. Out of the 400 judgements 26 were excluded. The mean absolute difference between the actual experienced frequencies and their estimates was $M = 2.45$ ($SD = 3.96$). In terms of their accuracy participants were again very well adjusted, $r\,(372) = .96$, $p < .001$; $R^2 = .91$, $F(1,372) = 3953$, $p < 0.001$). It is

obvious from Figure 4.2 that there is again no indication for potential underestimation of small frequencies. Rather, there are deviations from the diagonal in the direction of overestimation of small frequencies and underestimation of high frequencies. This is confirmed by a series of tests examining whether the means of the estimation errors are different from zero (from low to high frequency: $t(37) = 2.32$, $p = .026$; $t(61) = 1.86$, $p = .068$; $t(86) = 3.24$, $p = .002$; $t(68) = 2.26$, $p = .027$; $t(117) = 3.82$, $p < 0.001$, all two-sided).

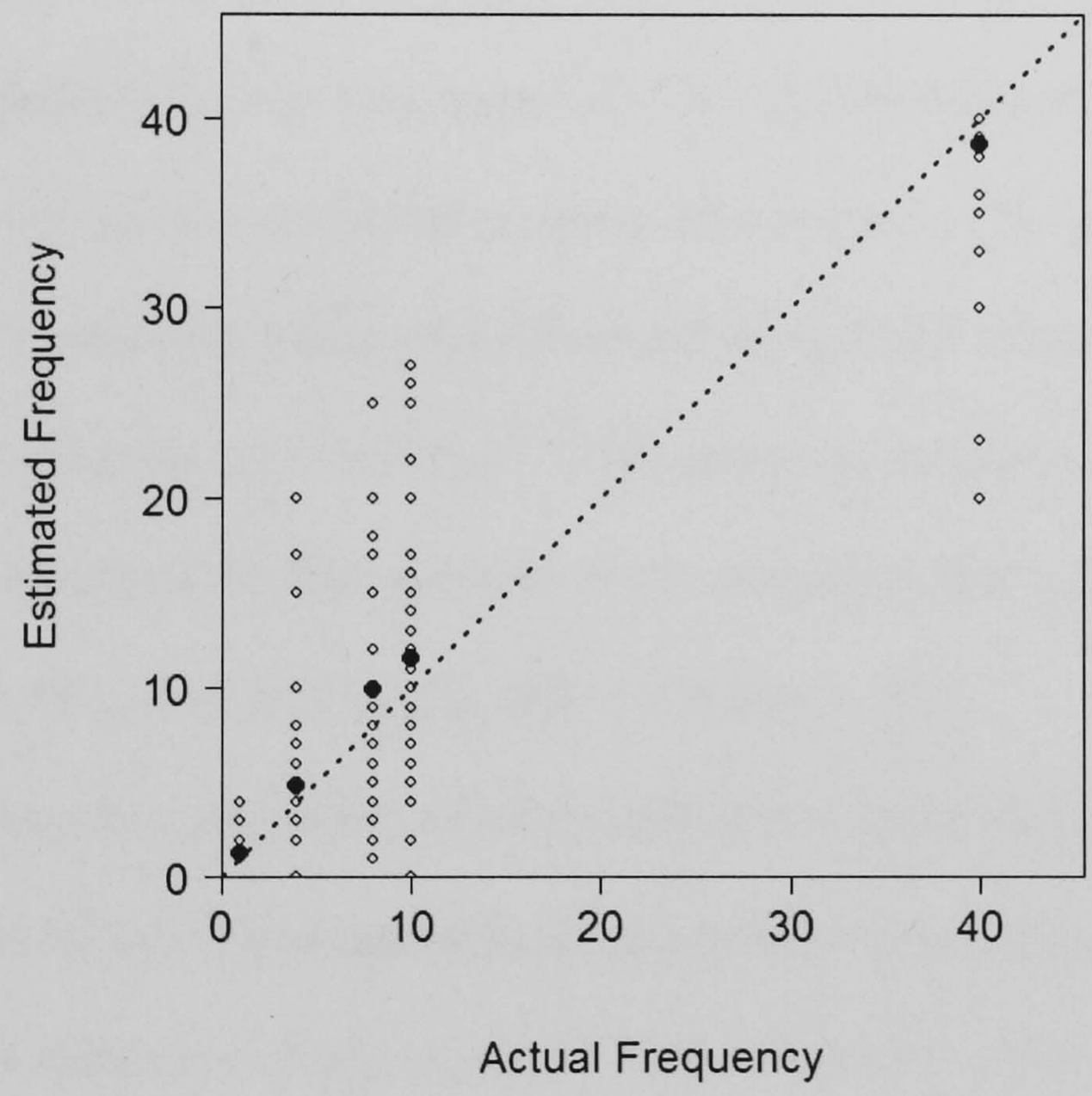


Figure 4.2. Deviations of the frequency judgements for the rare events in Experiment 5 plotted against the actually experienced frequencies. Due to the overlap of the probabilities in the six choice problems used there were only 5 different rare event frequencies. The dotted line indicates perfect calibration. The black dots are the mean estimates. The white dots indicate the observed estimation errors. One white dot may represent several data points from different participants.

Taken altogether, the results of the frequency estimations under fixed sampling order match the results presented in Experiment 3. A greater accuracy of the frequency judgements due to fixed sampling order was not found. Instead, the strength of the relationship between actual and estimated frequencies was the same for the two experiments ($z = 4.75$, $p = 1.00$), independent of sampling order.

4.2.2.3 Recency weighting

Fixing the sampling order did not make any difference in terms of recency weighting. As in the previous experiments, the rate of correct predictions based on different parts of the two sequences were just around 50% and did not differ significantly between the different splits. This applies to the predictions on the basis of the expected values of the first and second half of the sequences (50% vs. 49%, χ^2 McNemar (1) = 0.022, $p = 1.00$) and to the predictions based on the expected values of the four quartiles of the sequences (from 1 to 4: 55%, 50%, 53%, and 50%, Cochran's Q (3, 200) = 1.065, $p = .747$).

Also, the comparison of the proportions of maximising choices for the appearance of the 32 points within different parts of the sequence of choice problem 6 does not indicate any significant differences between earlier and later parts of the sequence. This is the case for encounters within the first or second half of the sequence (41% versus 29% correct predictions, Fisher's Exact $p = .518$ and the more extreme comparison of encounters within the first or last seven outcomes of the sequence (60% versus 50% Fisher's Exact $p = 1.00$). The actual trends within both comparisons point more towards a primacy effect.

4.2.2.4 Application of descriptive choice models

The rate of correct predictions on the basis of the experienced probabilities was 50% for EV maximisation and 47% for the PT model using the parameters by Tversky and Kahneman (1992). When using the valid frequency judgements instead, EV maximisation could account for 63% of the choices correctly, which is a significant increase (χ^2 McNemar (1) = 8.643, p = .003), whereas the two-stage model still predicted only 48% correctly (χ^2 McNemar (1) = .25, p = .617). Compared with the data from the Matched-Sampling Condition of Experiment 3 with free sampling order reported in Chapter 3, the rates were slightly higher. A significant difference though was only found for the EV maximisation rates based on the estimations (Fisher's exact p = .005). However, overall the fits remain still rather low, as illustrated in Figure 4.3.

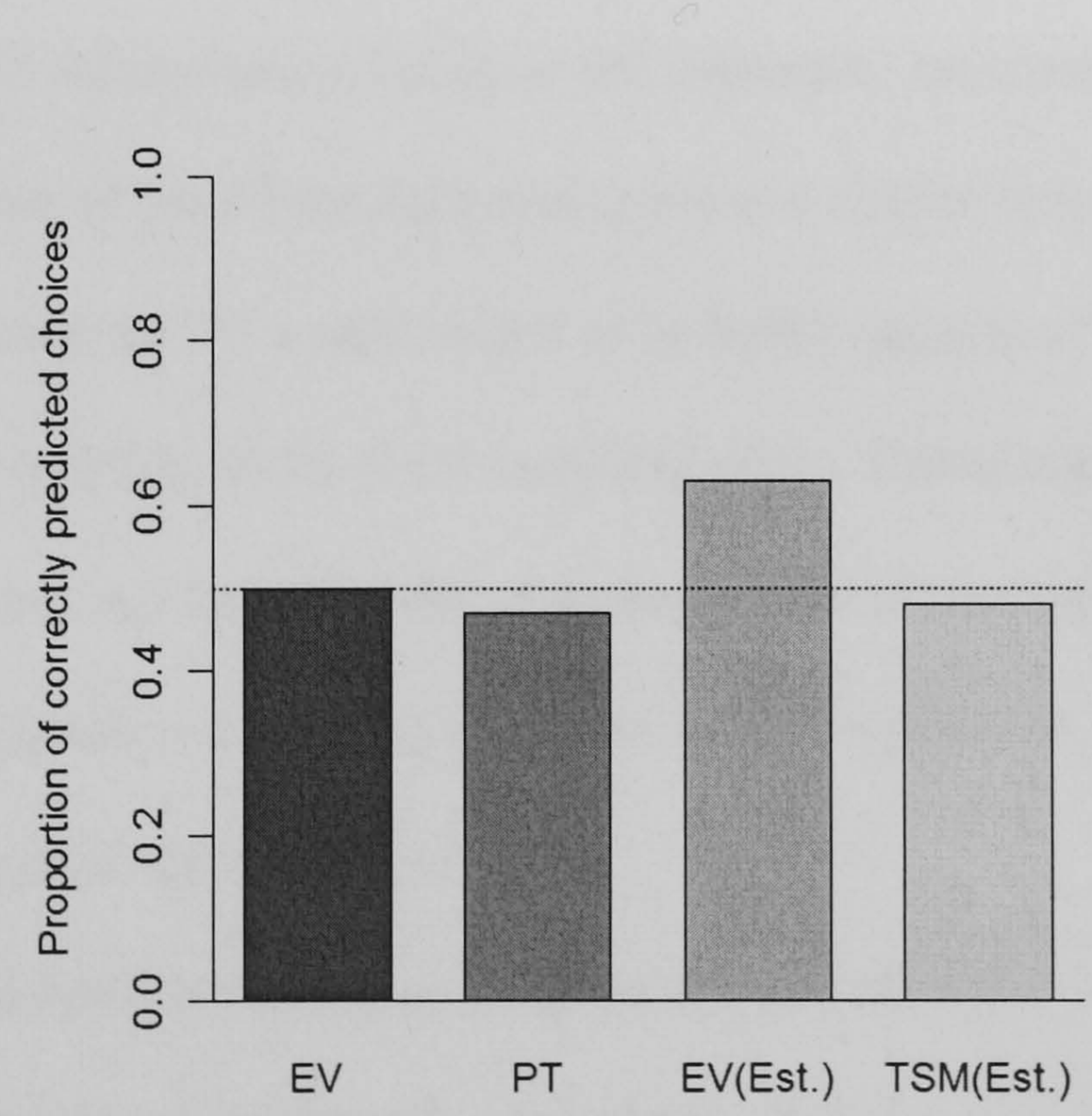


Figure 4.3. Percentage of correct predictions for the different choice models: expected value, prospect theory, expected value based on the estimated frequencies, and the two-stage model (also based on estimated frequencies). Only the expected value model based on the estimated probabilities performs above chance which is shown by the dotted line.

4.2.3 Discussion

The results of this experiment suggest that the exploration of the options one after the other acts as a moderating variable which makes the problem more similar to descriptive choice but does not reverse the underweighting into overweighting observed in DfD. Instead, we find the choice proportions to be between descriptive choice and Matched Sampling with free sampling order. Interestingly, the fixed sampling order does not affect any of the other properties observed in experiential choice. Although the fixed sampling order facilitates the aggregation of the sequential information, the accuracy of the frequency judgements was not found to be more accurate than under free sampling order. There was also no difference in terms of the absence of recency weighting. Only the rate of correct predictions of descriptive choice models did differ slightly, with an increased performance of EV maximisation based on the frequency estimations. However, the predictive power of the PT model remains around chance level. With linear probability weighting the EV model seems to be better capable of incorporating the reduced overweighting under fixed sampling order. Taken together, these results show that making the task even more equivalent to descriptive choice by facilitating the integration of the outcome information separately and as a whole is not enough to eliminate the DfXP pattern.

However, a few potential problems need to be addressed. Firstly, we have only looked at one choice problem per participant. In order to get an idea of the robustness of the choice pattern under fixed sampling it would be helpful to have participants completing all six decision problems. It is also important to keep in mind that a small percentage of participants in the Matched-Sampling Condition with free sampling order did actually sample in the same order as the fixed

exploration group (9%). Unfortunately, though, this group is too small to conduct a separate analysis with its data. The remaining participants are difficult to classify, and it might be inappropriate to compare such an amalgamation of different sampling orders. To provide clarification it would therefore be expedient to compare the fixed exploration with an alternative sampling strategy that is more distinct and also more homogeneous. Continuous alternation between options represents such a more extreme partitioning of the sequence. It has the highest degree of partitioning into sub-samples with a sample size of one for each sub-sample. The integration of the sequential information per option in this condition is most difficult due to higher working memory loads. The following experiment aims to address these issues with an extended sampling order design.

4.3 Within-Participant Analysis (Experiment 6)

The purpose of the second sampling order experiment was to address the methodological problems described above. Consequently, it includes the addition of an order condition that is more distinguishable from the 40-40 Sampling Condition, by forcing participants to switch after every single sample. Furthermore, it comprised the completion of all six choice problems in order to get a more reliable measure of the extent of the underweighting of small probabilities under fixed sampling order. The rationale was the same as above, assuming that observing the sampled outcomes in different orders could result in a different representation of the probabilities of the options. However, the experiment goes further by introducing a design that allows me to test whether the reversals between descriptive and experiential choice tasks can also be replicated

within participants. Until now, the experiments in the literature have only compared descriptive choice and experiential choice between participants.

4.3.1 Method

4.3.1.1 Participants

150 participants (56 men and 94 women, aged between 18 and 56 years, with an average age of 32) completed the Web-based experiment. The recruitment was conducted through 'ipoints', by sending out an invitation email to a random sample of their database. In exchange for their participation the participants received 200 ipoints, which is equivalent to £2.

4.3.1.2 Design

The six choice problems used in the previous experiments were presented to all participants in two formats, once in the form of a description format and once in a Matched Sampling format. Every participant therefore took part in the DfD condition and one of the three DfXP conditions (either Free Sampling, 40-40 Sampling or 1-1 Sampling). A summary of the design is provided in Table 4.4.

TABLE 4.4

Experimental design used in Experiment 6

		Between-participants conditions		
Within-participants measures	6 choice problems	1-1 Sampling Order	40-40 Sampling Order	Free Sampling Order
	6 choice problems	Descriptions		

The overall order of the resulting 12 decision problems in the two different presentation formats was mixed and completely randomised for each participant (e.g., XP3, Des4, XP5, Des2, Des1, XP6). Within the descriptions of the prospects, probabilities were expressed as percentages and presented in the following format:

p % chance to win/lose x points, 1-p % chance to win/lose 0 points.	or	q % chance to win/lose y points, 1-q % chance to win/lose 0 points.

In the Matched Sampling format, probabilities had to be inferred from 40 outcomes sampled from each of the two options available. The prospects were matched onto the sequences according to the Matched Sampling design introduced earlier. Whereas the descriptive choice tasks were identical for all the participants, the Matched Sampling task was implemented in three different order conditions, a Free-Sampling-Order Condition a 40-40 Sampling-Order Condition, and a Forced-Alternation Condition (1-1 Sampling-Order). The former two have already been used in the first sampling order experiment described above. New was the Forced-Alternation Condition in which participants had to alternate between the options after every single sample ({A,B,A,B,A,B,…}). Sampling within all the predetermined orders started with button 'A' and continued according to the specific schedule of the condition. Non-available buttons turned grey indicating when to switch to the other option. The assignment of participants to one of the three order conditions was randomised as was the placement of the options ('A' or 'B' button) within each prospect.

After the sampling phase, participants had to choose their preferred option which was played once at the end of the experiment. The points obtained were added to the final score. This phase was identical for both the descriptive and the experiential conditions. Participants were not asked to estimate frequencies in this task. As in previous designs, the experiment ended with the presentation of feedback including the points total and a summary of the outcomes of all the chosen lotteries.

4.3.2 Results

4.3.2.1 Information search

As the information search in the two predetermined sampling order conditions, the 40-40 Condition and the 1-1 Condition (Forced-Alternation), was restricted there was no need to analyse the information search within these conditions separately. Within the Matched-Sampling Condition with free sampling order, on the other hand, a separate analysis was necessary to explore the sampling strategies used. The median number of switches was higher than in Experiment 2 with 2 (M = 10.82) switches per choice problem. Only 45% of the sequences were explored in the same way as in the 40-40 Condition with only one switch after finishing the sampling from one option exhaustively. Exploration similar to the 1-1 Condition (79 switches) was observed in only 1% of the sequences. The majority of the sequences (54%) were explored with a number of switches somewhere between the two extremes. The overall distribution of the number of switches is provided in Figure 4.4. When put in relation to the possible number of switches the median switching ratio was .025 (M = .1).

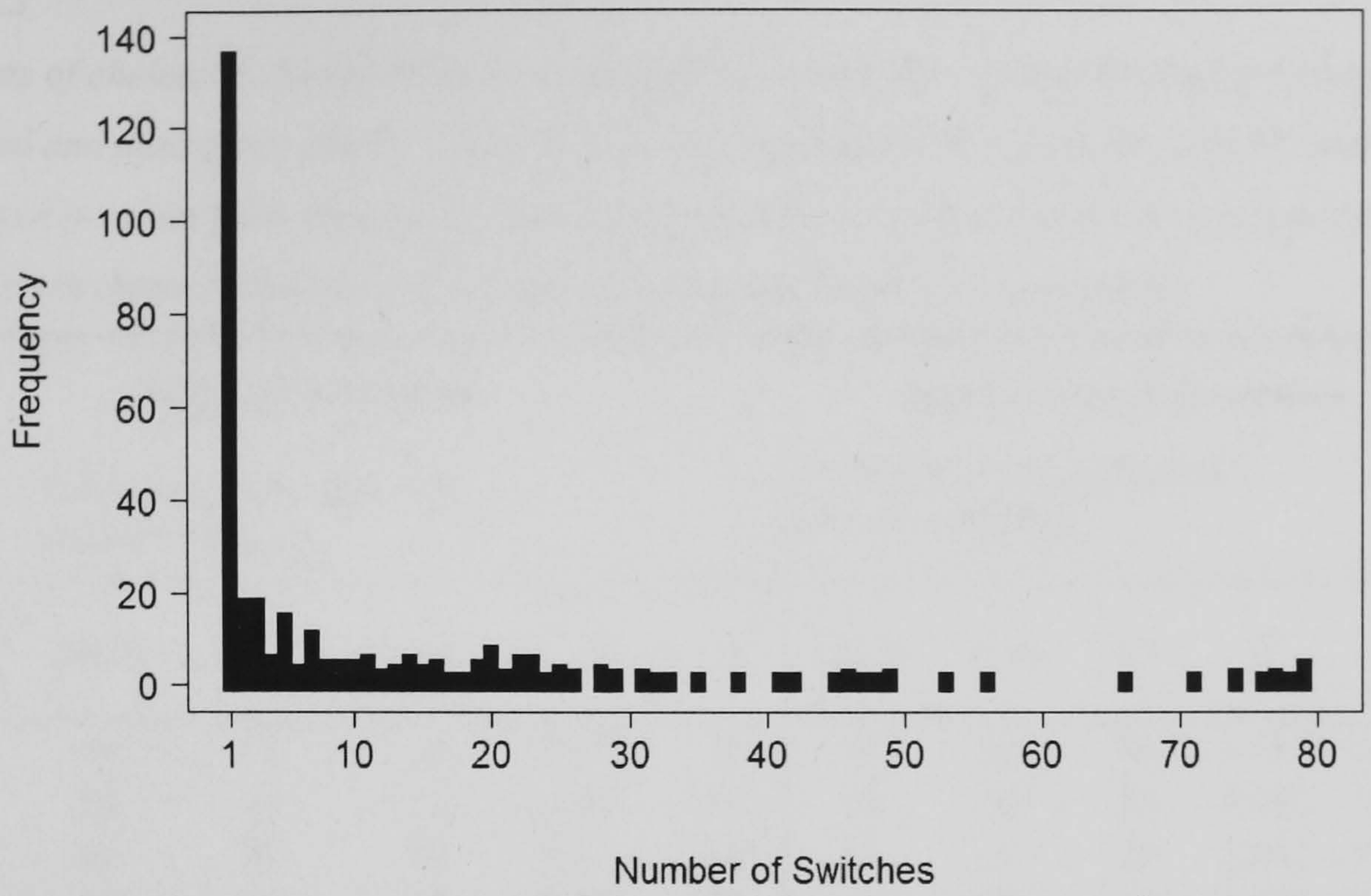


Figure 4.4. Distribution of the number of switches between buttons in the Matched-Sampling Condition with free sampling order in Experiment 6.

4.3.2.2 Choice behaviour

In a first analysis I tested whether there were any differences between choice behaviour in the three order conditions for descriptive and experiential choice problems. When analysing the data on the level of individual choice problems, the proportions of choices in the direction of overweighting within the individual choice problems did not differ between the three order conditions, neither within the experiential choice problems nor between the descriptive choice problems (see *p*-values in Table 4.5).

TABLE 4.5

Proportions of choices in the direction of overweighting across the individual choice problems for experiential and descriptive choice within the different experimental conditions. The chi square statistics and p-values from the tests of equality of proportions between the three conditions for the data from both choice formats are provided in the last two columns in each block.

	Matched Sampling					**Decision from Description**				
	% choices in the direction of overweighting					% choices in the direction of overweighting				
Decision Problem	40-40	Free	1-1	χ^2	*p*	40-40	Free	1-1	χ^2	*p*
1	56	40	52	2.774	.25	58	60	68	1.189	.552
2	56	62	72	2.813	.245	64	50	70	4.442	.109
3	46	28	46	4.5	.105	32	38	18	5.082	.079
4	40	60	44	4.487	.106	68	76	78	1.029	.598
5	46	32	34	2.451	.294	32	32	38	.535	.765
6	40	40	46	.493	.782	32	40	36	.694	.707

The proportions for the order conditions were therefore combined. A comparison based on these combined proportions shows that there was a significantly higher proportion of choices in the direction of overweighting under descriptive choice in two out of six choice problems (see Table 4.6). In three out of four problems where no significant difference could be found, the underlying reason seemed to be surprisingly low proportions in the direction of overweighting within descriptive choice. High proportions in the direction of overweighting within the experiential choice problems were only found in one of the six choice problems.

TABLE 4.6

Combined choice data for the individual choice problems including the p-values for the McNemar tests conducted.

	Choices in the direction of overweighting		
Decision Problem	Descriptive Choice	Experiential Choice	*p*
1	62	49	.027
2	61	63	.810
3	29	40	.056
4	71	48	.001
5	34	37	.576
6	36	42	.281

In order to test whether there were any differences between the mean proportions in the direction of overweighting of the three order conditions for descriptive and experiential choice problems a mixed design ANOVA was conducted with Sampling Order as a between-participant factor and Choice Format as a within-participant factor. The results revealed that there was no main effect of Choice Format ($F(1,147) = 1.202$, $p = .275$). There was also no main effect of the between-participants factor Sampling Order ($F(2,147) = .435$, $p = .648$) and no significant interaction between the two ($F(2,147) = .724$, $p = .486$). This is also illustrated in Figure 4.5.

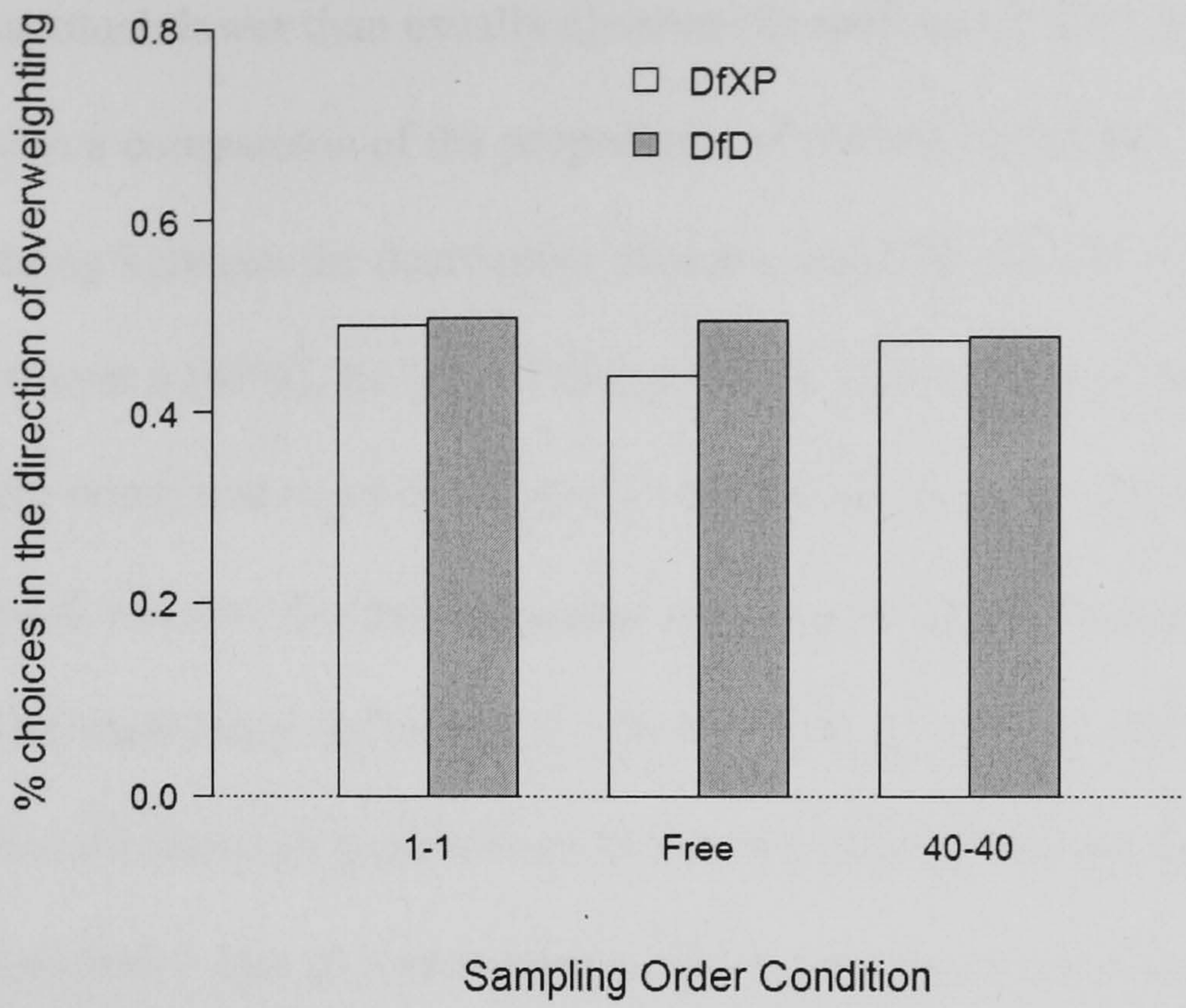


Figure 4.5. Interaction plot for the mean choices proportions within the different order conditions with different degrees of partitioning into sub-samples in Experiment 6.

Firstly, the mean percentages of choices in the direction of overweighting were very similar for all three order conditions around 50%. Secondly, the bars of both formats DfXP and DfD have also similar heights. It was therefore not possible replicate the between subjects difference between the mean proportions in the direction of overweighting for DfD and DfXP from earlier experiments. A higher rate of overweighting in the 40-40 Condition than under 1-1 Condition could also not be found. With no differences between the sampling order conditions, an analysis testing whether there was a significant ordering of the means in terms of overweighting of small probabilities was not indicated and thereby not conducted. The in-between position 40-40 Condition from the last experiment could consequently not be replicated either.

The main reason for the reduced gap between the percentages of choices in the direction of underweighting between experiential and descriptive choice seems to be related to fact that the extent of overweighting under descriptive

choice was much lower than usually observed in such tasks. This could be confirmed in a comparison of the proportions of choices in the direction of overweighting between the descriptive choice conditions of Experiment 2 (61%) and Experiment 6 (49%), $t(173) = 2.691$, $p = .011$. Conversely, a comparison between the combined experiential choice data of all order conditions in Experiment 6 and the Matched Sampling condition in Experiment 2 showed that there was no significant difference, $t(173) = -1.179$, $p = .246$. However, the proportions of choices in the direction of overweighting between the same experiential choice data in Experiment 6 and the descriptive choice condition in Experiment 2 did differ significantly, $t(173) = 3.2$, $p = .003$. Together, both results can be seen as evidence for a successful replication of choice in the direction of less overweighting within the experiential choice problems.

4.3.2.3 Within-participant reversals

In addition to the between-participant analysis described above comparing the means across the different choice problems between the two formats, I also conducted an analysis investigating preference reversals within participants. For this analysis, I counted the number of times participants preferred different options when faced with the same choice problem in the two different choice formats. The overall rate of such preference reversals across the six choice problems was 41%. In the rest of the cases participants chose the same option in both formats. The actual reversals were further classified in terms of whether they occurred in the direction predicted by DfXP or in the opposite direction. In choice problem 1, for example, a choice reversal in the direction of DfXP would mean that a participant choose the sure option when presented with a gamble description and the riskier option after sampling from the buttons. The distribution of

reversals across the different choice problems and their split into the two categories is shown in Table 4.7.

TABLE 4.7

Breakdown of the observed proportions of preference reversals for the six choice problems including the statistics from the McNemar's tests on the differences between the proportions.

Choice problem	H	L	% of observed reversals	% of reversals in direction of DfXP	% of reversals in direction opposite of DfXP	McNemar's $\chi 2$	p
1	4,.8	3,1.0	45	64	36	5.388	.020
2	4,.2	3,.25	46	48	52	0.130	.718
3	-3,1.0	-32,.1	41	37	63	4.129	.042
4	-3,1.0	-4,.8	43	77	23	18.063	.001
5	32,.1	3,1.0	34	45	55	0.490	.484
6	32,.025	3,.25	37	42	58	1.473	.225

In two out of the six choice problems (1 and 4), there were significantly more reversals in the direction of DfXP. For problem 3, on the other hand the opposite pattern was found.

4.3.2.4 Recency weighting

A comprehensive analysis of recency weighting was conducted comparing the mean percentage of correct predictions of the first and second half of the sampled sequence and the four quartiles of the sequence. Table 4.8 shows that the percentages of correct predictions are close together.

TABLE 4.8

Mean percentages of correct predictions for the different splits across the three order conditions

	Sequence Split					
Sampling Order	First 20	Last 20	Quart. 1 (1-10)	Quart. 2 (11-20)	Quart. 3 (21-30)	Quart. 4 (31-40)
1-1	.49	.52	.51	.49	.53	.47
Free	.48	.47	.51	.46	.47	.50
20-20	.45	.50	.50	.48	.47	.56
All	.47	.49	.51	.48	.49	.51

This is also confirmed by the set of repeated measures ANOVAs with the different sequence splits as a within-participant factor and Sampling Order as a between-participant factor testing whether recency weighting was facilitated by specific sampling formats. For the first and last half of the outcomes, there was no main effect of the part of the sequence the prediction was based on ($F(1,147) = .477$, $p = .491$), no main effect of the between-participants factor Sampling Order ($F(2,147) = .960$, $p = .385$), and no significant interaction between the two, $F(2,147) = .293$, $p = .746$). The same was confirmed for the percentages of correct predictions on the quartiles of the sequence. There was no significant main effect, neither for the different quartiles ($F(1,147) = .844$, $p = .470$, nor for Sampling Order ($F(2,147) = .545$, $p = .581$). A significant interaction was also not found, $F(2,147) = 1.210$, $p = .3$.

I then examined whether there was a significant difference in the number of maximising choices depending on the position of the rare event within the sequence. The analysis was conducted for the experiential choice data of choice problem 6 only using the combined Matched Sampling data (all sampling order conditions). The sequence-splits used were the same as in the previous analysis.

Again no indication for a potential impact of the serial position of the rare event on choice could be observed. There was no difference between the percentage of maximising choices for the cases in which the rare event occurred in the first or second half of the sequence (42% vs. 42%, Fisher's exact $p = 1.0$), nor for the cases in which the rare event was encountered in the first or last second seven outcomes of the sequence (48% vs. 38%, Fisher's exact $p = .572$).

4.3.2.5 Application of descriptive choice models

As all the experiential choice conditions were implemented in the form of a Matched Sampling design, the participants experienced exactly the same probabilities that were stated in the corresponding gamble descriptions. The number of correct predictions based on a model assuming that participants take the option with the highest expected value can therefore be directly calculated from the proportion of H choices. A breakdown of the rate of correct predictions of the two models across the two choice formats and the different sampling order conditions is presented in Table 4.9.

TABLE 4.9

Breakdown of the mean percentages of correct predictions

	Percent of correctly predicted choices			
Sampling Order	EV (Descriptive Choice)	EV (DfXP)	PT (Descriptive Choice)	PT (DfXP)
1-1	.38	.50	.50	.49
Free	.37	.44	.49	.44
40-40	.39	.49	.48	.47
All	.38	.48	.49	.47

Interestingly, expected value maximisation seems to predict choices better in DfXP than in descriptive choice. This is also confirmed by the results of a mixed

factorial ANOVA with Sampling Order as a between-participant factor and Choice Format and Type of Model as two within-participant factors. There was a significant main effect of the choice format on the rate of correct predictions of the models, $F(1,147) = 4.445$, $p < .042$), and a significant main effect of the Type of Model ($F(2,147) = 12.483$, $p = .001$). In addition, there was a significant interaction effect between Choice Format and Type of Model, ($F(2,147) = 20.54$, $p < .001$). However, a significant main effect of Sampling Order ($F(2,147) = .695$, $p = .501$) or a significant interaction with any other factors was not found. Figure 4.6 shows how the percentage of correct predictions of the two models differs depending on the choice format used. PT seems to provide better predictions in the context of descriptive choice, whereas EV maximisation and PT provide similar rates of correct predictions under Matched Sampling. As the rates of correct predictions of the PT model are identical to the proportions of choices in the direction of overweighting this ANOVA can be related to the ANOVA on the choice proportions presented further above. The bar chart on the right side of Figure 4.6 is therefore identical to the bar chart in Figure 4.5.

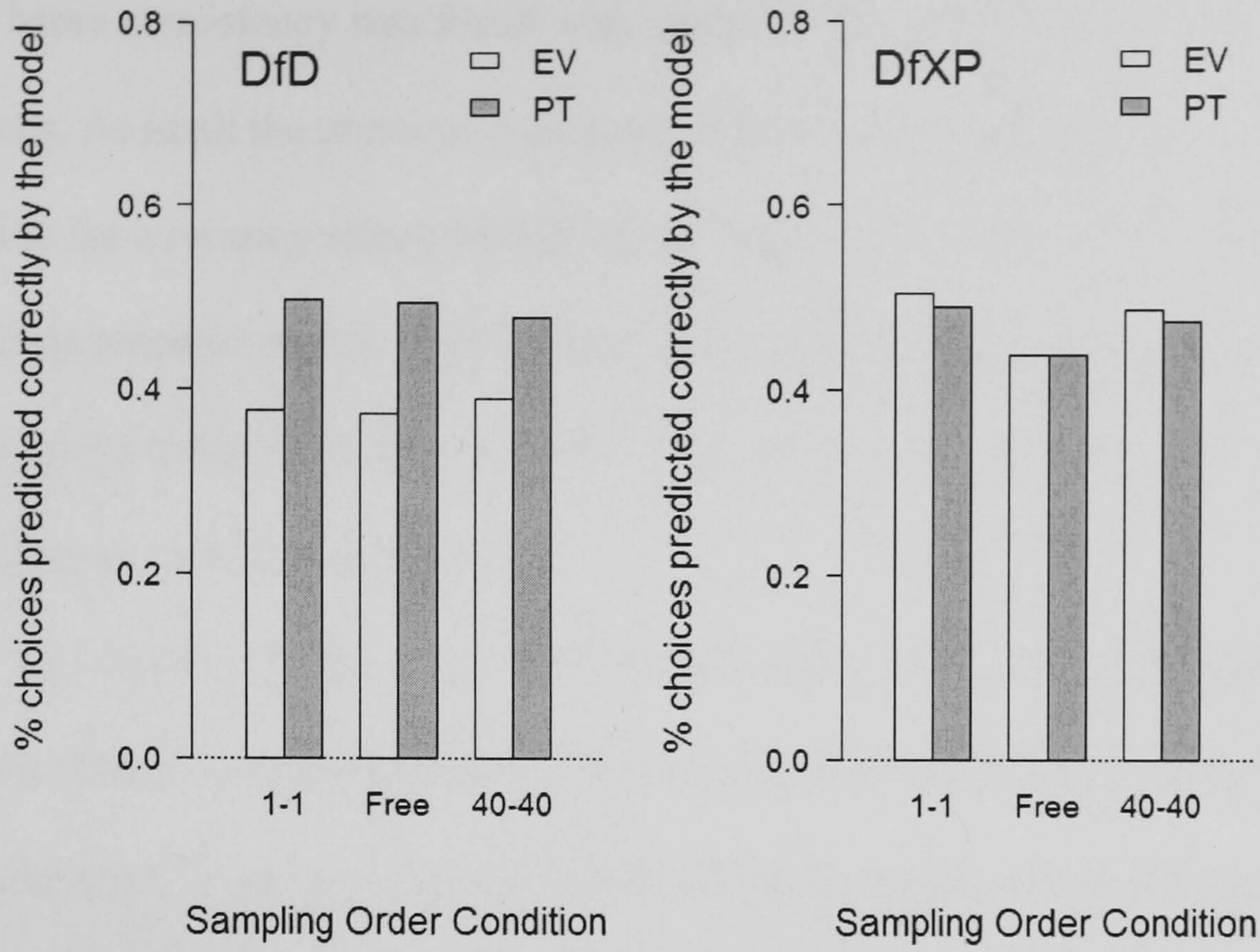


Figure 4.6. Interaction plots for the rates of correct model predictions in Experiment 6.

4.3.3 Discussion

Overall, this experiment provided no evidence for an effect of sampling order. Instead, similar proportions of choices in the direction of underweighting were found across all three order conditions. When compared with earlier descriptive choice data significant differences in the direction of underweighting of small probabilities could be replicated for the combined experiential choice data. Unexpectedly, the descriptive choice data of Experiment 6 also provided choice proportions that resembled underweighting of small probabilities more closely than overweighting, very similar to DfXP. Given the reversal within descriptive choice and its similarity with the DfXP proportions, there was also no strong evidence for preference reversals from overweighting to underweighting between the two formats. This applies to both the analysis across the means of all participants and the count of preference reversals within individuals.

More consistency was found with regard to the analysis of recency weighting. As in all the previous experiments, there was not the slightest indication for a recency effect, neither in the comparison of the rate of correct predictions based on earlier or later parts of the sampled sequence, nor in the analysis of the effect of the position of the rare event within the sequence on the proportions of maximising choices.

The results from the analysis on the predictive power of the descriptive choice models are slightly different to the results found in previous experiments, where expected value maximisation predicted choice under Matched Sampling better than PT. The superiority of PT over EV in the context of descriptive choice tasks is less surprising though. Yet, the low absolute rate of the correct predictions of both models within descriptive choice is quite unusual and far below the rates usually reported in the literature.

One potential reason for the unusual choice proportions found under descriptive choice could be some kind of an interaction between the two formats resulting from the mixed presentation of both formats. This could be either the dominance of the representation of one format over the other or the usage of alternative choice heuristics due to the exposure to both formats. Alternatively, participants might attempt to maintain consistency across the two formats and adjust their choice behaviour accordingly. To exclude the impact of such interaction an additional sampling order experiment was conducted which will be described in the closing section of this chapter.

4.4 Within-Participant Reversals II (Experiment 7)

This experiment provides a replication of the experiment above re-examining sampling order effects and within-participant reversals under further experimental control, eliminating the potential impact of the mixed presentation of the two choice formats. Moreover, an additional sampling order condition was introduced to have an intermediate degree of partitioning (5-5) in addition to the extreme cases of 1-1 and 40-40 sampling.

4.4.1 Method

4.4.1.1 Participants

The Web-based experiment was completed by a total of 250 participants (164 male and 86 female). Their age ranged from 18 and 76 years with a mean of 32 years. The recruitment was again organised through 'www.ipoints.com' and participants received 200 ipoints (= £2.0) for the completion of the experiment.

4.4.1.2 Design

The design was very similar to the previous experiment. All participants had to complete the whole set of the same six gambles, both in a descriptive format and in a Matched Sampling format with a specific sampling order. The main difference was the order in which the choice problems of the different formats were presented. For the different sampling order conditions, participants first received the six problems in the Matched Sampling format in one block before they saw the six choice problems in the form of gamble descriptions. The order of the choice problems within each block was randomised.

Furthermore, two more conditions were added resulting in a total of 5 conditions. The first three conditions were the same as described above; Free Sampling Order (Matched Sampling) Condition, Fixed Order Condition (40-40) and a Forced Alternation Condition (1-1). One of the new conditions was a 5-5 Condition in which the two buttons had to be explored in clusters of 5 samples from each button, starting with button 'A' ({A,A,A,A,A,B,B,B,B,B,A…}). The second new condition (Reversed 5-5) was added as a control group which was similar to the 5-5 Condition but with the reversed order of choice formats (DfD first, DfXP second). This condition allowed to test whether the presentation of one format had an impact on the choice behaviour in the subsequent choice formats. The rest of the procedure was the same as described in the previous experiment. A summary of the design is given in Table 4.10.

TABLE 4.10

Experimental design used in Experiment 7

		Between-participants conditions				
		4 Sampling Order Conditions				Control Group (Reversed 5-5)
Within-participants measures	First block of 6 choice problems	1-1	5-5	40-40	Free	Descriptions
	Second block of 6 choice problems	Descriptions				5-5

4.4.2 Results

4.4.2.1 Information search

Within the Matched-Sampling Condition with free sampling order, the switching behaviour was very similar to the one in the previous experiment with a median of 2 ($M = 9.8$) switches per choice problem. Put in context of the possible number of

switches (79), the median switching ratio was .025 (M = .12). Similarity was also found with regard to the observed sampling strategies. 45% of the sequences were explored in the same way as within the 40-40 Condition with only one switch after two blocks of 40 samples from each button. Exploration under free sampling order similar to the 1-1 Condition, resembling 79 switches, was again observed in only 1% of the sequences. The overall distribution of the number of switches is provided in Figure 4.7.

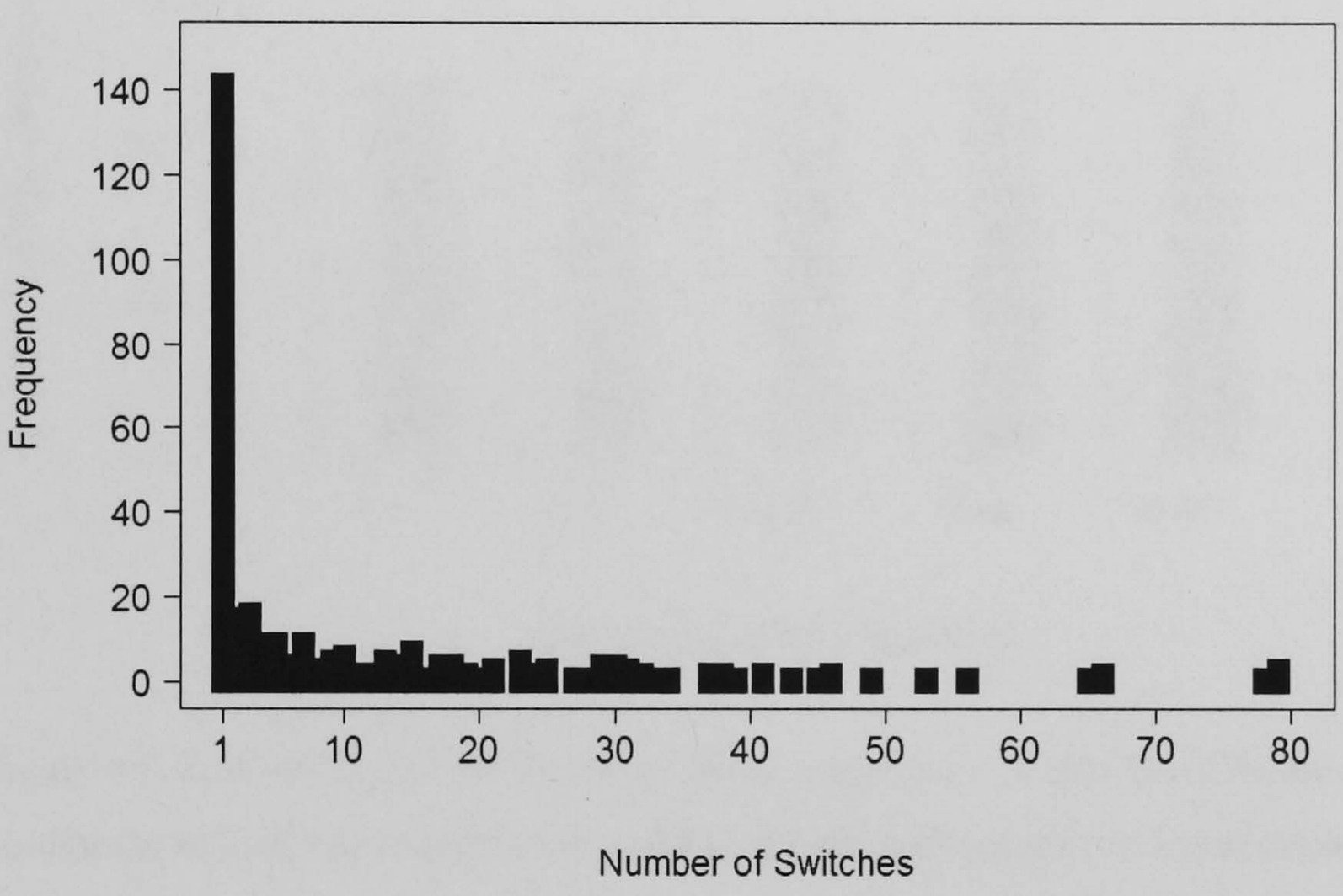


Figure 4.7. Distribution of the number of switches between buttons for the Matched-Sampling Condition with free sampling order in Experiment 7.

4.4.2.2 Choice behaviour

Following the analysis of the previous experiment, a mixed design ANOVA was conducted first with Sampling Order as the between-participant factor and Choice Format as the within-participant factor in order to test whether there were any differences between the mean proportions of choices in the direction of overweighting of small probabilities. Again, there was no main effect of Choice Format ($F(1,245) = .965$, $p = .327$), no main effect of Sampling Order ($F(4,245) =$

.153, p =.961) and no significant interaction (F(4,245) = 1.107, p =.354). The interaction plot in Figure 4.8 shows how close the means for both choice formats lie together just below the 50% line across the different order conditions.

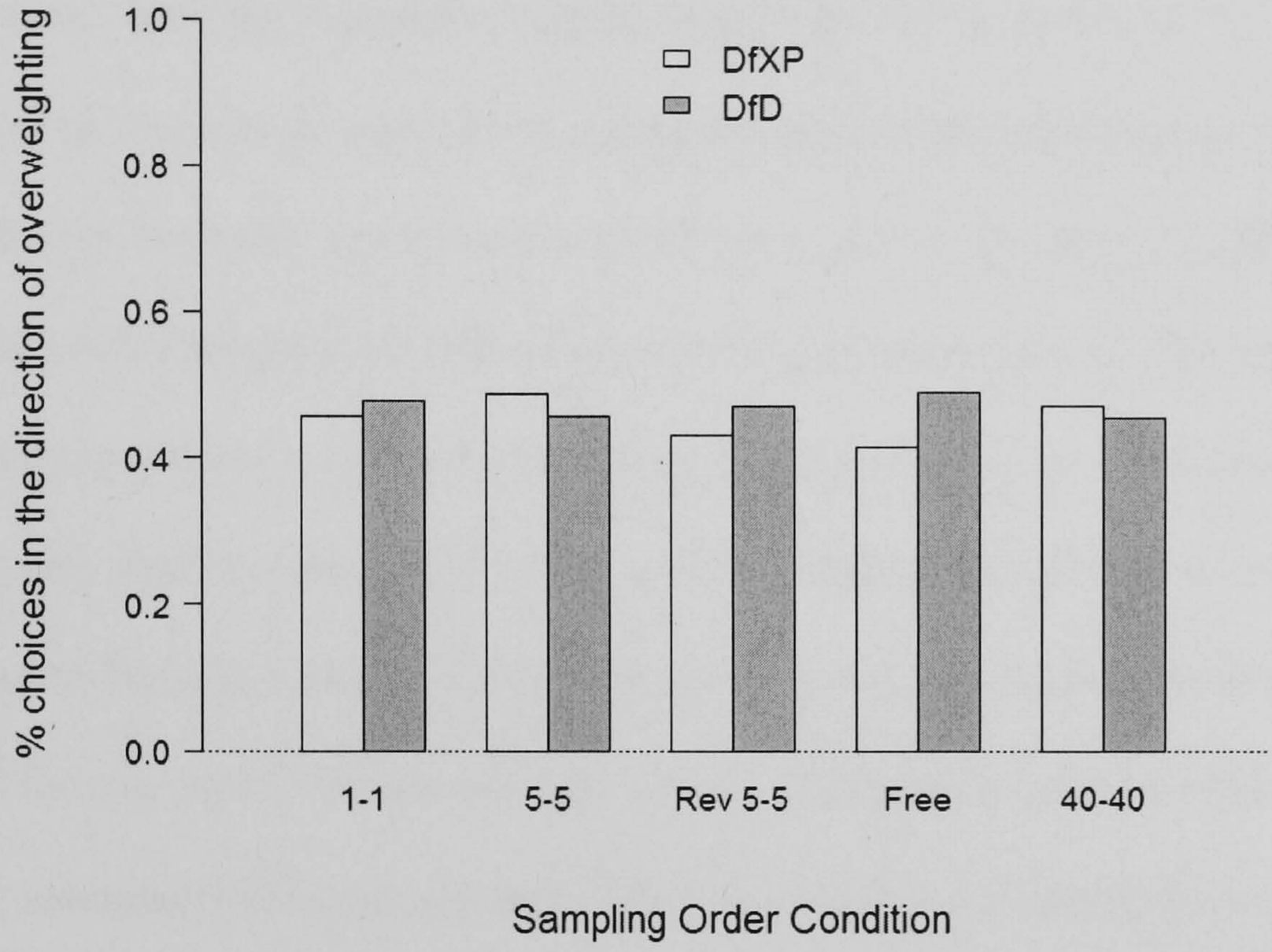


Figure 4.8. Interaction plot for the mean choice proportions within the different order conditions with different degrees of partitioning into sub-samples in Experiment 7.

As in the previous experiment, the underlying reason for the similarity of the choice behaviour in both formats seems to be reduced overweighting within the descriptive choice problems. Again, this was confirmed in a comparison of the proportions of choices in the direction of overweighting between the descriptive choice conditions of Experiment 2 (61%) and Experiment 7 (47%), which revealed that there was significantly less overweighting within the descriptive choice problems of Experiment 7, t(273) = -3.097, p =.002. The experiential choice data in Experiment 7, on the other hand, exhibited proportions of overweighting similar to the ones observed in Experiment 2, t(273) = .655, p

=.513. Furthermore, we find a significant reversal in the direction of underweighting under DfXP when comparing the experiential choice data from Experiment 7 with the descriptive choice data from Experiment 2 ($t(273) = -3.221$, $p = .001$), but not when comparing the descriptive choice data from Experiment 7 with the experiential choice data from Experiment 2, $t(273) = 1.122$, $p = .263$. The description-experience gap for the experiential choice data could therefore only be replicated between experiments, as both formats in Experiment 7 provide choice proportions in the direction of less overweighing. The hypothesis in the last experiment was that this might be due to the mixed presentation of the two formats. This does not apply to the current design as the formats were presented in separated blocks. What remains, then, is a potential effect of the order of the two blocks. However, there was no evidence for such an effect. Instead, contrasts performed as part of the mixed ANOVA reported above comparing the 5-5 Order Condition with the Reversed 5-5 Condition, revealed that there was no significant difference between the proportions of choices in the direction of overweighting within these two conditions.

The absence of any effect of sampling order on the proportions of choices in the direction of overweighting was also found on the level of individual choice problems (see Table 4.11). The only exceptions were choice problem 3 under experiential choice and choice problem 4 under descriptive choice.

TABLE 4.11

Choices in the direction of overweighting under Matched Sampling

	Decision Problem	1-1	5-5	Reversed 5-5	Free Sampling Order	40-40	$\chi 2$	p
DfXP	1	34	40	42	38	50	2.915	.572
	2	76	56	68	58	66	5.717	.221
	3	38	52	20	38	42	11.375	.023
	4	54	54	48	36	42	6.453	.168
	5	52	50	40	36	38	4.336	.362
	6	42	40	42	42	44	.164	1.0
DfD	1	74	58	56	60	60	4.295	.368
	2	50	58	52	70	62	5.335	.255
	3	34	40	24	32	40	3.922	.417
	4	62	54	84	58	60	11.991	.017
	5	38	30	30	40	20	5.811	.214
	6	28	34	32	34	30	.629	.96

However, a comparison of the data with all order conditions combined shows that there were significant differences between the proportions of choices in the direction of overweighting between the two formats on the level of individual choice problems. This stands in contrast to the results on the comparisons of the means across all six choice problems mentioned above. A closer look at Table 4.12 allows us to explain the coexistence of both findings, though. In four of the six choice problems the choice proportions in the direction of overweighting were even lower under descriptive choice than under DfXP. In two problems this difference was significant (see problems 5 and 6).

TABLE 4.12

Combined choice data for the individual choice problems including the p-values for the McNemar tests conducted

	Choices in the direction of overweighting		
Decision Problem	Descriptive Choice	Experiential Choice	*p*
1	62	41	.001
2	58	64	.18
3	34	38	.353
4	64	43	.001
5	32	43	.003
6	32	42	.018

In choice problems 1 and 4, on the other hand, we find the typical gap between of choices in the direction of underweighting for DfXP and choices in the direction of overweighting under descriptive choice. Taken together, both results on the level of individual gambles may reconcile the absence of a difference for the mean proportions across choice problems.

4.4.2.3 Within-participant reversals

The overall rate of within-participant preference reversals between choice formats in one direction or the other across the six choice problems was 42%. An inspection of the Table 4.13, showing the breakdown of the reversals across the different choice problems and their split into the two categories of DfXP and Non-DfXP reversals, again provides inconclusive results. Similar to the results reported previously there were significantly more reversals in the direction of DfXP observed within choice problems 1 and 4. However, I also found significantly more reversals in the opposite direction for choice problems 5 and 6.

TABLE 4.13

Breakdown of the observed proportions of preference reversals for the six choice problems including the result from the McNemar's test on the differences between the proportions

Choice problem	H	L	% of observed reversals	% of reversals in direction of DfXP	% of reversals in direction opposite of DfXP	McNemar's $\chi 2$	p
1	4,.8	3,1.0	46	73	27	24.426	.001
2	4,.2	3,.25	44	43	57	2.064	.151
3	-3,1.0	-32,.1	38	45	55	1.064	.302
4	-3,1.0	-4,.8	47	73	27	24.009	.001
5	32,.1	3,1.0	37	34	66	9.043	.003
6	32,.025	3,.25	41	38	62	6.068	.014

4.4.2.4 Recency weighting

The potential impact of recency weighting on choice was examined in the same way as is in Experiment 6 using a set of mixed ANOVAs comparing the mean percentage of correct predictions based on the expected values of different sequential splits (see Table 4.14) across the five order conditions. For the 20-20 split there was no significant main effect of the sequence ($F(1,245) = 3.43$, $p = .065$), no significant main effect of Sampling Order ($F(1,245) = 1.382$, $p = .241$), and no significant interaction between the two ($F(4,245) = .06$, $p = .993$).

TABLE 4.14

Mean percentages of correct predictions based for the different splits across the three order conditions

Sampling Order	Sequence Split					
	First 20	Last 20	Quant. 1 (1-10)	Quant. 2 (11-20)	Quant. 3 (21-30)	Quant. 4 (31-40)
1-1	.54	.50	.52	.50	.48	.54
5-5	.51	.46	.47	.54	.51	.50
Reversed (5-5)	.49	.45	.54	.45	.43	.49
Free (matched)	.54	.49	.61	.46	.49	.48
40-40	.52	.49	.54	.52	.54	.43
All	.52	.48	.54	.50	.49	.49

The mixed ANOVA using the quartiles provided a significant main effect of the Sequence Splits, $F(3,245) = 3.05$, $p = .028$. It also showed a significant interaction effect between the Sequence Splits and Sampling Order ($F(4,245) = 2.54$, $p = .003$), indicating that the occurrence of a primacy effect that depending on the sampling order during exploration. The inspection of the interaction plot (Figure 4.9) shows that the rate of correct predictions based on the expected values derived from the outcomes of quartile 1 were higher than in the later quartiles and that this difference is more pronounced in the Matched-Sampling Condition with free sampling order, in the Reversed 5-5 Condition and in the 40-40 Condition. A significant main effect of Sampling Order itself was not found, $F(1,245) = 1.07$, $p = .372$.

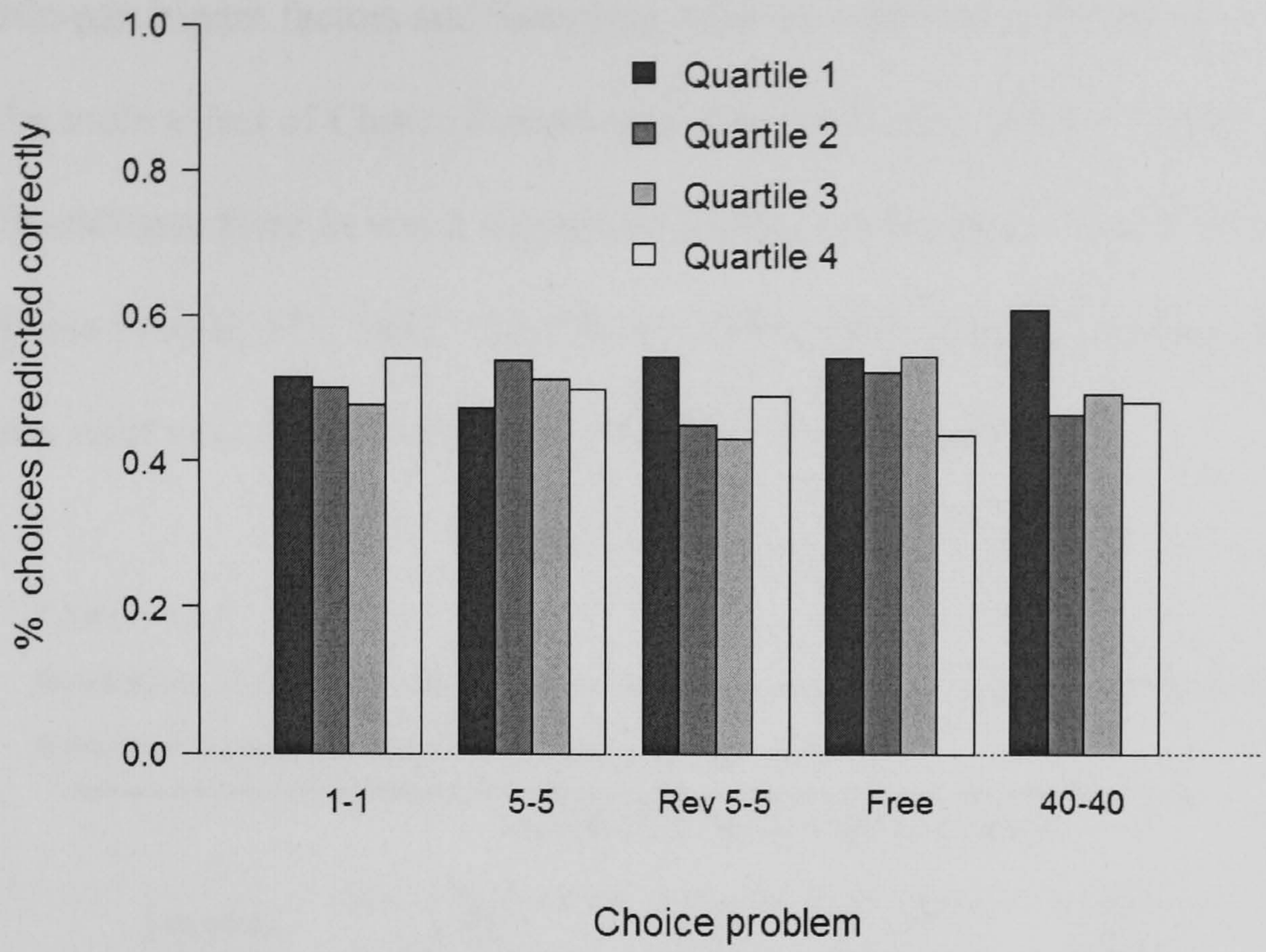


Figure 4.9. Interaction plot for the mixed ANOVA on the quartile splits.

The analysis of the proportions of maximising choices conditional on the position of the rare event within the experiences samples of choice problem 6 provided different results. Here, the percentage of maximising choices was higher for the cases in which the rare event occurred in the first half of the sequence than for encounters in the second half (50% vs. 32%, Fisher's exact $p = .007$). This means that encountering the rare event of 32 points in the last 20 outcomes from that option coincided with a higher rate of choosing the other option with a sure outcome of 3 points. The same was observed for the first or last second seven outcomes of the sequence (53% vs. 29%, Fisher's exact $p = .029$).

4.4.2.5 Application of descriptive choice models

In order to assess whether the descriptive validity of the EV and PT model vary across the different formats and order conditions another mixed ANOVA was conducted with Type of Model (EV vs. PT) and Choice Format (DfD vs. DfXP)

as within-participant factors and Sampling order as a between-participant factor. Only the main effect of Choice Format was significant, $F(1, 245) = 12.367$, $p = .001$. In addition, there as was a significant interaction between Type of Model and Choice Format, $F(1, 245) = 46.508$, $p < .0001$. Both findings are illustrated by the mean rates of correct predictions in Table 4.15 and Figure 4.10.

TABLE 4.15

Breakdown of the mean rates of correct predictions of the two models across choice formats and sampling order conditions for Experiment 7

Sampling Order	Percent of correctly predicted choices			
	EV (Descriptive Choice)	EV (DfXP)	PT (Descriptive Choice)	PT (DfXP)
1-1	.36	.57	.48	.46
5-5	.42	.51	.46	.49
Reversed (5-5)	.32	.46	.47	.43
Free (Matched)	43	.50	.49	.41
40-40	39	.50	.45	.47
All	.38	.51	.47	.45

The interaction shows that the model predictions based on expected value maximisation provide better predictions than PT in experiential choice and that PT can account better for choices under descriptive choice tasks. The former finding is also in agreement with the results reported in the previous chapters. None of the models performs at above chance level. This is rather unusual, especially for descriptive choice data.

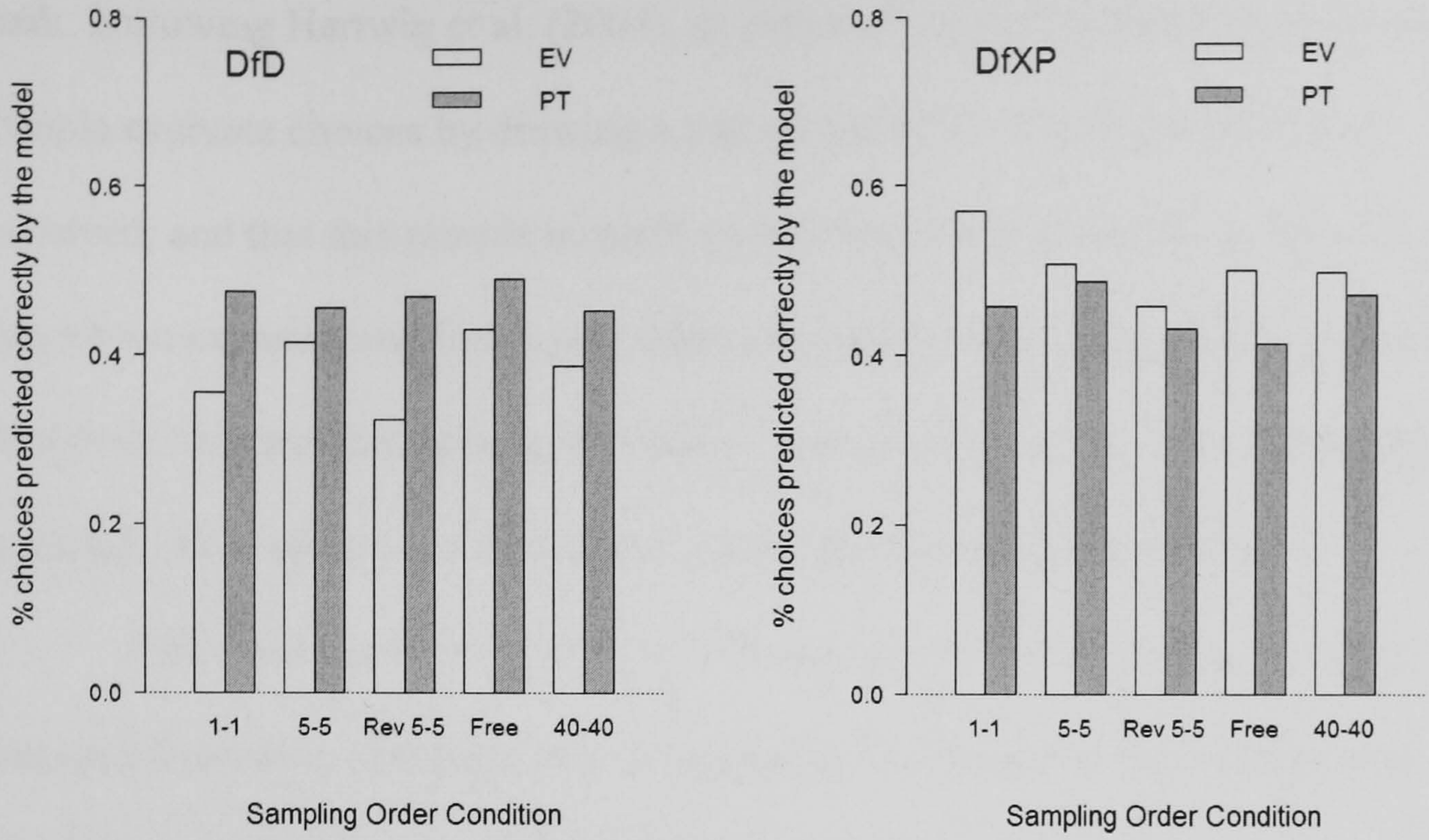


Figure 4.10. Interaction plots for the rates of correct model predictions in Experiment 7.

4.4.3 Discussion

Although there was no evidence for an interaction between the two blocks of choice formats a significant order effect was still not observed. Instead, apparent underweighting in DfXP was found across all the different sampling order conditions. The consistency of this finding suggests that the partitioning due to sampling order cannot explain the differences in choice behaviour between DfD and DfXP. On the other hand, there might be a difference between partitioning the sequence independently and being forced to sample a sequence according to an externally fixed schedule. Whereas in the former case the sub-samples are a result of different hypothesis testing strategies, the schedules in the latter case might not be able to induce similar processes. Interestingly, both modes lead to similar DfXP choice patterns. More important is the finding that it is not possible to eliminate choices in the direction of underweighting by fixing the sequence in a way that facilitates the separate evaluation of the options as it was done in the 40-40 sampling order which is structurally more equivalent to a descriptive choice

task. Following Hertwig et al. (2004), an explanation for this result could be that people evaluate choices by drawing a mental 'sample' of the input they have received; and that this sample is much smaller than the full sample to which they have been exposed, and hence may often not contain low probability events at all. However, this would contradict the observation, shown earlier, that participants are well aware of the overall frequencies within experienced samples.

Also puzzling is the finding of reduced overweighting in the descriptive choice proportions of the last two experiments. This could be due to increased noise in the data or due to specific differences in the characteristics of the demographics of the populations of participants. In the case of the last two experiments, the average age was slightly higher with 32 years which could coincide with differences in risk preference.

The most consistent finding across all seven experiments has been the absence of any form of overweighting of recent items. However, Experiment 7 has provided some indication for a primacy effect. Given the absolute rate of correct predictions observed, this effect does not exceed the predictive power of the sequence as a whole. Unusual are also the findings from the analysis of choice problem 6 as they indicate that the risky option with the rare event is avoided more often when participants have experienced the rare event more recently. This also contradicts Hertwig et al.'s assumption that people behave as if they underweight small probabilities as a consequence of not encountering the rare event recently. Instead, the results of Experiment 7 point in the opposite direction, indicating that people chose as if they underweight small probabilities when they have actually seen the rare event more recently, which is counterintuitive. One potential interpretation of this phenomenon could be that participants have

specific assumptions regarding the underlying sampling mechanism which makes it more likely for them to assume the reoccurrence of the event after a series of non-occurrences than after a recent encounter. I will pick up this idea again in the discussion of a run-based model in Chapter 6.

Additional confirmation of earlier findings could also be obtained with regard to the applicability of descriptive choice models. The accumulated results seem to indicate that the superiority of the PT model in terms of its predictive power usually found under descriptive choice does not hold under experiential choice in which expected value maximisation seems to be able to account for the data slightly better but still unsatisfactorily low.

Finally, although the examination of the within-participant reversals did not provide conclusive evidence regarding the description-experience gap within participants, the results from Experiment 6 and7 have shown consistent evidence for DfXP reversals in choice problems 1 and 4. Taken together, both reversals constitute a reversed reflection effect under DfXP which was observed in 12% of the participants in Experiment 6 and in 15% of the participants of Experiment 7. This way of analysing the data, focusing on the comparison across different choice problems has been omitted so far, but the following chapter will be dedicated to provide a re-analysis of the data collected so far in terms of established choice paradoxes like the reflection effect.

CHAPTER 5

ANALYSIS OF COMMON DECISION MAKING BIASES

5.1 Introduction

In all the previous chapters I have examined the choice behaviour within individual decision problems between decisions from description and various experiential choice formats, following the approach of Hertwig et al. (2004). This has mainly been done in the form of between-subject comparisons, with the exception of Chapter 4, which also provided a within subjects analysis of differences between DfD and DfXP. For example, I have compared choice proportions for the choice "100% chance of 3 or 80% chance of 4" between description and experience formats. However, this is not the only way that this data can be examined. Another approach is to look at specific biases based on preference reversals within sets of choice problems that have been observed in descriptive choice tasks, following Gottlieb et al. (2007). In the introduction to this thesis I reviewed a variety of such decision making biases, including the common ratio effect, the common consequence effect, and the reflection effect. Continuing the above example, in DfD, the common ratio effect is a reversal in preference for 100% chance of 3 over 80% chance of 4 to a preference for a 20% chance of 4 over a 25% chance of 3. Will this reversal also be observed under DfXP? Furthermore, the choice problems used to examine these biases allow one to draw conclusions regarding specific properties of the value- and probability weighting function within the PT framework. Investigating such biases under experiential choice could provide further insights on the causes of the differences

between the two formats and clarify whether these are due to different properties of the weighting function, as has been suggested in the literature. Consequently, this chapter aims to provide such a re-analysis of the data obtained in the experiments presented earlier. By rearranging the six choice problems used, two pairs of common-ratio effect problems and two pairs of reflection effect problems can be formed. Common-consequence problems were not included in the original Hertwig et al. (2004) set that I have been using, and can therefore not be investigated.

5.2 The common ratio effect

The interesting property of common ratio type problems is that they allow for a direct examination of non-linearity of the probability weighting under PT. To illustrate how underweighting or overweighting of probabilities is inferred from people's choices, consider the impact of underweighting or overweighting of small probabilities on the choice between two lotteries, (A_1) a .8 chance of £4 and a .2 chance of winning nothing, or (B_1) £3 for sure. If we divide the probabilities by a common factor of 4 we get the resulting lotteries A_2 (.2 chance of £4 otherwise nothing) and B_2 (.25 chance of £3 otherwise nothing). From a normative point of view, the division should be irrelevant: The preference for A or B should be the same in both cases, because the second lottery can be viewed as identical with the first, except that it has a ¾ chance of being 'called off'. As the values are the same, any change in preferences between the prospects must be due to decision weights not scaling linearly with probabilities. Therefore, the observation that people prefer B_1 over A_1 and A_2 over B_2 can be explained by an inverse S-shaped weighting function in which the ratio between the weights of .2

and .25 may be higher than between .8 and 1.0. That is, $w(.20)/w(.25) \geq w(.80)/w(1.0)$. This is prospect theory's explanation of Allais' (1953) paradox. In other words, in the PT framework the choice behaviour within this problem allows us to make inferences regarding the slopes of the nonlinear weighting function between particular points which, in the case of DfD, implies that the function is steeper between the weights of .8 and 1.0 than between the weights of.2 and .25. The traditional interpretation of this result is an inverse-S-shaped weighting function, though there are other - perhaps more contrived - functional forms that would also have this property.

As pointed out earlier, behaviour in experiential choice tasks has been shown to deviate from this descriptive choice pattern and complete reversals of this bias have been reported. Barron and Erev (2003), for example, found a mirrored common ratio effect which underlined their hypothesis of underweighting of rare outcomes in small feedback-based decisions. Hertwig et al. (2004, 2006) replicate a reversed common-ratio effect under DfXP which they explained by an underrepresentation of the rare events and recency weighting. Using the example provided above (4, .8; 3, 1.0), the underrepresentation of the probability.2 of receiving 0 will make the riskier option more attractive, resulting in a reversed common ratio effect. Following the logic of the common ratio effect within the PT framework, such a reversal of the effect would imply that $w(.20)/w(.25) < w(.80)/ w(1.0)$, thereby constraining the weighting function in the opposite way than under DfD with a function steeper nearer 0 than nearer 1. If the underrepresentation is the only reason for the reversal then it should not be observed in experiential choice problems where sampling error is eliminated. In a very recent paper Gottlieb et al. (2007) have examined the effect of different

presentation formats of uncertainty information on common ratio and common consequence type decision making biases. One of the presentation formats implemented was an experiential sampling task, similar to the Matched Sampling paradigm described earlier, involving decks of cards. In another format participants were provided with the actual percentages of the outcomes within the decks resembling a descriptive choice task. Instead of the inverse common ratio effect reported by Hertwig et al. (2004), Gottlieb et al. (2007) found the usual common ratio effect within their experiential format. However, they failed to replicate the common ratio effect in the descriptive choice task (an extremely well replicated effect) which prevents a comparison of the description and experience formats and thus makes it hard to interpret their data.

Throughout the earlier chapters, I have been able to show preference reversals within different individual choice problems under conditions where sampling error is eliminated. Taken together these effects could imply reversed common ratio effects in contrast to the findings of Gottlieb et al. (2007). Further clarification regarding the existence of common ratio effects under experiential choice with and without inherent sampling error will therefore be provided in the following section in which the presented data is re-analysed in the context of the common ratio effect.

In order to do so, four of the six choice problems used across the experiments presented in Chapters 2 to 4 were regrouped into two pairs. The first pair consists of Problems 1 (4,.8; 3,1.0) and 2 (4,.2; 3,.25) and the second pair is based on Problems 5 (32, .1; 3,1) and 6 (32,.025; 3,.25). In both pairs the second problem can be generated by dividing the probabilities within the first problem by four. A summary of both problems is provided in Table 5.1.

TABLE 5.1

Common ratio problems used in the reported experiments

	High probability context		Low probability context	
	High risk	Low risk	High risk	Low risk
CR1	4,.8	3,1.0	4,.2	3,.25
CR2	32,.1	3,1.0	32,.025	3,.25

The data were analysed separately for each experiment and pooled together across experiments wherever the same experimental conditions were implemented in order to increase the power of the tests. As different numbers of gambles were presented within the experiments the equivalence of the observed choice proportions within the common ratio problems was analysed using either McNemar's Chi^2 test (multiple choices within subjects) or Fisher's Exact test (single choice between subjects).

5.2.1 The common ratio effect under descriptive choice

As pointed out earlier, the common ratio effect under descriptive choice is denoted by people preferring the riskier option more often in the low probability context of the problem (p/x) than in the high probability context (p). In the Problems 1 and 2, for example, the majority of the participants have been shown to prefer (3,1.0) when presented together with (4,.8), but chose (4,.2) when presented together with (3,.25). Figure 5.1 presents the proportions of riskier choices within the two common ratio problems (CR1 and CR2) from the experiments deploying descriptive choice conditions (Experiment 2 and 7). As both designs demanded participants to make choices within all choice problems,

McNemar's Chi^2 test was used to determine whether the two proportions within problems were equivalent. For comparison, the choice proportions from other experiments involving gamble descriptions (Hertwig et al., 2004; Kahneman & Tversky, 1979) have been included where available. As the raw data for this external data were not available no inferential statistics could be computed.

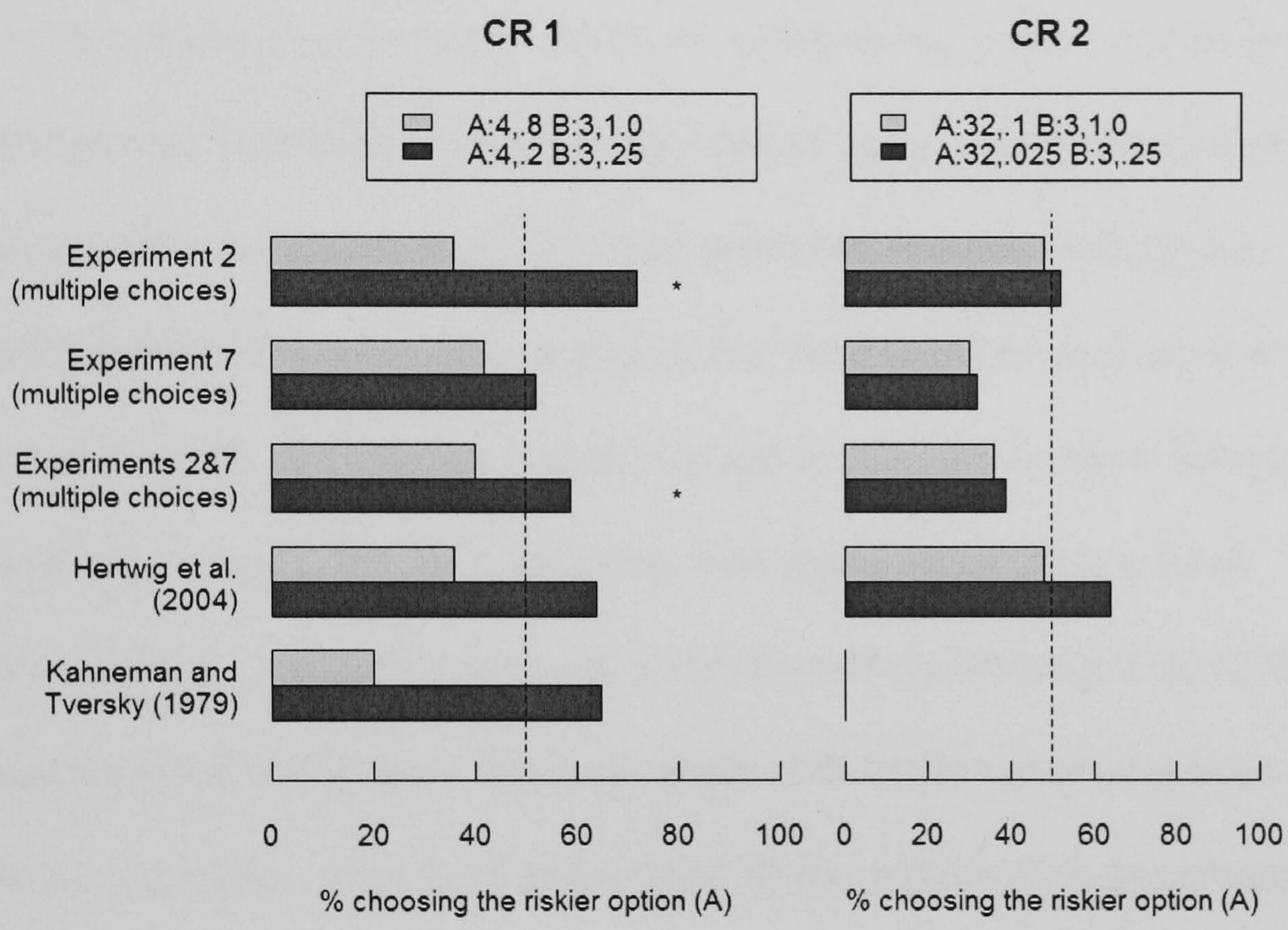


Figure 5.1. Proportions of riskier choices within the two common ratio problems under descriptive choice across the different experiments. Significant differences are highlighted by asterisks

Within the first set of problems (CR1) the typical common ratio effect could be replicated, consisting of a higher percentage of H choices in the low probability context. Significant reversals crossing the 50%-line were found for Experiment 2 and for the data set pooled across experiments. However, in the second set of problems (CR2) no significant differences between the two choice proportions were found.

5.2.2 The common ratio effect under Free Sampling

A different pattern was observed for the same set of problems within the Free-Sampling Condition, the original DfXP design used by Hertwig et al. (2004) in which participants are allowed to draw with replacement from the underlying distribution as often as they wanted. As we have seen earlier, due to the underrepresentation of rare events in this task, preference reversals in comparison to descriptive choice are commonly observed. Consequently, this can also lead to reversed common ratio effects with the risky option being chosen more often in the context of the sure alternative than when presented together with the low probability options. Figure 5.2 shows that such a trend could be replicated within CR1 across all the Free-Sampling Conditions and the Comprehensive-Sampling Condition, which is also based on sampling with replacement using a fixed number of samples. For comparison, the proportions from Hertwig et al. (2004) have been included in the chart. However, none of the differences have been found to be significant. Also, both proportions lie beyond the 50%-line, indicating that the riskier option is preferred in both contexts.

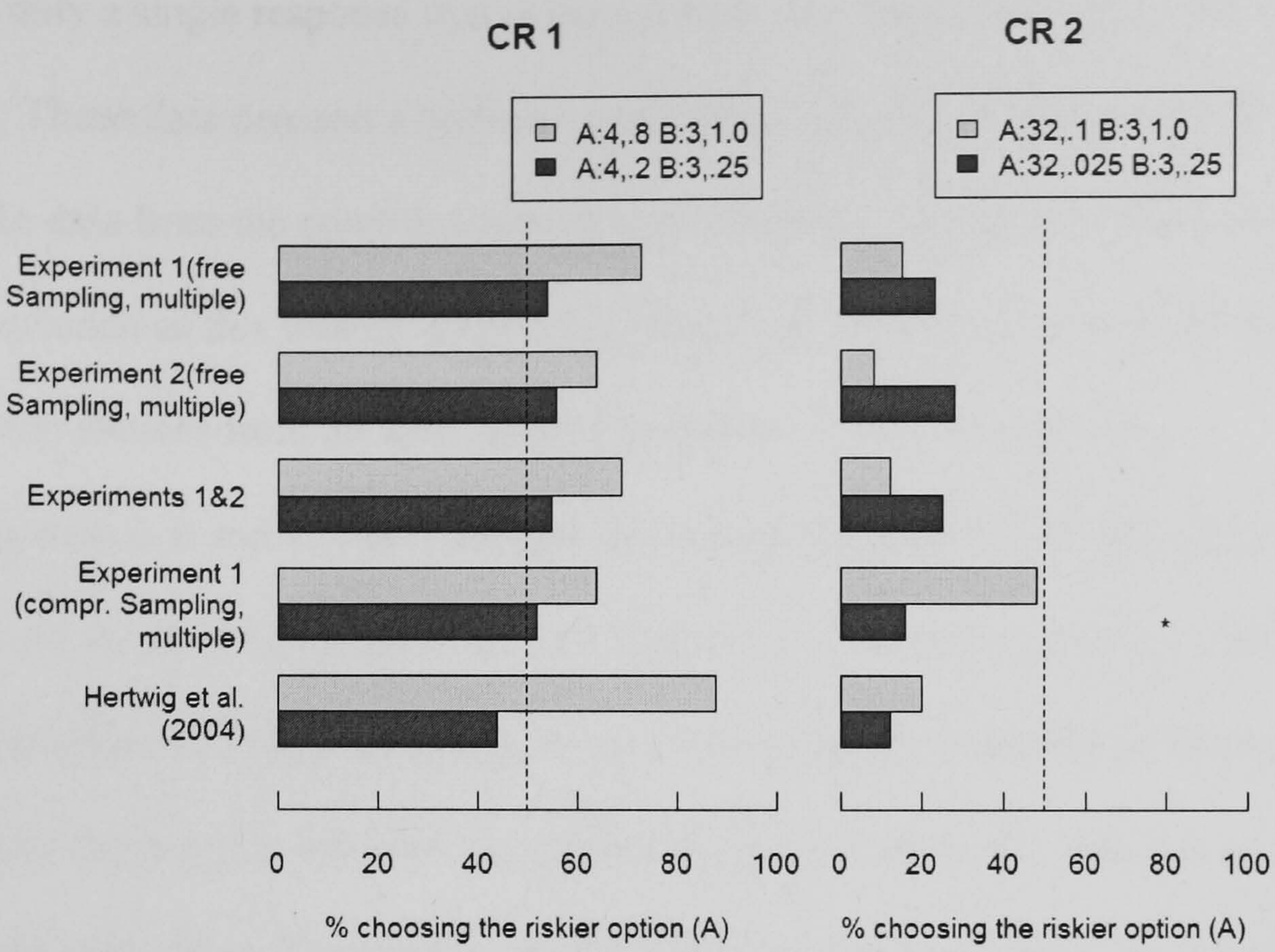


Figure 5.2. Proportions of risky choices for the CR problems in the Free-Sampling Conditions

For CR2 the results are less clear. Across all data sets, including the Hertwig et al. data, the proportion of riskier choices remains below 50%. Within the Free-Sampling Conditions the trend points in the direction of a typical common ratio effect. Under Comprehensive Sampling, on the other hand, a significant reversal of the common ratio effect was observed. Overall, the results seem to indicate a confirmation of earlier results within the Free Sampling paradigm. In order to assess whether this finding is due to the underlying sampling error the findings have to be contrasted with the proportions obtained under Matched Sampling which will be described in the following section.

5.2.3 The common ratio effect under Matched Sampling

For the Matched Sampling design a variety of different data sets has been collected including data from experiments involving frequency estimations for

which only a single response was collected from each participant (Experiments 3 and 4). These data demand a between-participants analysis. For Experiment 4 only the data from the condition asking for probability judgements 'after choice' were included as this was the only design that was comparable with Experiment 3. Data with choices from all four relevant problems were available from Experiments 2, 6 and 7. The results are presented in Figure 5.3. Within CR1 the results are not as coherent as in the previous cases. Across most of the conditions the proportions of riskier choices in both contexts are at a similar level beyond the 50% line. Experiment 2 shows non-significant deviations in the direction of a common ratio effect. However, a significant preference reversal in the direction of a reversed common ratio effect was found in Experiment 3. As in the earlier analysis, the proportions of risky choices in the second pair of common ratio problems were generally below 50%. For the experiments involving all 4 problems there was a trend in the direction of a common ratio effect.

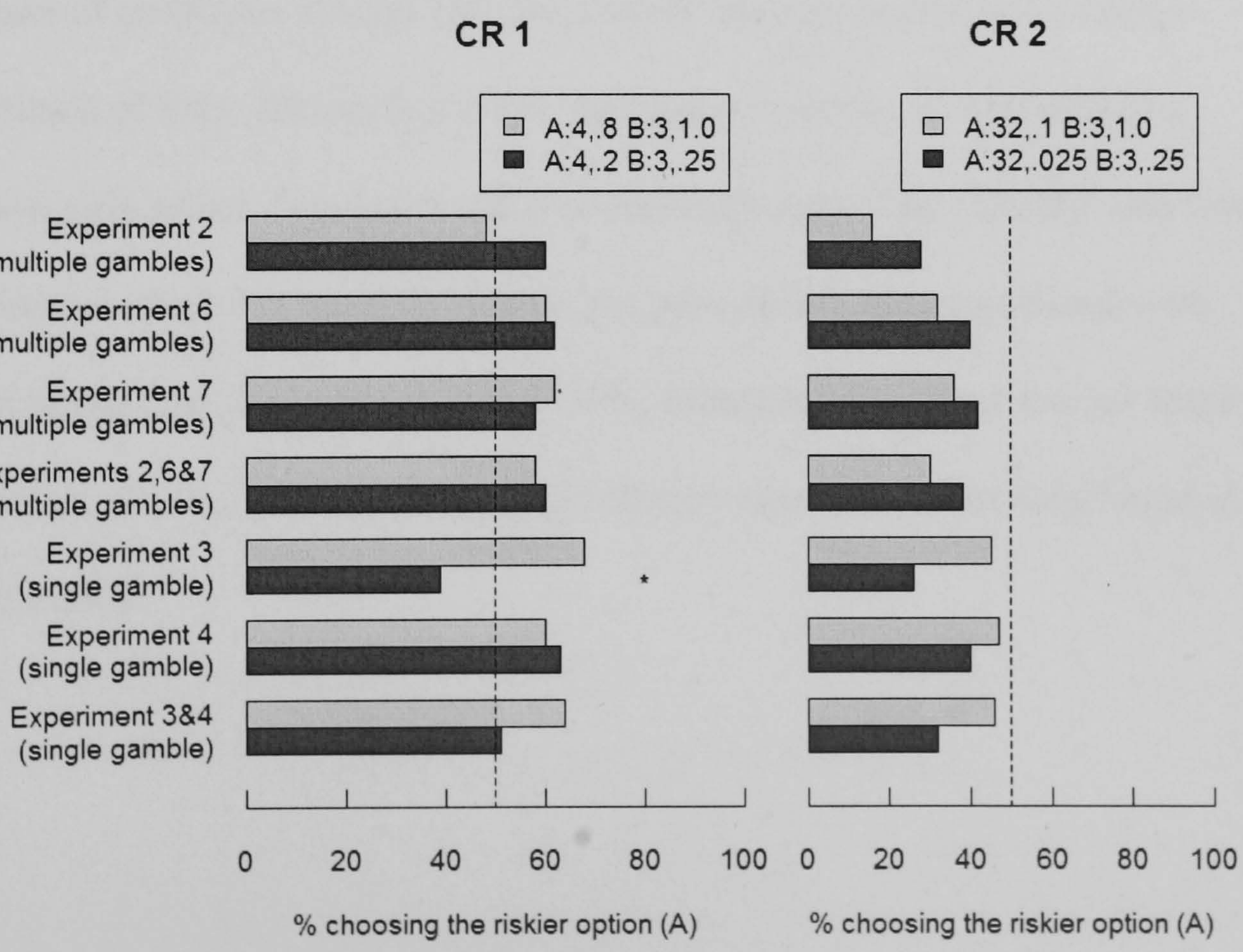


Figure 5.3. Proportions of risky choices for the CR problems under Matched Sampling

The opposite was found for Experiments 3 and 4 which are based on comparisons between the choices of different subjects. None of the differences was significant though. In summary, the data from the majority of the Matched-Sampling Conditions seems to indicate that there is no difference between the proportions of riskier choices in the high and low probability contexts, meaning that there was neither a common ratio effect nor a clear reversal of it.

5.2.4 The common ratio effect under fixed sampling order

A separate analysis was conducted for experiments involving Matched Sampling under a restricted 40-40 sampling order. Experiment 5 provided an experiment with only one choice problem per person whereas Experiments 6 and 7 contributed complete sets of data allowing for within-participant analyses. Figure 5.4 illustrates the results. No significant differences were found for either of the two pairs of problems. Within CR1 the overall trend showed slightly higher proportions of risky choices in the low probability context, conforming to a common ratio effect. No clear trend was observed under CR2. For the data from experiments involving multiple choices per participants the proportions were similar under both contexts and below 50%, indicating that there was no common ratio effect. A trend towards a reversed common ratio effect was only found in Experiment 5.

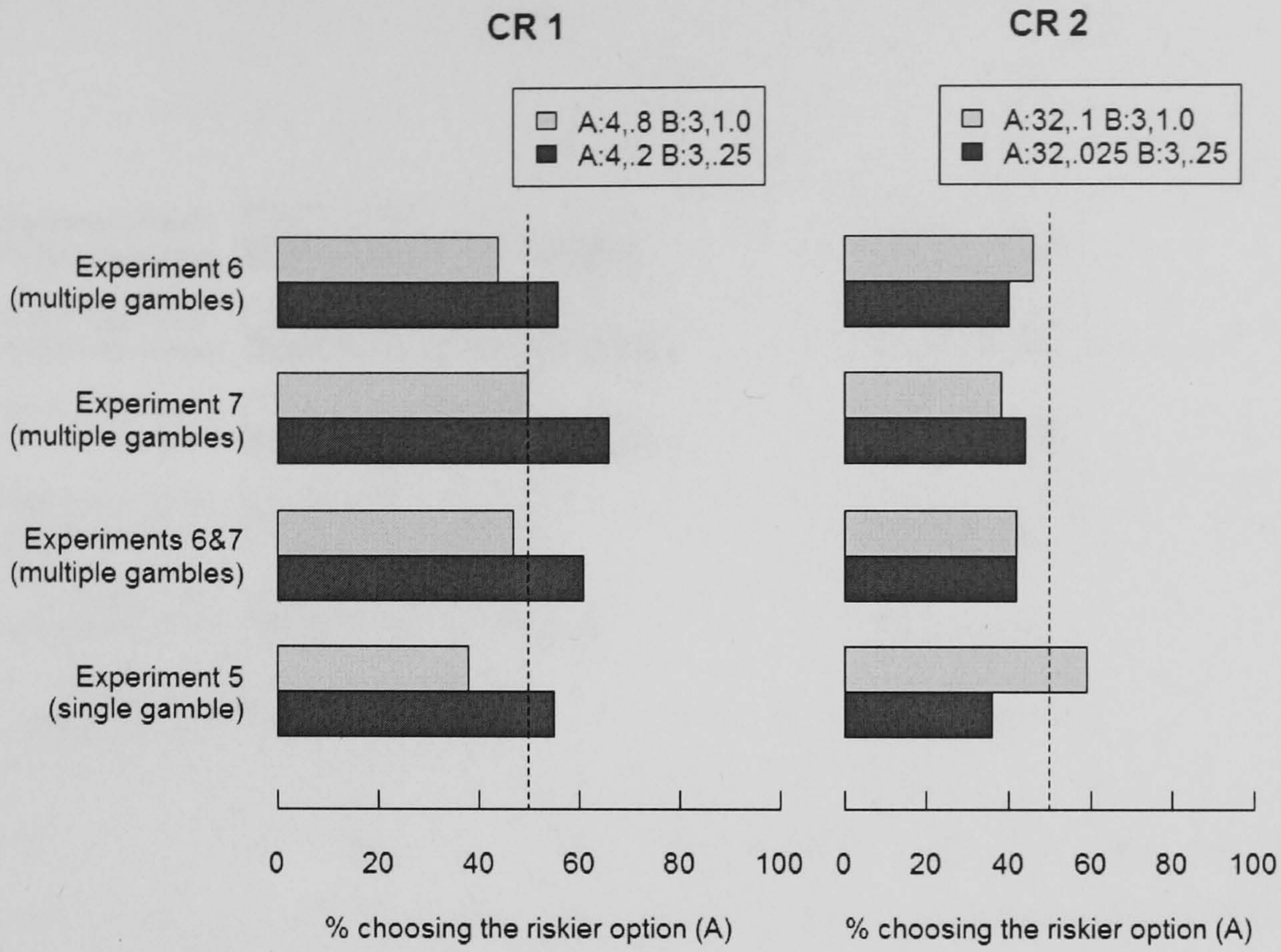


Figure 5.4. Proportions of risky choices for the CR problems under 40-40 sampling order

The results for the remaining sampling order conditions have been summarised in Figure 5.5. Under 1-1 sampling order significant differences in the direction of a common ratio effect were found for problem CR1. For the second problem the results were incoherent within the two different experiments. The pooled data from both experiments shows similar proportions under both contexts. For the 5-5 sampling order condition there was a non-significant difference in the direction of a reversed effect under both problems. As in previous results, the proportions of riskier choices were above the 50% line in CR1 and below 50% in CR2.

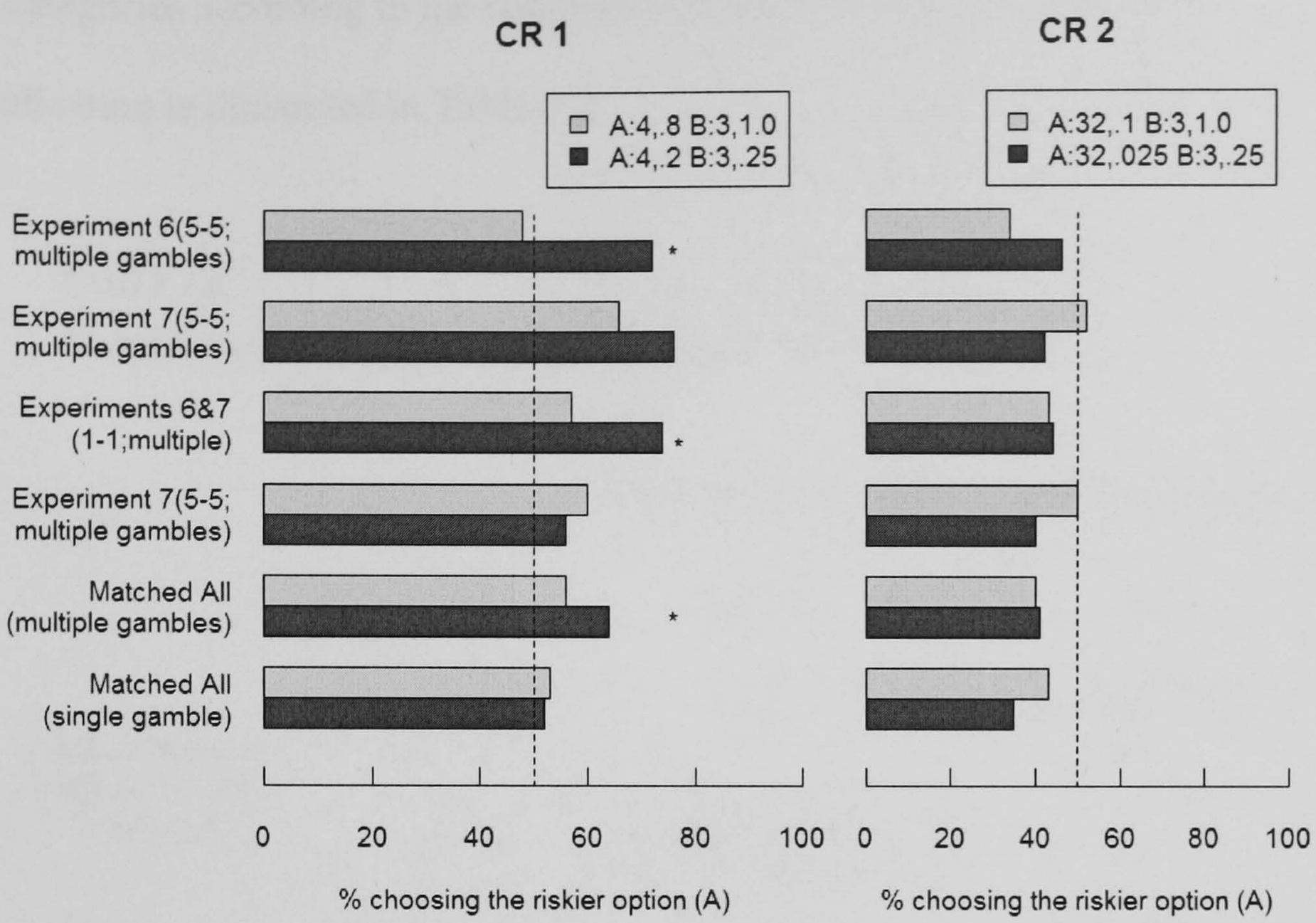


Figure 5.5. Proportions of risky choices for the CR problems under the remaining order conditions and across all Matched-Sampling Conditions (all orders)

Finally, the data from all available Matched-Sampling Conditions, including all different sampling orders, were pooled together. This was done separately for single gamble conditions and multiple gamble conditions. The results of this analysis (also in Figure 5.5) show a significant common ratio effect for the data using multiple gambles in CR1. No difference was found for the proportions from the data based on a single gamble per participant. Under CR2 no significant differences were found.

5.2.5 Within-participants analysis of the common ratio effect

Furthermore, the within- participants data from Experiments 6 and 7 allowed an examination of whether there were any within-participants reversals for the CR problems between DfD and DfXP. Before analysing the data, the two preferences within the choice problems under each choice format were combined to one of

four categories according to the common ratio pattern observed. This first classification is illustrated in Table 5.2.

TABLE 5.2

Classification for the CR problems, separately for each choice format.

		Choice Problem 2 (Low Probability Context)	
		High risk	Low risk
Choice Problem 1 (High Probability Context)	High risk	HH	Common Ratio Effect (CR)
	Low risk	Reversed Common Ratio Effect (revCR)	LL

With the classification within the two 2 × 2-tables, one for each format, one new 4×4-table was produced, reflecting the agreement of preferences within the common ratio problems under DfD and DfXP. The scheme for classification within this 4×4-table is illustrated in Table 5.3.

TABLE 5.3

Classification for the CR problem agreement between the two choice formats. In the actual analysis only the corresponding non-diagonal cells with similar shades of grey are compared.

CR classification within DfD	CR classification within DfXP: CR	HH	LL	revCR
CR	CR_CR	CR_HH	CR_LL	CR_revCR
HH	HH_CR	HH_HH	HH_LL	HH_revCR
LL	LL_CR	LL_HH	LL_LL	LL_revCR
revCR	revCR_CR	revCR_HH	revCR_LL	revCR_revCR

To test whether this resulting contingency table was symmetric, a generalisation of McNemar's test for $I \times I$-contingency tables with $I > 2$ by Bowker (1948) was used. For this analysis only, the corresponding non-diagonal frequencies are considered, which indicate actual changes in preference. In the example in Table 5.3, this corresponds to the six pairs of cells with similar shades of grey. Following the suggestion of Zwick, Neuhoff, Marascuilo and Levin (1982), I also conducted all six possible individual a priori contrasts for these corresponding cell pairs to identify the causes for a potential asymmetry within the contingency table. Of particular interest was the comparison between the cells with participants showing either a common ratio effect or a reversed common ratio effect within the two formats ('revCR_CR' and 'CR_revCR'). The critical values for these comparisons were obtained from the Bonferroni inequality tables by Dayton and Schafer (1973).

For the data from Experiment 6, the choice behaviour within the common ratio problems was found to be symmetrical, both for CR1 (Bowker's χ^2 (6,

N=150) = 8.616, p = .196) and CR2 (Bowker's χ^2 (6, N=150) = 5.743, p = .453). However, for Experiment 7 the null hypothesis of internal symmetry had to be rejected for CR1 (Bowker's χ^2 (6, N=250) = 32.126, $p < .0001$) and CR2 (Bowker's χ^2 (6, N=250) = 17.774 (6), p = .007). The inspection of the results from the planned comparisons of the six possible 2 × 2 breakdowns of the original table (see Table 5.4.) shows that the found asymmetry for CR1 in Experiment 7 seems to stem from shifts between the pattern of always choosing the riskier option and never choosing the riskier option (Δ_{HH_LL}). More specifically, in nearly all of the cases involving shifts between these two patterns (21 out of 23) participants choose both sure options under DfD and the riskier options under DfXP.

TABLE 5.4

Results for all six non-directional comparisons of the 4 × 4 McNemar Tables for Experiment 7. Significant χ^2 values are highlighted by asterisks.

		Z_{CR_HH}	Z_{CR_LL}	Z_{CR_revCR}	Z_{HH_LL}	Z_{HH_revCR}	Z_{LL_revCR}
Experiment 7	CR1	2.496	1.121	1.091	3.962*	1.177	2.524
	CR2	.5	.686	.47	2.2	1.569	3.087*

For the six non-directional comparisons of the off diagonal cells in the table and their corresponding mirror image counterparts (the cells in Table 5.3 with corresponding shades of grey) with a family wise alpha level of .05, equally divided, the critical value was 2.649 (from the tables by Dayton & Schafer, 1973).

For the second common ratio problem (CR2), significant shifts were observed between the pattern of always choosing the sure option and the reversed common ratio effect (Δ_{LL_revCR}). Similar to finding in CR1, there were more shifts from choosing the sure option in both contexts under DfD to a reversed common ratio effect under DfXP than the other way round (68 times vs. 9 times). Statistically significant asymmetries regarding direct shifts between common ratio effects and reversed common ratio effects (Δ_{CR_revCR}), which were found in the between-participants analysis, could not be confirmed within subjects in either of the two experiments.

5.3 The Reflection Effect

The rationale behind the reflection effect is very similar. The pairs of prospects compared are of the form (*x,p*; *y,q*) *and* (*-x,p*; *-y,q*). This means that the probabilities are the same in both pairs but the signs of the outcomes involved are reversed. According to Kahneman and Tversky (1979), such a reflection of the values around 0 results in a reflection of the preferences between the two prospects, hence the name reflection effect. To provide an example, consider the choice between the first pair of prospects used above, (A_1) a .8 chance of £4 and a .2 chance of winning nothing, or (B_1) £3 for sure. As mentioned earlier, in decision from description, people prefer the sure gain (B_1) over the riskier option (A_1). However, if we now reverse the sign of the values, moving into the domain of losses but keeping the probabilities constant, (A_3) a .8 chance of losing £4 and a .2 chance of losing nothing, or (B_3) a sure loss of £3, people prefer the riskier option (A_3) over the sure loss (B_3). This reversal implies $v(3) > w(.8)v(4)$ and $v(-3) < w(.8)v(-4)$, which is in line with a value function under PT that is concave for

gains and convex for losses, reflecting diminishing sensitivity for both domains with increasing distance from the reference point.

Evidence for this reversal under DfD has been provided in a number of studies (e.g., Abdellaoui, 2000; Schoemaker, 1990; Tversky & Kahneman, 1992). In experiential choice, on the other hand, the reflection effect has not received much attention. As pointed out earlier, the causes for the differences in choice behaviour under DfXP and the apparent underweighting of small probabilities have been mostly attributed to differences in the probability weighting function. Barron and Erev (2003) have provided related results in the context of feedback-based decisions where participants showed a reversed reflection pattern which they referred to as a *reversed payoff domain effect*. An inspection of the raw proportions reported by Hertwig et al. (2004) does not provide a clearly identifiable trend. Within one of the problems there seems to be a reversed reflection effect whereas a second pair of problems exhibits the reflection effect usually observed under descriptive choice. Studies systematically examining the reflection effect similar to the approach used for common ratio problems by Gottlieb et al. (2007) in DfXP are therefore missing.

Consequently, this section will provide a further contribution by providing an initial investigation of the reflection effect patterns and the properties of the value function under DfXP. These results may have ramifications for the interpretation of the differences in choice behaviour found under experiential choice within the PT framework. From the six choice problems employed, two pairs of reflection effect problems (REF1 and REF2) could be extracted, which are summarised in Table 5.5.

TABLE 5.5

The two reflection effect problems in the set of choice problems used

	Gain domain		Loss domain	
	High risk	Low risk	High risk	Low risk
REF1	4,.8	3,1.0	-4,.8	3,1.0
REF2	32,.1	3,1.0	-32,.1	-3,1.0

The analysis was conducted in the same way as the in the context of the common ratio problems, separately and on pooled data wherever possible. Choice proportions from multiple problems were analysed using McNemar's Chi^2 test. Fisher's Exact tests were used for the between-participants analyses involving choices in a single problem.

5.3.1 The reflection effect under descriptive choice

Within the descriptive choice conditions used the reflection effect could be replicated in both pairs of problems (see Figure 5.6.). Subjects preferred the sure option (B) in the set of gains and the riskier option (A) within the set of losses. In the first pair of problems (REF1) all the observed differences stretched across the 50% line. Significant differences were found for Experiment 7 and within the pooled data for Experiments 2 and 7. The only case in which the reflection effect could not be replicated was in Experiment 2 under REF2. Instead, these proportions matched the results of Hertwig et al. which are presented in the same figure.

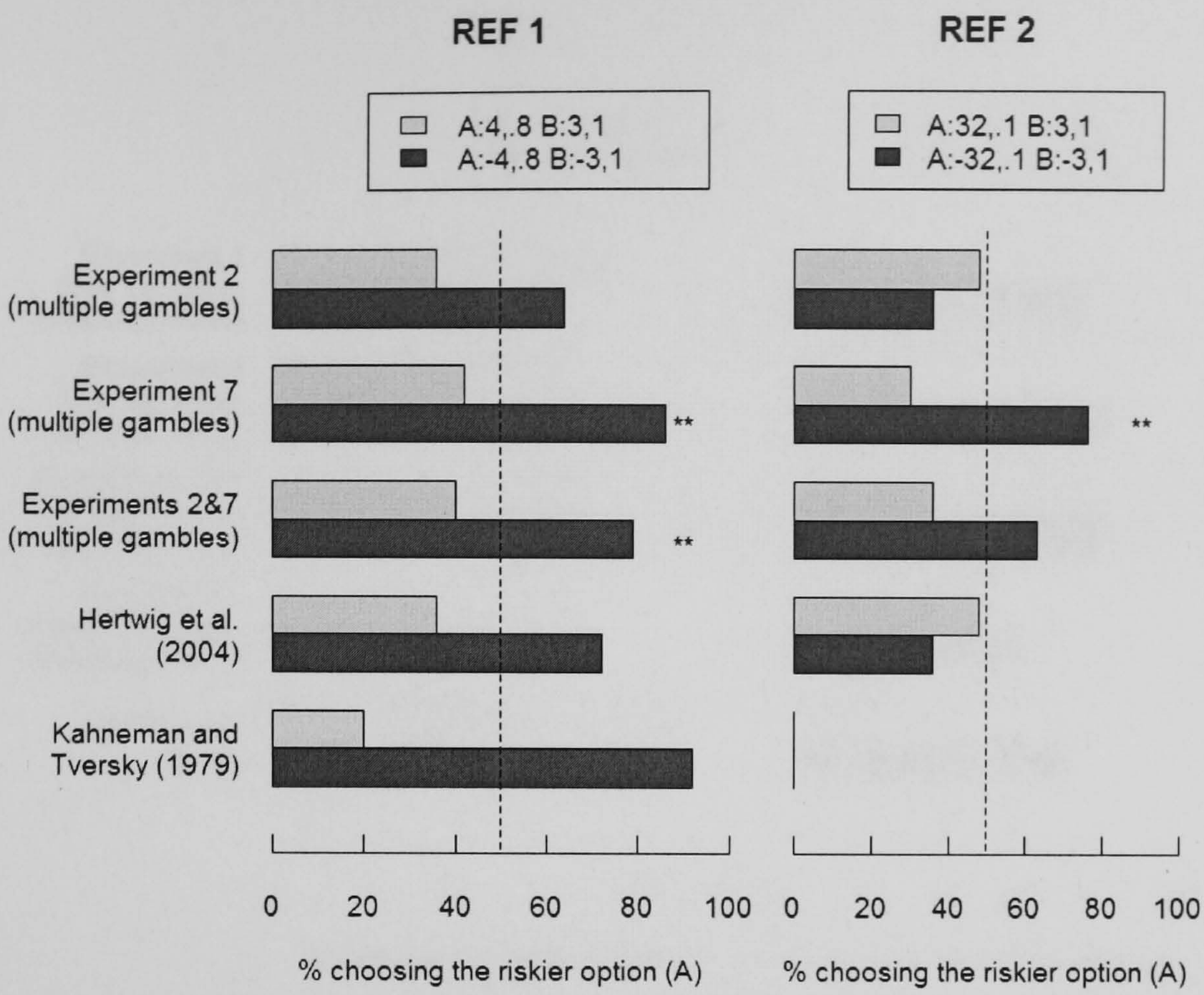


Figure 5.6. Proportions of riskier choices within the reflection effect problems across the different Description Conditions. Significant differences are highlighted with asterisks.

5.3.2 The reflection effect under Free Sampling

The data collected within the Free Sampling paradigms in Chapter 2 and 3 shows a reversed reflection effect with differences across the 50% line for REF1 within all the experiments, including the Comprehensive-Sampling Condition. Statistically significant differences were obtained within Experiment 2 and within the combined data set. However, in REF2 the opposite result, a strong reflection effect was found across all the experiments, again crossing the 50% line indicating full reversals in preferences. The contrasting results obtained in both problems exactly match the results apparent from Hertwig et al.'s proportions.

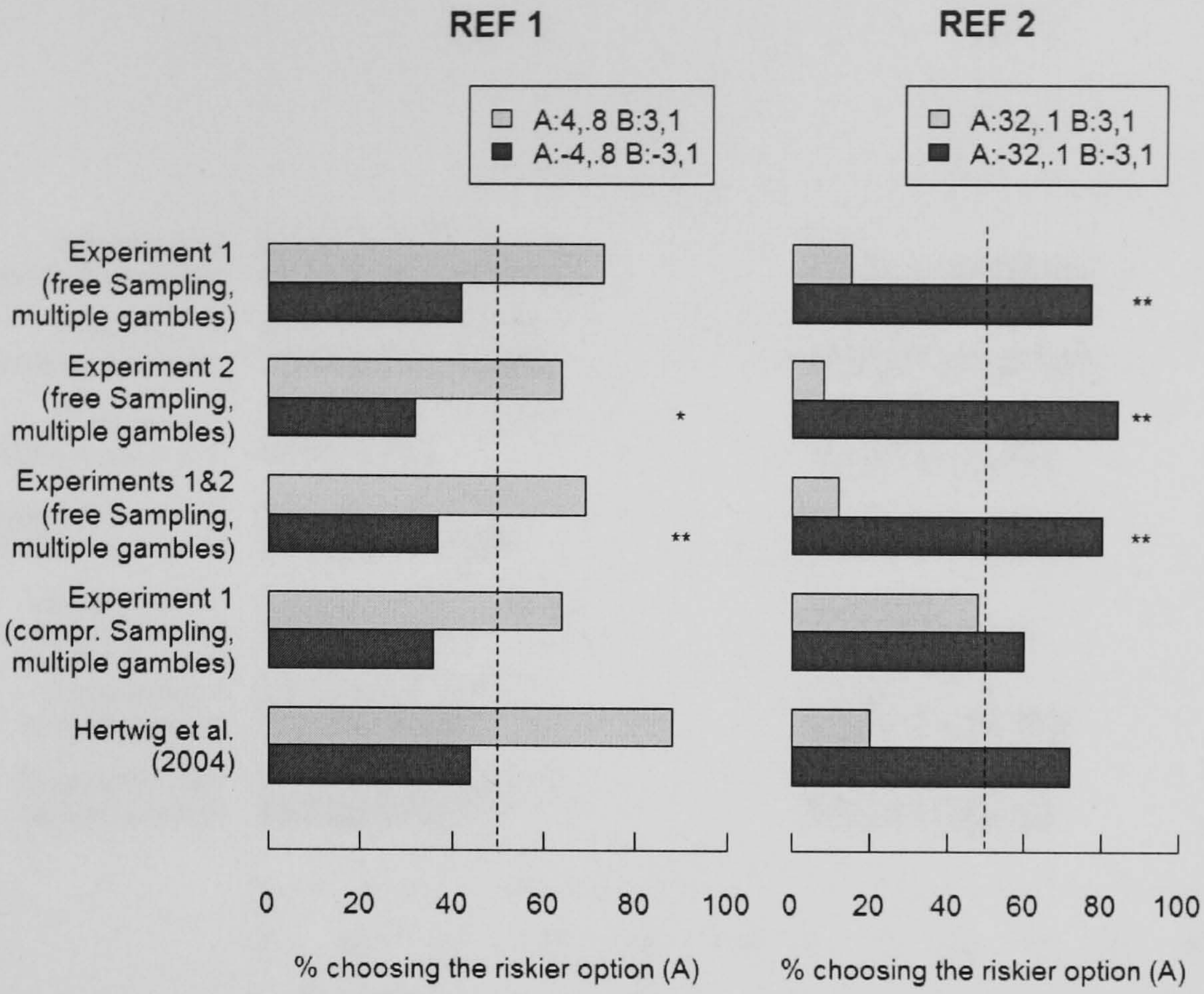


Figure 5.7. Proportions of risky choices for the REF problems under Free Sampling

5.3.3 The reflection effect under Matched Sampling

The same pattern of a reversed reflection effect under REF1 and the usual reflection effect under REF2 was also observed within the Matched-Sampling Conditions (see Figure 5.8). However, within REF1 the reversed trend was only significant for the data from Experiment 7. Experiment 2 was the only case of a trend in the direction of a reflection effect. Identical proportions were observed in Experiment 6. More consistency was found within single gamble conditions (Experiment 3 and 4). The differences for the reflection effect proportions observed in REF2 crossed the 50% line without exception and were significant for all the experiments employing multiple decision problems.

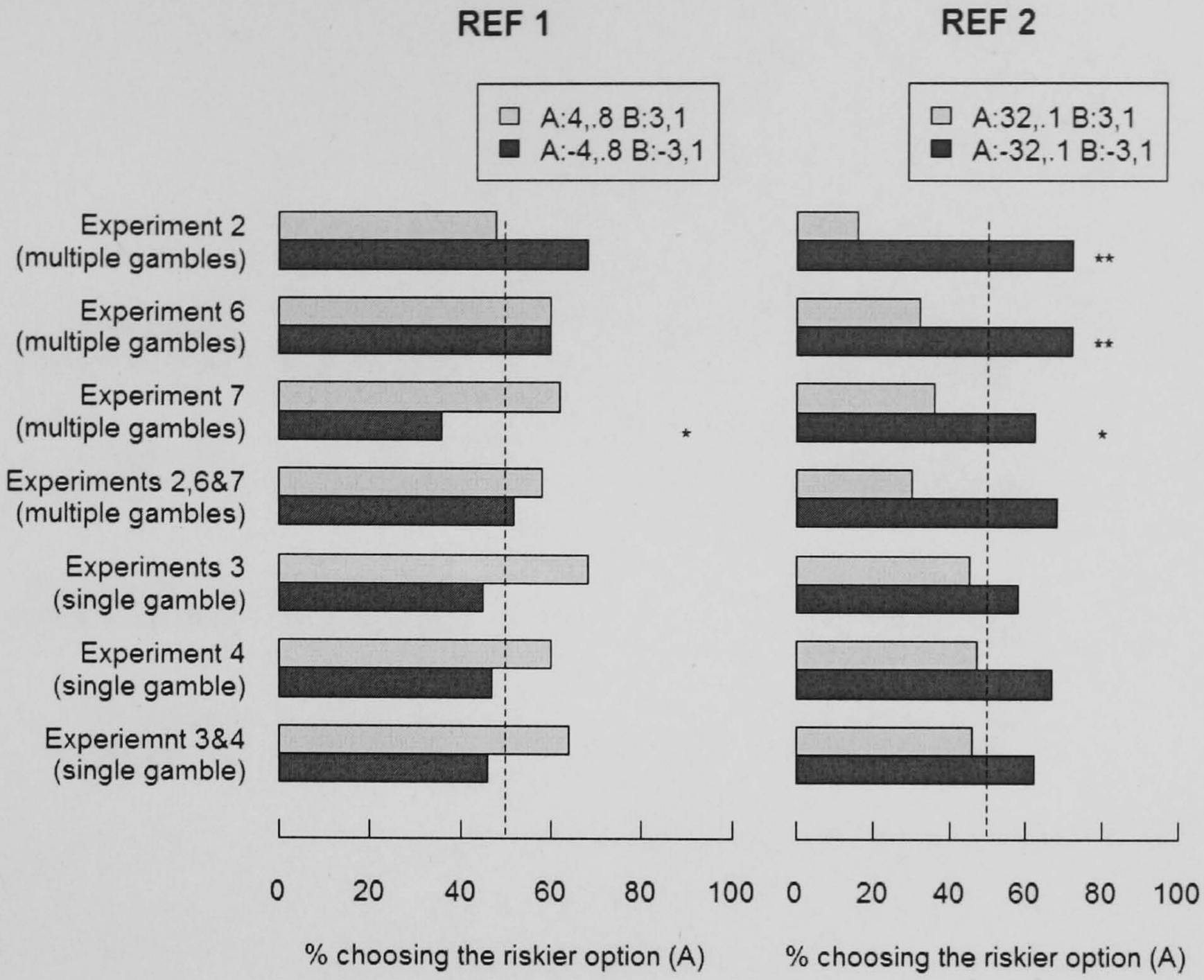


Figure 5.8. Proportions of risky choices for the REF problems under Matched Sampling

5.3.4 The reflection effect under fixed sampling order

As can be seen in Figure 5.9, the two different trends are observed in the conditions with 40-40 sampling order as well. Whereas all choice proportions in REF1 stay below 50%, reversals crossing the line are observed in REF2. The differences are much smaller though and none of them is statistically significant.

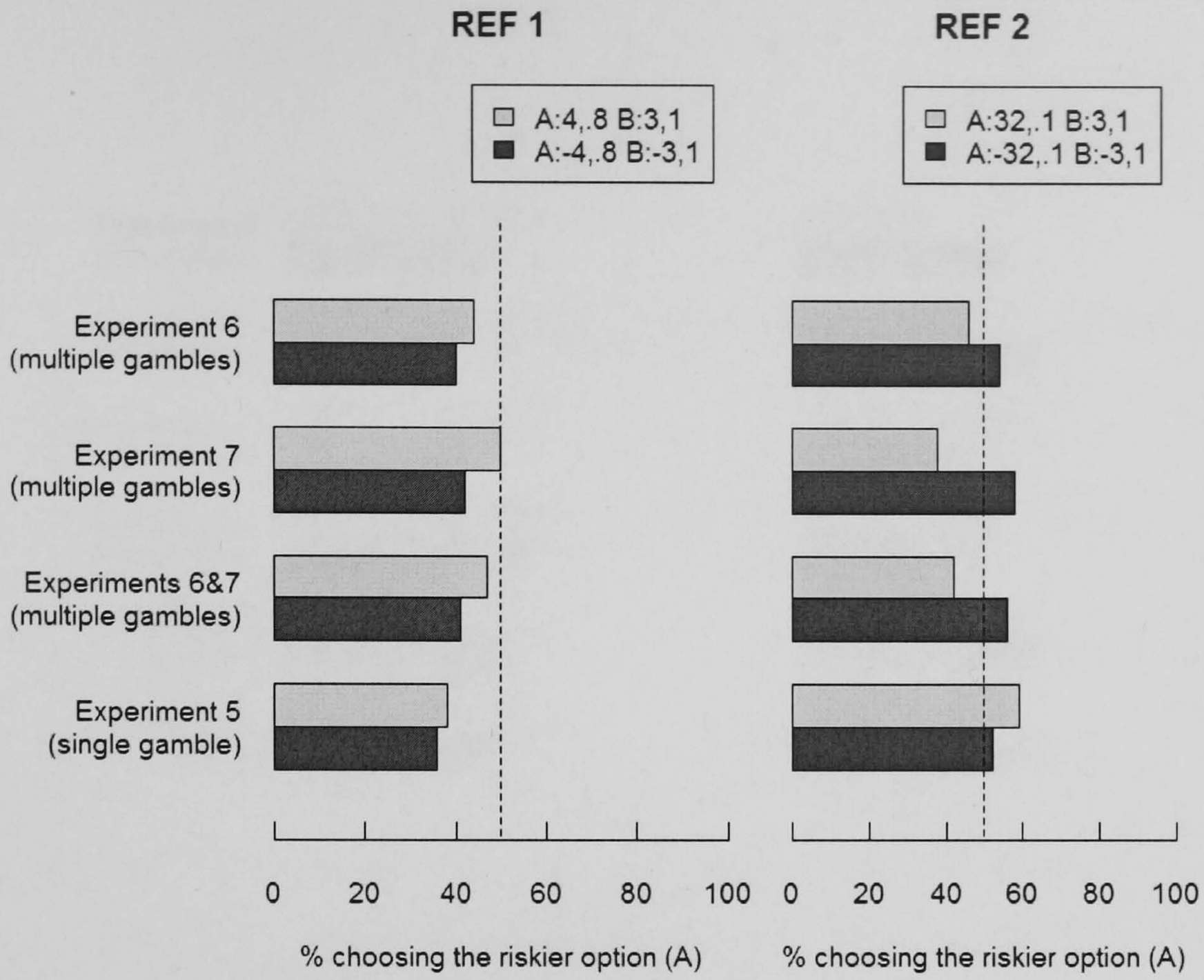


Figure 5.9. Proportions of risky choices for the REF problems under 40-40 Sampling

Additional confirmation for the reversed choice behaviour in REF1 and REF2 was obtained within the rest of the conditions. Significant reversed reflection effects under REF1 were found for 1-1 sampling order in (Experiment 7) and within the combined data from all the Matched-Sampling Conditions involving the presentation of multiple decision problems. Significant reversals in accordance with the reflection effects were observed for the combined Matched Sampling data in REF2, both for the experiments involving single and multiple decision problems (see Figure 5.10).

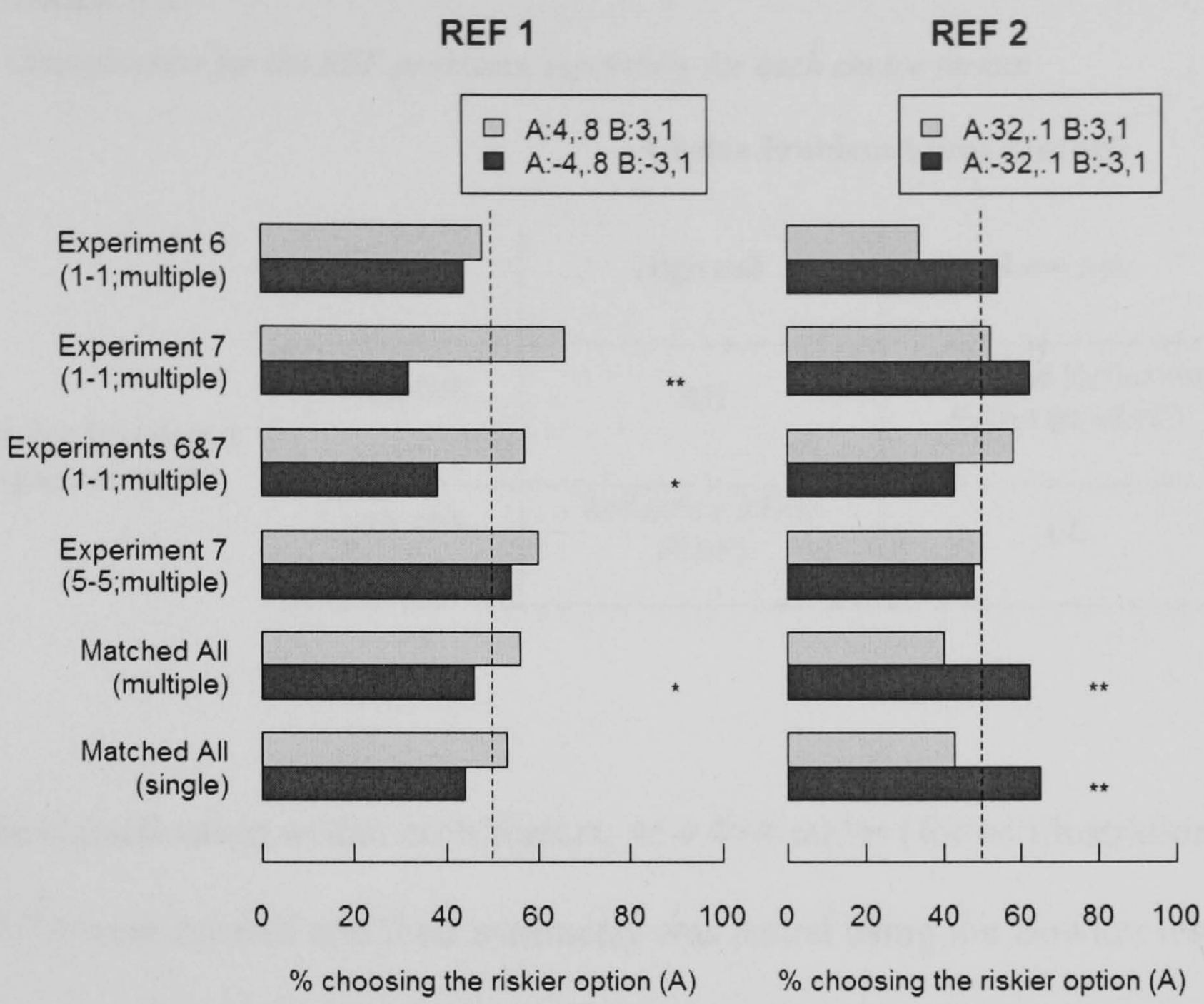


Figure 5.10. Proportions of risky choices for the REF problems under the remaining order conditions and across all Matched-Sampling Conditions (all sampling orders).

5.3.5 Within- participants analysis of the reflection effect

Finally, similar to the analysis conducted earlier on the common ratio problems, the within-participants reversals between descriptive and experiential choice for the REF problems were examined using the data from Experiments 6 and 7. Again, prior to the analysis, the two preferences within the choice problems under each format were combined to a single category, similar to the CR analysis above. The scheme for this classification is presented in Table 5.6.

TABLE 5.6

Classification for the REF problems, separately for each choice format

		Choice Problem 4 (loss context)	
		High risk	Low risk
Choice Problem 1 (gain context)	High risk	HH	Reversed Reflection Effect (revREF)
	Low risk	Reflection Effect (REF)	LL

With the classification within each format, new 4×4-tables (for an illustration see Table 5.7.) were created and their symmetry was tested using the Bowker test with a priori contrasts on the six individual 2×2 breakdowns of the table (the six pairs of cells with similar shades of grey).

TABLE 5.7

Classification for the REF agreement between the two choice formats. In the actual analysis only the corresponding non-diagonal cells with similar shades of grey are compared.

		REF classification within DfXP			
		REF	HH	LL	revREF
REF classification within DfD	REF	REF_REF	REF_HH	REF_LL	REF_revREF
	HH	HH_REF	HH_HH	HH_LL	HH_revREF
	LL	LL_REF	LL_HH	LL_LL	LL_revREF
	revREF	revREF_REF	revREF_HH	revREF_LL	revREF_revREF

For the data from Experiment 6 the choice behaviour within the reflection effect problems was found to be symmetrical for REF2, Bowker's χ^2 (6, N=150) =

5.166, $p = .523$. However, for REF1 a significant asymmetry was observed within the table, Bowker's χ^2 (6, N=150) = 21.903, $p = .001$. For Experiment 7 the null hypothesis of internal symmetry had to be rejected for both pairs of problems, REF1 (Bowker's χ^2 (6, N=250) = 41.375, $p < .001$) and REF 2 (Bowker's χ^2 (6, N=250) = 16.168, $p = .013$). To identify the sources for the asymmetries within the tables the planned comparisons were examined. It can be seen from Table 5.8, that in all three cases significant shifts were found between the reflection effect and the reversed reflection effect (Δ_{REF_revREF}). Specifically, more shifts were observed between a reflection effect under DfD to a reversed reflection effect under DfXP, than in the opposite direction. This confirms the differences found for the reversal of the reflection effect between subjects on a within-subject level. In addition, within Experiment 7 under REF1 an asymmetry was also observed for the comparison of choosing the low risk option in both cases and the reflection effect (Δ_{LL_REF}). Again, more participants shifted from a reflection effect pattern under DfD to the pattern of consistently choosing the low risk option under DfXP than the other way round. Together, both findings show that there seems to be a consistent pattern of changing the choice behaviour within reflection effect problems when faced with different choice formats. Instead of a reflection effect observed in DfD, participants shift to different patterns under DfXP which has been found to be either a reversed reflection effect or consistent preference of the low risk alternatives.

TABLE 5.8

Results for all six non-directional comparisons of the asymmetric 4 × 4 McNemar Tables. Significant χ^2 values are highlighted by asterisks.

		Z_{HH_LL}	Z_{HH_REF}	Z_{HH_revREF}	Z_{LL_REF}	Z_{LL_revREF}	Z_{REF_revREF}
Experiment 6	REF1	1.508	1.706	1.387	2.357	.577	2.985*
Experiment 7	REF1	1.0	1.915	2.2	3.053*	1.964	4.323*
	REF2	.0	.906	.853	1.298	2.324	2.746*

For the six non-directional comparisons of the off diagonal cells in the table and their corresponding mirror image counterparts (the cells in Table 5.8 with corresponding shades of grey) with a family wise alpha level of .05, equally divided, the critical value was 2.649 (from the tables by Dayton & Schafer, 1973).

5.4 Discussion

Overall, the findings presented here suggest that in order for PT to account for the choice behaviour under DfXP the shape of its value and weighting function must be different to the ones inferred from decisions under DfD. In terms of common ratio effect under Matched Sampling, there seems to be some support for a common ratio effect under CR1, similar to the results by Gottlieb et al. (Gottlieb et al., 2007), at least for the combined data including all different sampling orders. Within CR2, on the other hand, no effect was found. The proportions of riskier choices were similar in both the high and low probability context and remained generally at a lower level than in CR1, below the 50% line. Furthermore, the comparison between the different sampling conditions seems to provide a similar picture for the decision biases as seen for the raw choice proportions in the previous chapters. Whereas the usual common ratio effect was replicated under DfD, a reversed common ratio effect was observed under Free Sampling and

intermediate results were obtained under Matched Sampling. This shows once more that Matched Sampling attenuates the effect but it doesn't eliminate it.

The reason for difference between the two pairs of problems remains unclear but the result matches the trend within the Hertwig et al. data. The within-participants analysis shows that differences in terms of the behaviour in common ratio type problems under different choice formats can also be observed within participants, but the observed asymmetries are not direct shifts from a common ratio effect under DfD to a reversed common ration effect under DfXP, as suggested by the between-participants comparisons. These results imply that the weighting function must have a shape for which holds $w(.20)/w(.25) \geq w(.80)/w(1.0)$ and $w(.025)/w(.25) = w(.1)/w(1.0)$. The second part of this constriction is unfortunately much less informative as one of the slopes stretches from .1 to 1.0. How the function behaves in-between remains unclear. A generally flatter and more linear weighting function might be sufficient to explain the pattern. This would be in line with the findings presented earlier indicating that EV provides better predictions than PT. An alternative interpretation for the differences between CR1 and CR2 could be an interaction with the value function which possesses different properties for experienced values.

However, with only two pairs of common ratio type problems no further constraints regarding the functional form can be derived. A more systematic examination with specially selected problems accompanied by common consequence problems similar to the procedure used by Wu and Gonzalez (1996) would be useful to get a finer grained picture of different segments of the weighting function.

More consistency across the different choice conditions was observed for the reflection problems where a shift from a reflection effect under DfD to a reversed reflection effect under all experiential choice conditions was found for REF1 between and within participants. For REF2, on the other hand, a reflection effect was confirmed across all DfXP conditions. Both effects therefore replicated independently of the coexistence of sampling error. Again, there was an asymmetry between the two pairs of problems. The finding of a reversed reflection effect under REF1 on its own would seem to suggest a value function with a form opposite to the one found under DfD with $v(3) < w(.8)v(4)$ and $v(-3) > w(.8)v(-4)$, implying convexity for gains and concavity for losses. However, with the second and equally consistent finding of a reflection effect in REF2 with $v(3) > w(.1)v(32)$ and $v(-3) < w(.1)v(-32)$, implying the usual shape of the value function with concavity for gains and convexity for losses, the interpretation seems to be more complicated. An alternative explanation, reconciling the reversed risk preferences in the two problems, could be a dependence of the pattern on the properties of the probability weighting function and its impact on the different probabilities involved, as in the fourfold pattern of risk (Tversky & Kahneman, 1992), where the relationship between risk aversion and risk seeking for gains and losses is reversed for small probabilities. However, as similar probabilities occur in REF1 and REF2 the explanation seems to lie within the value function itself, which on its own would only be able to incorporate the results by assuming a more complicated functional form.

In summary, this chapter has provided a few new and interesting insights that will help to get a clearer idea regarding the shapes of the transformations of values and probabilities for a PT model in order to account for the experiential

choice data presented here. The patterns emerging are noticeably different to the ones usually observed under descriptive choice. In contrast to the existing literature which attributes the description-experience gap solely to a different weighting function, the findings observed here suggest that the differences might instead be the result of different properties of both the weighting and the value function. This is an entirely new insight that could give the search for the potential reasons behind the observed choice phenomena in DfXP an important new direction.

Furthermore, with regard to the interpretation of the apparent underweighting of small probabilities under DfXP, as the examined choice problems do not allow constraint of the functions any further there still remain a variety of different shapes for the weighting function that can accommodate the findings. As the involvement of over- or underweighting under DfXP is a model dependent statement, such conclusions cannot be drawn at this point. In order to further narrow down the set of potential shapes, I will use the next chapter to estimate the best fitting sets of parameters for the different experimental conditions that have been examined here. In addition, this analysis may also help to develop some hypotheses regarding the mechanism behind the changes of the functional forms when gamble descriptions are translated into experiential choice tasks.

CHAPTER 6

MODELLING DECISIONS FROM EXPERIENCE

6.1 Introduction

Throughout the experimental chapters of this thesis, I have presented various results indicating that PT based models are not very well suited to explaining experiential choice behaviour. When modelling DfXP results using cumulative prospect theory with the median value- and weighting-function parameters from Tversky and Kahneman (1992) the model performs at about chance level. In addition, I have provided evidence reinforcing the current view expressed in the literature that there is a difference in terms of the weighting of probabilities between DfXP and DfD. Furthermore, the reanalysis of the data in the context of decision biases in the last chapter has shown that the PT framework is more likely to account for the observed data when assuming different parameter values for both transformations. A number of constraints regarding the shapes of potential candidate functions have also been identified. However, in the light of these new results it is obvious that the initial model fits in the earlier chapters, which were based on a very limited range of parameter values, do not allow for an accurate assessment of the potential accountability of PT-based models for the reported data. In order to remedy this methodological shortcoming, and to test the convergence of evidence regarding the identified constraints, it is necessary to conduct a more comprehensive parameter estimation for the PT model. The evaluation of such a complete model fit will be the objective of the first part of this chapter.

Once it has been established what specific forms of the PT model give the best explanatory account of the experimental data this information can be used as a benchmark to evaluate the ability of a wider range of alternative models to explain the observed choice behaviour. This will be the objective of the second part of the chapter. However, given the scope of the thesis I do not claim this to be an exhaustive survey of all available models. Instead, I will try to explore the applicability of approaches that are the focus of the current literature, including simple heuristics and approaches derived from stochastic learning models.

A further contribution will be the exploration of a completely new theoretical approach that has not previously been applied in the context of DfXP. This model has some tradition in the probability learning literature and tries to use different aspects of the sequential information that are inherent to experiential choice tasks and that might not be captured by the alternatives considered so far. The chapter will close with a discussion of the obtained results.

6.2 Models based on prospect theory

Before I describe the modelling procedure and its results, I want to briefly go through the parameterisation of both the weighting and the value function of the PT model again, in order to illustrate the impact of these parameter values on the shape of the functions and their interpretation. This will also set the foundations for the predictions regarding the parameter values in the context of decisions from experience (DfXP) that follow from the results presented so far.

The value function $v(\cdot)$ with the parameterization of Tversky and Kahneman (1992), has the following form:

$$v(x_j) = \begin{cases} x_j^{\alpha} & ,\text{if } x_j \geq 0 \\ -\lambda(|x_j|)^{\beta} & ,\text{if } x_j < 0. \end{cases} \qquad (6.1)$$

There are three adjustable parameters (α, β, and λ) of which α and β separately fit the curvature of the function for gains and losses. The third parameter, λ, scales loss aversion and is only needed for mixed gambles which contain both gains and losses. As the problems used throughout the experiments presented here contain either all gains only or all losses only, this parameter cannot be estimated and is omitted. This leaves us with only two parameters (α, β) and the following function:

$$v(x_j) = \begin{cases} x_j^{\alpha} & ,\text{if } x_j \geq 0 \\ -(|x_j|)^{\beta} & ,\text{if } x_j < 0. \end{cases} \qquad (6.2)$$

The weighting function $w(\cdot)$, according to Tversky and Kahneman (1992), has two parameters (γ, δ) which fit the shape of the weighting functions, again separately for gains and losses:

$$w(p_j) = \begin{cases} \dfrac{p_j^{\gamma}}{\left(p_j^{\gamma} + (1-p_j)^{\gamma}\right)^{1/\gamma}}, \text{if } x_j \geq 0, \\ \dfrac{p_j^{\delta}}{\left(p_j^{\delta} + (1-p_j)^{\delta}\right)^{1/\delta}}, \text{if } x_j < 0. \end{cases} \qquad (6.3)$$

Under the two-stage model (Fox & Tversky, 1998; Tversky & Fox, 1995) the same function is applied to subjective probability estimates instead of objective probabilities.

Taken together, this leaves us with four parameters, which in the context of descriptive choice, are usually assumed to take on values between 0 and 1, resulting in both an S-shaped value function, concave for gains and convex for losses, and an inverse S-shaped weighing function, implying overweighting for small probabilities and underweighting of high probabilities. This is illustrated by the grey lined functions in Figure 6.1.

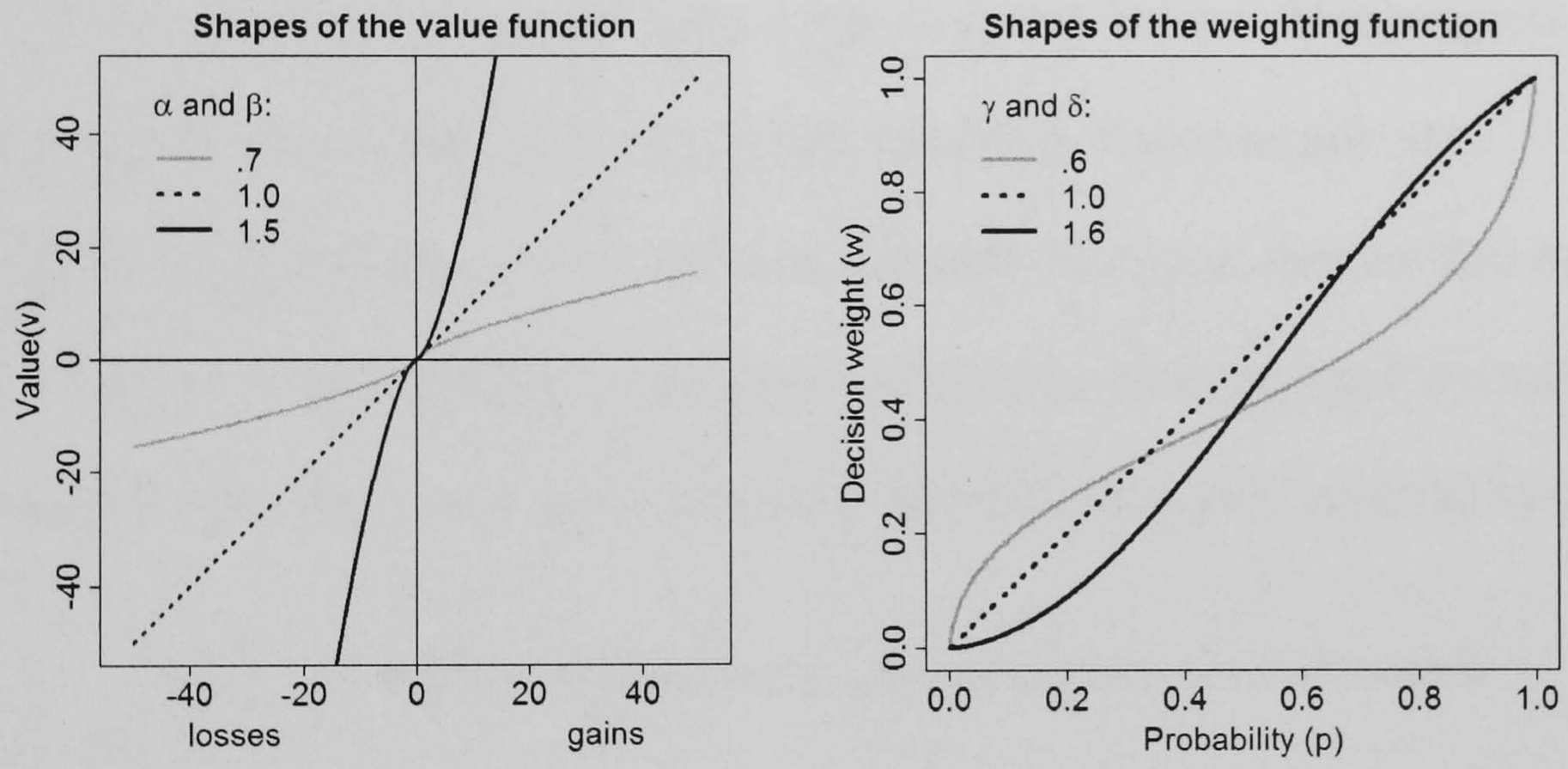


Figure 6.1. Different shapes of the value- and weighting function for parameter values between 0 and 2.

As I have shown in the previous chapter there is accumulating evidence that both functions have different properties under DfXP. The reversed reflection effect makes it necessary to assume a value function that is concave for losses and convex for gains (black lines in Figure 6.1A). A reversed common ratio effect, together with the choice behaviour in the earlier chapters, seems to imply a weighting function that is steeper between .2 and .25 than between 1.0 and .8, and

which accommodates underweighting of small probabilities (black lines in Figure 6.1B).

Interestingly, the only parameter estimation for the PT model reported in the DfXP literature so far (Hau et al., in press) has only allowed parameters to vary between 0 and 1. Given such a limitation of the parameter range the authors clearly bypassed an opportunity to explore the fit of alternative shapes that would actually describe their findings of choice behaviour in the direction of underweighting of small probabilities within the PT framework. Instead, they found an optimised performance of 69% for parameters indicating linear transformations for both probabilities and values. Effectively, the best fitting parameter values for gamma and delta in the range 0-1 are at 1. This suggests that better fitting values might be found outside this range. Consequently, the estimations reported here will circumvent this methodological shortcoming by providing the first estimation across an extended range between 0 and 2 which incorporates the functional forms that are in line with the results reported earlier.

6.2.1 Parameter estimations for prospect-theory-based models

Given the inherent dependencies of these two functions it is necessary to fit both transformations together. For example, risk averse responding is typically interpreted as a concave value function, but equally good fits can be obtained assuming a linear value function but a convex probability weighting function (e.g., Birnbaum's (2007) TAX model).

Using the decision problems involving only gains, I estimated α and γ and using the decision problems involving only losses, I estimated β and δ. The number of correct predictions was used as the optimisation criterion. For each

choice parameter values were optimised using an exhaustive grid search. Parameter values were selected to maximize the number of model predictions in the same direction as the modal preference. Given the small number of decision problems per person (1 or 6), there was not enough information to estimate the parameters separately for each subject. Rather, the choices from different subjects were pooled together and parameters were estimated across all choices at the population level.

Note that as the dichotomous outcome predicted by the model (A or B) stays constant across a wide range of parameter values for both the value- and the weighting function and does not change continuously. The same is true for the overall rate of correct predictions. As a result, there is not one unique set of parameters that produces the maximum number of correct predictions. Instead, there are discrete steps with the same rate of correct predictions across a range of different pairs of parameter values. In a three dimensional space the resulting surface will therefore have various plateaus with the same fit (e.g., like a hill with terraces of rice fields). Optimisation algorithms like the Nelder-Mead method (Nelder & Mead, 1965) or, more broadly, methods that depend on local gradient information, can therefore not be employed here as the simplex would get stuck on one of these plateaus, failing to find the true best fit. Given the small number of parameters, I therefore avoided this problem by calculating the number of the correct predictions for each combination of parameter values of the value- and weighting function between 0 and 2 in steps of 0.01. This will provide an estimate for the upper limit of the predictive power that this model can achieve on the basis of experiential choice data. To avoid overfitting, the parameter estimations are conducted for both combined sets of data, wherever possible, and separately for

each individual set of data. The same methodology was used for the calculation of the model fits of the two-stage model. Due to the reliance on subjective probability estimates, the predictions of the two stage model could only be tested on data from the experiments involving either frequency or probability estimations (Experiments 3, 4 and 5).

6.2.2 General performance of the prospect theory model

Although the performance of the model was examined across a wider range of parameter values than usual, the mean maximum fit obtained across the different experiential data sets remained still low with 65% (*SD* = 9%). This is also lower than the performance for PT usually reported under DFD.

The maximum proportions of correct predictions obtained across the different sets of data are provided in Figure 6.2. The best performances were observed for the Free Sampling data of Experiment 1 (83%) and Experiment 2 (81%). 71% correct predictions were obtained for the descriptive choice data in Experiment 7 and the Comprehensive-Sampling data in Experiment 1. However, for the Matched Sampling data the performance was found to be much lower (*M* = 60%, *SD* = 3%). The lowest fits (58 % - 55%) were observed within the data from conditions with restricted sampling orders in Experiments 5 and 7.

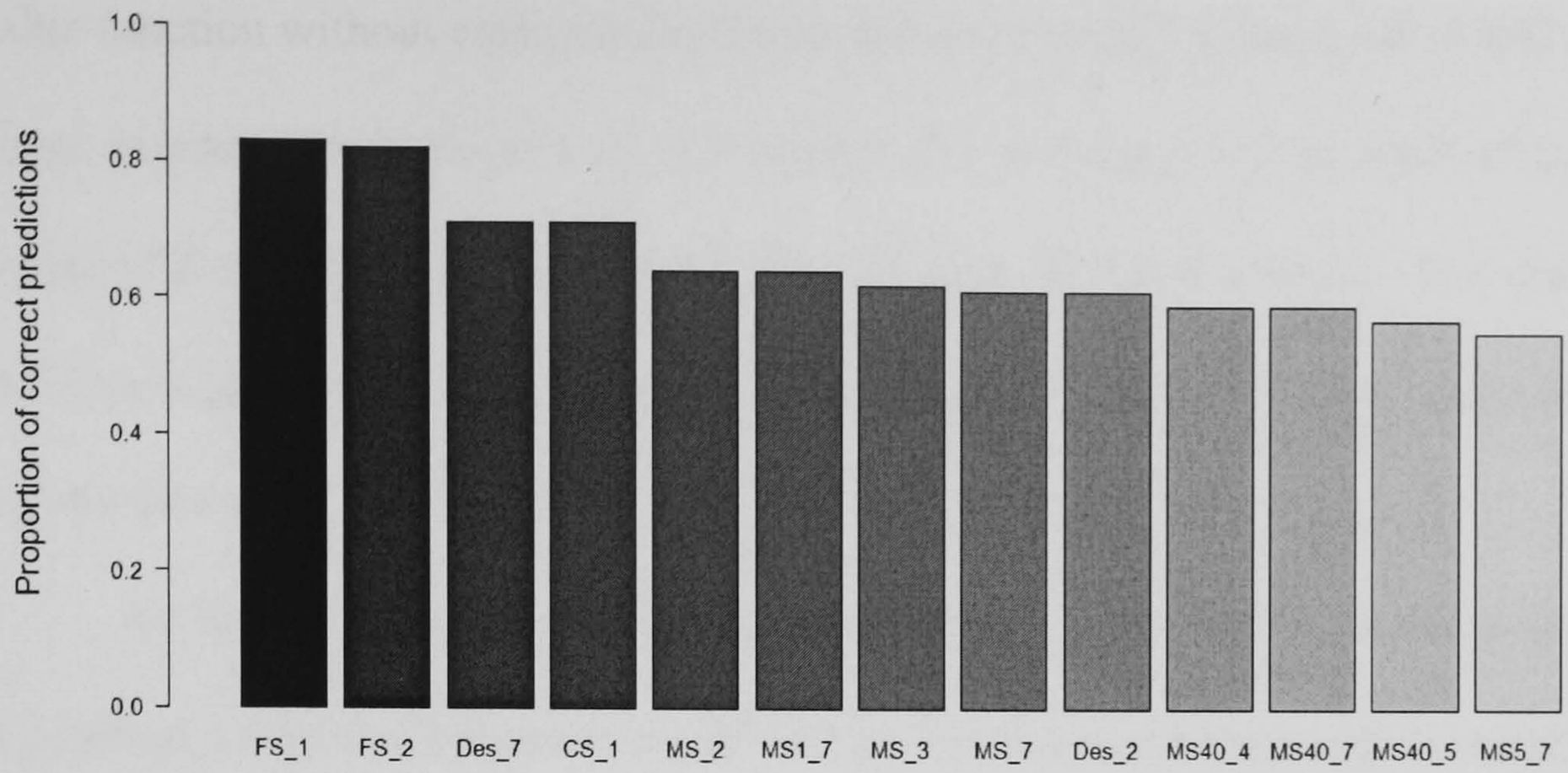


Figure 6.2. Summary of the maximum rates of correct predictions for the PT model. The conditions sorted by performance are: FS = Free Sampling, Des = Description, CS = Comprehensive Sampling, MS = Matched Sampling (1,5 and 40 are the restricted sampling orders). The numbers at the end indicate the numbers of the experiment.

As important as the fits are the sets of parameter values under which these values were obtained. This information will also tell us to what extend the best fitting shapes of the functions overlap with the constraints inferred from the decision biases in the previous chapter. The following sections will therefore explore the parameters of these maximum fits in more detail, separately the different experimental conditions.

6.2.3 Prospect theory in the free sampling paradigm

As shown in the previous section, the best fits for PT could be observed within the Free-Sampling Conditions. But where do these best fitting parameter values actually lie? In Figure 6.3 I have plotted the rates of correct predictions as a function of the two parameters across the entire range (0-2), separately for gains on the left and losses on the right. Areas of the same shade indicate that equally good fits can be obtained by trading off value-function and probability-weighting-function parameters. Thus, one cannot talk of the best fitting parameter for the

value function without constraining the probability weighting parameters and vice versa. In other words, there is no best fitting set of parameters, but rather a large region of parameter space over which equally good fits are obtained – this type of problem is known as the problem of flat maxima, in the literatures on optimization and statistics.

To facilitate the interpretation, I have added a vertical and a horizontal line at position 1.0 for both parameters, dividing each range of values into two halves, thus forming four quadrants. Firstly, this division helps to detect inherent characteristics of the best fitting parameters and their interaction. Secondly, the lines help to identify the regions that are tangent to a linear value- and weighting function.

Within all eight plots the darkest regions, representing the best fits, lie predominantly within the top left and top right quadrants which are marked by weighting function parameters greater than 1. Conversely, the brightest areas with the lowest fit are mostly found for weighting function parameters smaller than 1, especially within the estimates for losses on the right side, for which the rate of correct predictions seems to be generally lower. This confirms the earlier results of a potential underweighting of small probabilities within the observed choice data and would also roughly fit the restrictions implied by the reversed common ratio effect that was found for this data.

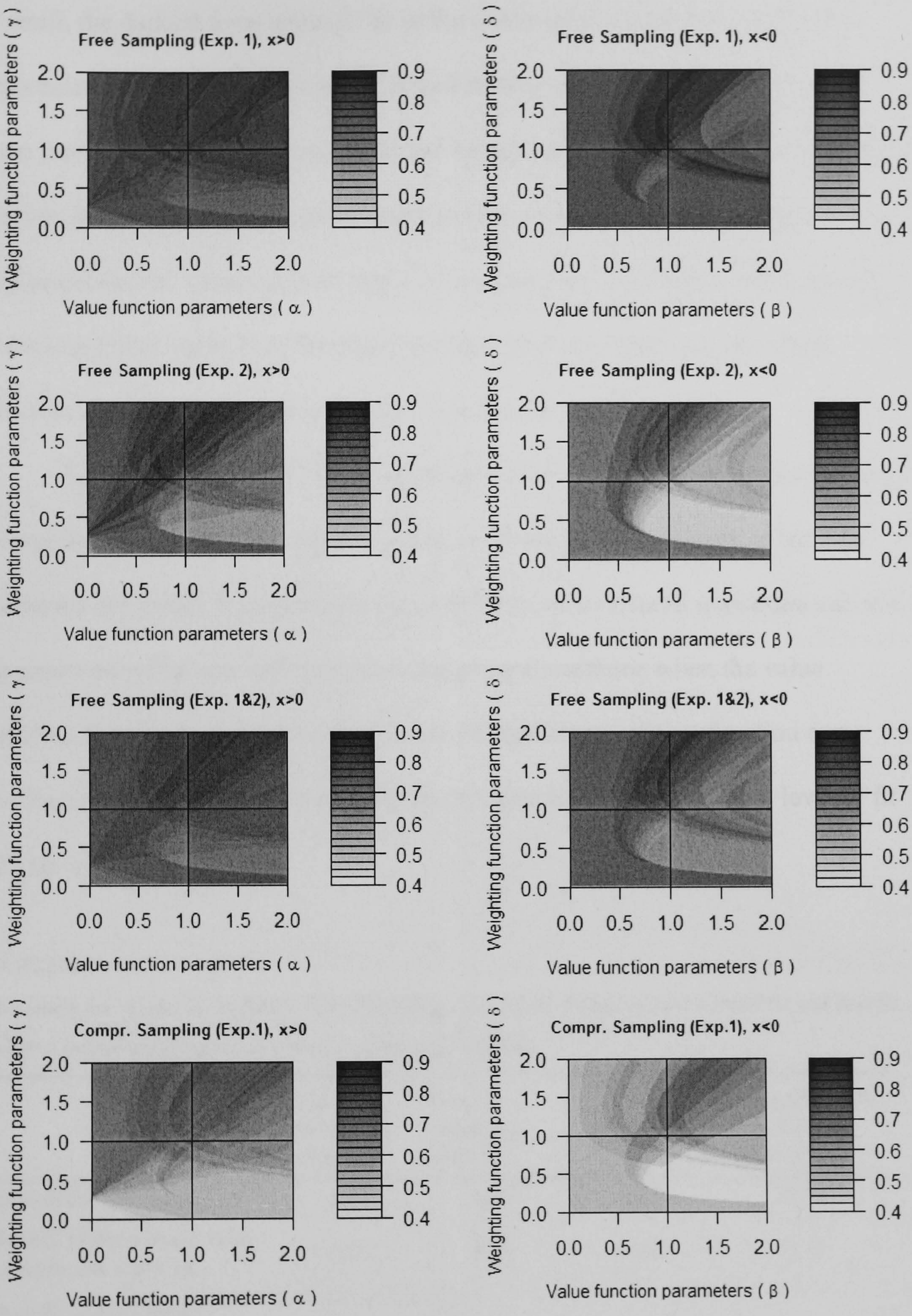


Figure 6.3. Filled contour plots with the rates of correct predictions for the tested combinations of parameter values in the context of the Free Sampling data. Left column: Fits to the gains-only gambles. Right column: Fits to the losses-only gambles.

In terms of the value function parameters the interpretation is less clear. Overall, the darkest areas seem to lie in the centre of the upper half of the plots with extensions into the right half. This clustering along the centre seems to be even more pronounced for losses and can be seen as an indication that the best fits are found for value function parameters around 1 and slightly greater than 1. As explained earlier, values greater than 1 are in line with the finding of a reversed reflection effect under Free Sampling because they result in a convex value function for gains and a concave value function for losses.

Table 6.1 shows a different summary of the results with the highest overall fit and the best fit obtained under conditions in which either or both of the two functions are linear. It can be seen that a fit close to the overall maximum can still be observed within the different Free-Sampling Conditions when the value function is set to be linear. Under a linear probability weighting function or a combination of both linear functions, on the other hand, the maximum level of fit can no longer be obtained.

TABLE 6.1

Maximum fits within the different Free-Sampling Conditions including best overall fit and best fit under a linear weighting and/or linear weighting function

	Free Sampling (Exp 1)	Free Sampling (Exp 2)	Free Sampling (Exp 1&2)	Comprehensive Sampling (Exp 1)
Best fit (all gambles)	0.83	0.81	0.82	0.71
Best fit under a linear value function ($\alpha = \beta = 1$)	0.83	0.81	0.82	0.70
Best fit under a linear weighting function ($\gamma = \delta = 1$)	0.81	0.77	0.76	0.67
Best fit under a linear value and weighting function ($\alpha = \beta = \gamma = \delta = 1$)	0.81	0.68	0.75	0.65

Taken together, the optimum performance for the Free Sampling data can be expected from PT models with a weighting function that incorporates underweighting of small probabilities and a value function that is either close to linear or slightly S-shaped.

6.2.4 Prospect theory under Matched Sampling

Equivalent error surfaces for the Matched-Sampling Conditions are shown in Figure 6.4. A few general remarks have to be made regarding this set of plots. First, they are all much brighter than the previous ones, demonstrating that the rates of correct predictions are generally much lower than within the Free-Sampling Conditions. Second, there are smaller numbers of areas with much bigger and clearer shapes. This is a result of the matching process which reduces the number of different probabilities experienced during the sampling process. Furthermore, the plots show a remarkable consistency regarding the form of the resulting shapes, providing a more consistent picture between experiments than the observed preferences. It can be seen from Figure 6.4 that the parameter combinations yielding the highest model performance for gains and losses lie predominantly in the top-right quadrant with both parameters greater than 1. Within a smaller number of plots these areas also reach into the bottom-left quadrant (e.g. Experiment 2 for gains and most of the loss plots). In terms of the weighting function parameters this means again that the best fits for this data are generally found for parameters incorporating an S-shaped weighting function with underweighting of small probabilities.

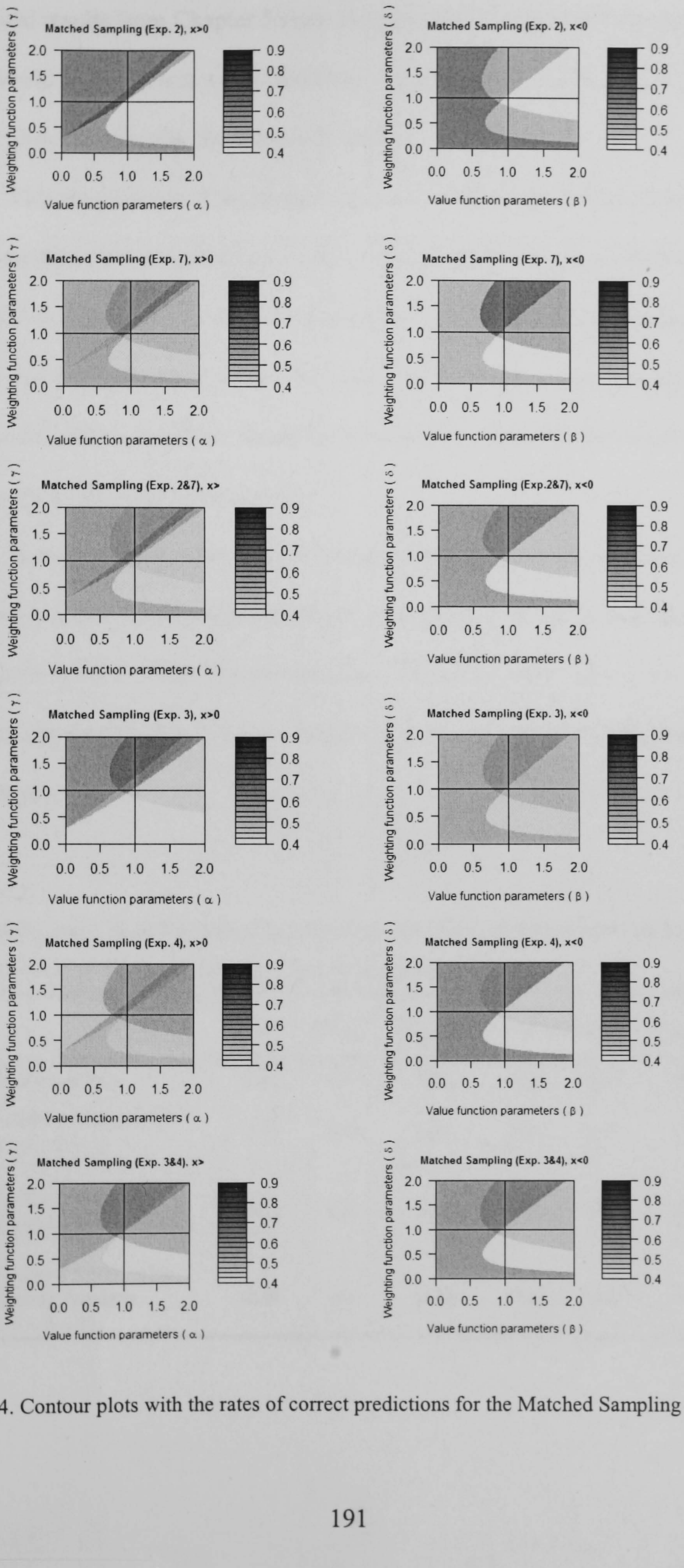


Figure 6.4. Contour plots with the rates of correct predictions for the Matched Sampling data.

The mixed results from Chapter 5 showing either no or a reversed common ratio effect would be consistent with probability weighting parameters close to one or slightly above, restricting the shape still further.

With the plateaus of the highest fits covering both the left and the right side of the plots, the interpretation in terms of the value function is even more difficult than in previous cases. Within most of the plots the areas with the highest fit seem to cover parameter values between .7 and 1.8. However, in order to match the reversed reflection effects found for this data the values can be narrowed down to the right of the vertical line.

An examination of the fits for the different combinations of linear functions in Table 6.2 provides additional confirmation for this interpretation. Under both, a linear weighting or value function performance close to the optimum can be obtained. Their combination, however, results in suboptimal performances.

TABLE 6.2

Maximum fits within the different Matched-Sampling Conditions including best overall fit and best fit under a linear weighting and/or linear weighting function

	Exp 2	Exp 7	Exp 2&7	Exp 3	Exp 4	Exp 3&4
Best fit (all gambles)	0.64	0.61	0.61	0.62	0.59	0.59
Best fit under a linear value function ($\alpha = \beta = 1$)	0.63	0.61	0.61	0.62	0.59	0.59
Best fit under a linear weighting function ($\gamma = \delta = 1$)	0.63	0.61	0.61	0.58	0.59	0.59
Best fit under a linear value and weighting function ($\alpha = \beta = \gamma = \delta = 1$)	0.39	0.5	0.46	0.46	0.49	0.47

6.2.5 Prospect theory under restricted sampling order

The maximum PT fits for these sets of data were found to be the lowest and this is also reflected in the low intensity of the plots. However, the shapes of the different plateaus look remarkably similar to the other Matched Sampling plots (see Figure 6.5). The first reading is therefore the same with optimal probability weighting function parameters greater than one and value function parameters close to one and greater than one. The constraints, on the other hand, have been found to be slightly different which needs to be taken into account when interpreting these values. For most of these conditions there was indication for a common ratio effect. As a result, the set of potential parameters within the domain of gains has to be found on the extension of the diagonal with the highest fit somewhere within bottom-left quadrant. For losses, however, the area with the highest fit does not reach into the lower half of the plot. Given these constraints we would therefore have to assume the actual performance to be lower than the obtained maximum. The only exception seems to be the 5_5 Condition. In terms of the value function, the spread is again very wide, especially for the 40_40 data for which no reflection effect of any form could be found. For the other conditions, the highest fitting value parameters have values greater than one which matches reversed reflection effects as they were observed within this data.

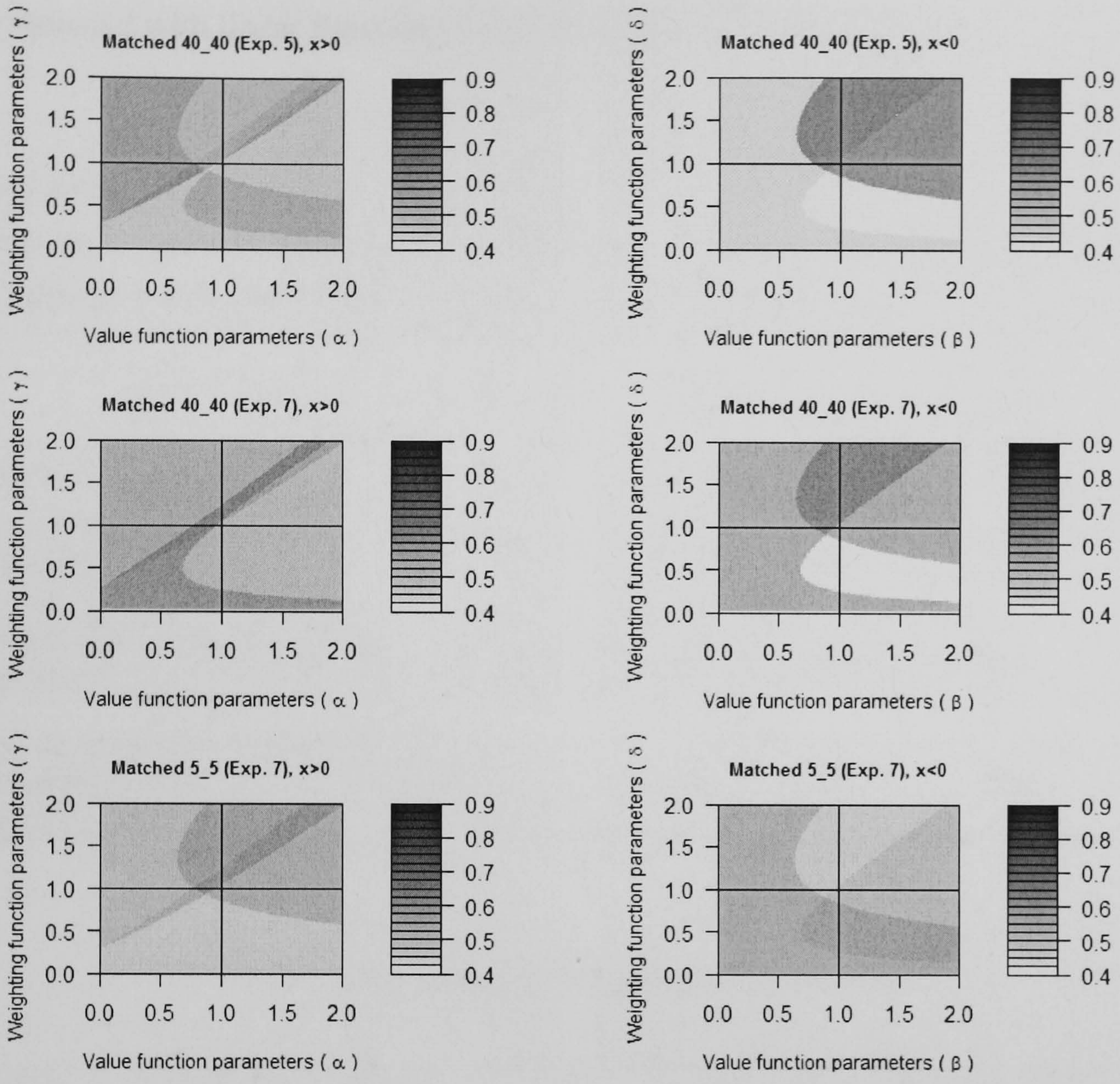


Figure 6.5. Filled contour plots with the rates of correct predictions for the Matched Sampling data with restricted sampling order.

Nevertheless, as the fits are generally rather low and vary only slightly, the figures in Table 6.3 show that for most of the conditions a similar level of fit can be obtained with linear functions and even for their combination.

TABLE 6.3

Maximum fits within the different Matched-Sampling Conditions with restricted sampling order for different combinations of linear value and weighting functions

	40_40 (Exp 5)	40_40 (Exp 7)	1_1 (Exp 7)	5_5 (Exp 7)
Best fit (all gambles)	0.57	0.58	0.64	0.55
Best fit under a linear value function ($\alpha = \beta = 1$)	0.56	0.58	0.64	0.55
Best fit under a linear weighting function ($\gamma = \delta = 1$)	0.55	0.58	0.63	0.55
Best fit under a linear value and weighting function ($\alpha = \beta = \gamma = \delta = 1$)	0.5	0.5	0.57	0.51

6.2.6 Performance under descriptive choice

In order to provide an additional validation of the observed difference between DfXP and DfD, I also tested the performance of the model in the context of the data from the Descriptive-Choice Conditions. As already pointed out above, the obtained fits for this data were not as high as usually reported ($M = 65\%$) but still higher than the Matched Sampling fits. More important though, is the distribution of the actual parameter values providing the best fit for this data. It can be seen from Figure 6.6 that the distribution of the grey scale values is actually inversed for the descriptive choice data, compared to the results under experiential choice. In the context of losses the resulting areas of optimum fit is rather large which allows for a wide range of equally likely combinations.

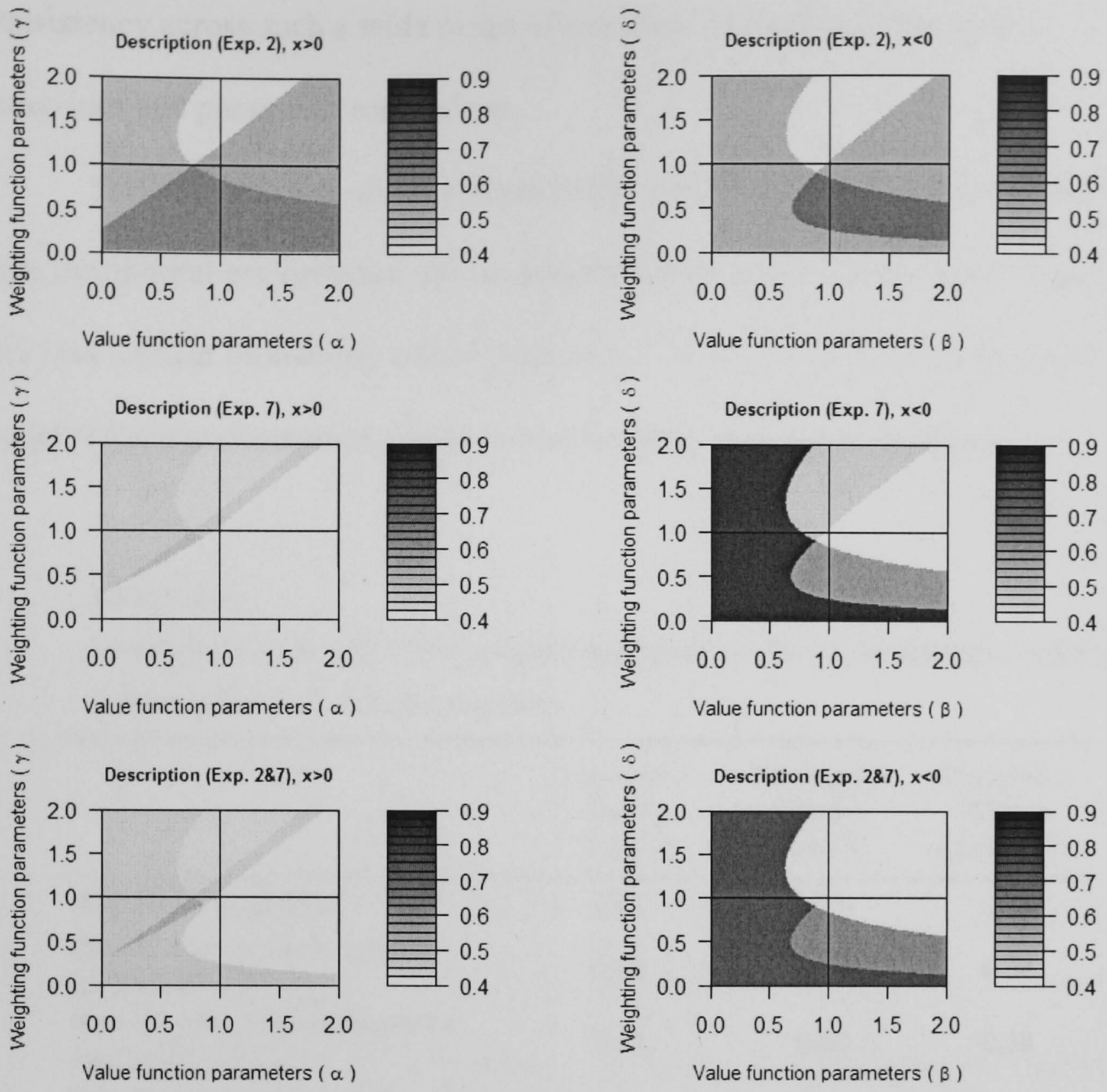


Figure 6.6. Filled contour plots with the rates of correct predictions under descriptive choice.

However, most of the top-right quadrant, which provided the best fitting parameter combinations for experiential choice, is now white, indicating the areas of the lowest fit under descriptive choice. For the gains, on the other hand, there is clear indication that the optimal values can be found within the bottom-left quadrant which overlaps with the parameter values usually reported in descriptive choice. All these findings illustrate again the obvious shift in terms of the parameter values that is necessary for the PT framework to account for choice behaviour within both formats. It is important to mention that this is actually the first time this difference between the formats has been found with such clarity and

consistency across such a wide range of evidence, including actual choice behaviour and parameter estimations.

Furthermore, it is obvious from Table 6.4 that for the descriptive choice data the optimal performance can be achieved even under a linear value function, whereas a linear probability transformation on its own or in combination with linear value transformation results in a substantial drop in performance.

TABLE 6.4

Maximum fits within the different descriptive choice conditions for different combinations of linear value and weighting functions.

	Descriptive Choice (Exp 2)	Descriptive Choice (Exp 2)	Descriptive Choice (Exp 2&7)
Best fit (all gambles)	0.61	0.71	0.64
Best fit under a linear value function ($\alpha = \beta = 1$)	0.61	0.71	0.64
Best fit under a linear weighting function ($\gamma = \delta = 1$)	0.51	0.60	0.58
Best fit under a linear value and weighting function ($\alpha = \beta = \gamma = \delta = 1$)	0.51	0.29	0.35

6.2.7 The performance of the two-stage model

In order to examine whether the performance of the PT-framework can be improved when using subjective probability estimates instead of the objective probabilities I also estimated the parameters for the two-stage model. However, the best fit observed on the basis of the available probability estimates did not provide any improvement. Instead, with a mean of 65% an upper limit close to the 63% reported by Fox and Hadar (2006) was confirmed.

As the number of different individual probability and frequency judgements exceeds the number of objective probabilities, the gradation of colouring for the two-stage plots in Figure 6.7 is much more complex, again with more variance and smaller areas. Similar to the earlier parameter estimations, the best fitting weighting function parameter values can be found above the horizontal line with values greater than one. Thus, even when taking sampling error out of the equation using probability estimates the best performance of the PT framework is still found under a probability transformation that allows for underweighting of small probabilities. This provides additional support for the claim made on the basis of the observed choice behaviour that the elimination of sampling error does not eliminate the apparent underweighting of small probabilities.

Again, it is more difficult to interpret the plots in terms of the value transformations. In the case of the two-stage model plot this is made even more difficult due to the inconsistency between plots, especially within the plots for losses on the left of Figure 6.7.

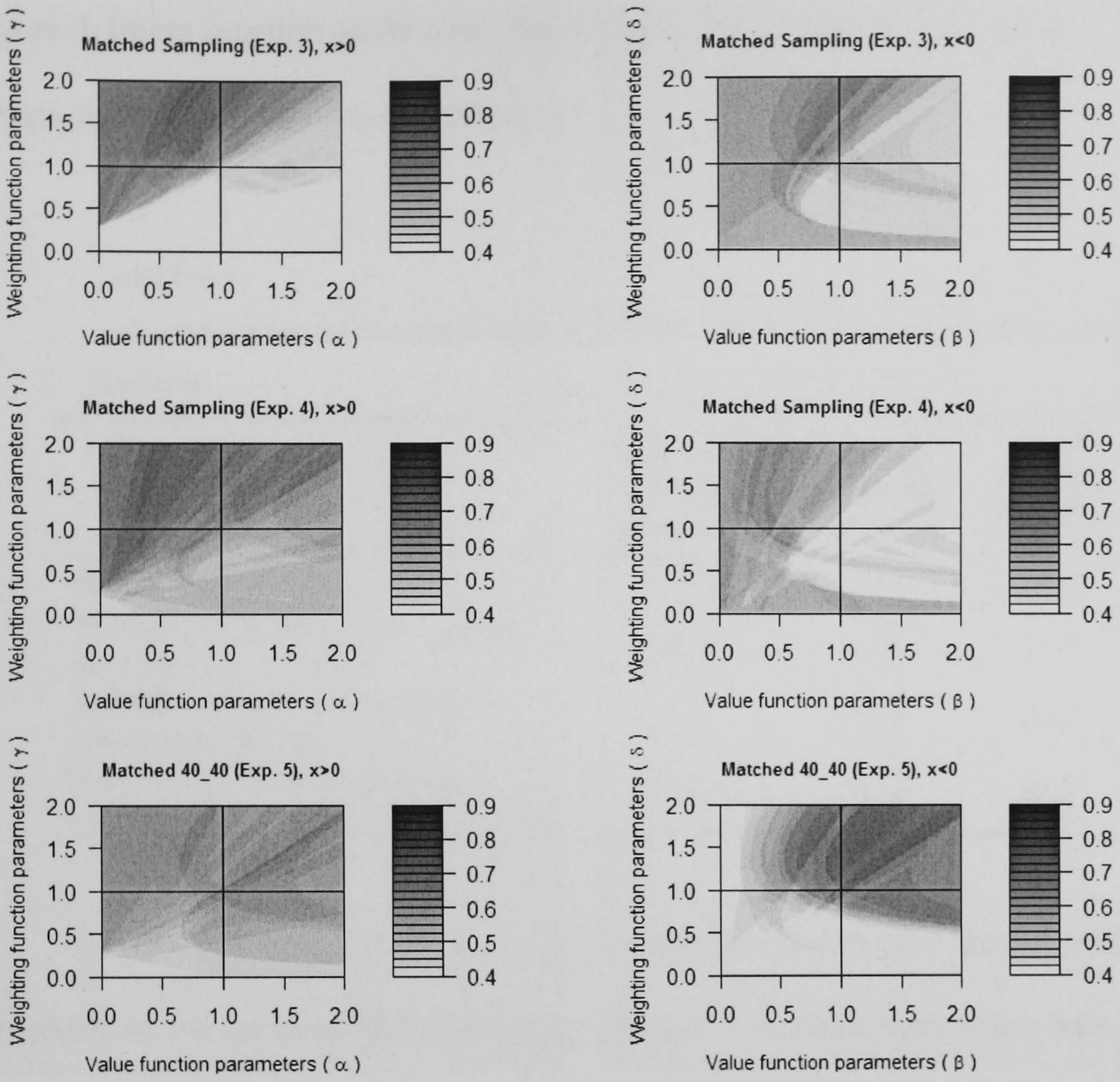


Figure 6.7. Filled contour plots with the rates of correct predictions for the two-stage model under Matched Sampling.

The performance under linear transformations seems to be close to the maximum for each linear function on its own. When combined though, the performance drops significantly again (see Table 6.5).

TABLE 6.5

Maximum fits for the two-stage model with combinations of linear value and weighting functions

	Matched Sampling (Exp 3)	Matched Sampling (Exp 4)	40_40 (Exp 5)
Best fit (all gambles)	0.65	0.64	0.64
Best fit under a linear value function ($\alpha = \beta = 1$)	0.63	0.60	0.63
Best fit under a linear weighting function ($\gamma = \delta = 1$)	0.61	0.63	0.63
Best fit under a linear value and weighting function ($\alpha = \beta = \gamma = \delta = 1$)	0.48	0.53	0.61

In summary, this first section of the chapter has shown how the different results from the previous analyses can be integrated in the context of the best fitting parameter values for PT models under DfXP. Furthermore, the optimisations have helped to narrow down the shape of functions that a PT model would need to accommodate in order to describe choices under DfXP. The emerging picture confirms the initial hypothesis that the shapes must be different to the ones that have been established under DfD. For the first time this could also be shown in a direct comparison of the best fitting parameter values obtained for these two formats. Finally, the results of the two-stage model suggest that these findings can be obtained independent of the elimination of sampling error.

6.3 The application of an adaptive learning model

The upper limits for the predictive power of PT-based models reported in the first part have left plenty of scope for improvement. It might therefore be indicated to compare these first results with the performance of alternative models that capture different aspects of the available information. What all PT-based models have in common is the usage of aggregated information from the experienced sequences including the overall probabilities of the different outcomes. They therefore do not take into account specific properties of the sequential accumulation which is also a characteristic of the task.

One alternative class of models that incorporates the sequential updating of the experienced information and which has already been discussed in the context of decisions from experience are associative learning models. As mentioned in the introduction, there has been a tradition of using different variants of simple reinforcement models formerly developed in the animal and human learning literature to examine repeated-choice problems and learning phenomena in bandit problems (e.g., Barron & Erev, 2003; Denrell, 2007; Erev & Barron, 2005; March, 1996; Sarin & Vahid, 2001). A systematic comparison of the assumptions underlying these different models has been presented by Yechiam and Busemeyer (2005). All of these models can be seen as derivatives of the standard Bush-Mosteller type stochastic learning models (Bush & Mosteller, 1955; Estes, 1959). The basic assumption is that the probability of choosing on option increases or decreases depending on how choosing the option is rewarded. The decision maker's assessment of the expected value of an uncertain alternative at time *t* is a weighted average of the previous estimate and the most recent payoff which is usually weighted by a learning parameter that determines the rate of adaptation.

6.3.1 The value-updating model

Hertwig et al (2006) have tried to capture the sequential updating process within DfXP by applying a particular variant of a weighted adjustment model (March, 1996), which they refer to as the *value-updating model* and which seems to follow an earlier form proposed by Barron and Erev (2003). Again, the basic assumption is that people update their expectation regarding the value of the option j at time t, $A_j(t)$, with every new piece of information according to the following mechanism:

$$A_j(t) = (1 - \omega_t)\, A_j(t\text{-}1) + (\omega_t)\, v(x_t), \qquad (6.4)$$

where the value at time t is modelled as a weighted average of the expectation according to the previously encountered outcomes from this option $A_j(t-1)$, and the latest value drawn, x_t. Instead of a learning parameter, this model employs a recency parameter, φ, to weight the value of the latest outcome: $\omega = (1/t)^{\varphi}$. Values for $\varphi = 1$ indicate equal weighting whereas $\varphi < 1$ imply recency weighting and $\varphi > 1$ primacy weighting. With no prior knowledge about the available options the initial expectations $A_j(0)$ are set to 0. In addition, the last part of Equation 6.4 shows that Hertwig et al. also incorporated a prospect theory type value function $v(\cdot)$ to transform the experienced outcomes. Following Barron and Erev (2003), this function has the following form:

$$v(x_i) = \begin{cases} x_i^{\alpha} & , \text{if } x_i \geq 0, \\ \lambda |x_i|^{\alpha} & , \text{if } x_i < 0. \end{cases} \qquad (6.5)$$

Compared with the function used in the previous section, this parameterisation is unusual as it contains the loss aversion parameter λ but only one parameter to determine the curvature of the transformation of both gains and losses. As

Hertwig et al. did not employ any mixed gambles λ can be omitted and leaves only α as a free parameter. A slightly different variant was implemented by Hau et al. (in press) who dropped the value function altogether, assuming a linear value transformation.

Rather than conducting a complete optimisation, Hertwig et al. (2006) implemented the Tversky and Kahneman (1992) parameters for the value function and estimated only the recency parameter φ. With a value of .29 for this parameter they could obtain a correlation of .91 between predicted and actual choice proportions. Hau et al. (in press) found the maximum rate of correct predictions for their estimations to be 66% under a parameter value indicating much less recency weighting (φ =.75). However, both studies looked only at DfXP under Free-Sampling Conditions. Given the consistent absence of any form of recency weighting in the analyses reported throughout this thesis, I expected this parameter to be much closer to 1 in the context of the Matched Sampling data.

With the limitations of these original procedures, I have adapted the model slightly for my own test of the performance of the value-updating model and used a more comprehensive estimation procedure. First, instead of the value function parameterisation of Hertwig et al. (2006), I used the same function that I have employed for the PT model fits (see Equation 6.2). Second, given the earlier results underlining the unsuitability of the Tversky and Kahneman (1992) value function parameters in the context of experiential choice data I estimated α from the data (as well as φ). As in the first part of this chapter, the estimations were therefore conducted separately for gains and losses, with the rate of correct predictions as the optimisation criterion. Finally, for each set of parameters the estimations were conducted in steps of 0.01 within the limits of 0 and 2.

6.3.2 The value-updating model under Matched Sampling

Although the model uses different properties of the information experienced by the subjects, the performance observed under optimal value and recency parameters was comparable to the maximum fits reported for the PT model with a mean of 66% (*SD* = 3%). These rates of correct predictions can also be seen as a confirmation of the results by Hau et al. (in press) within Matched Sampling. A summary of the best fits obtained across the different data sets is provided in Table 6.6.

TABLE 6.6

Maximum fits within the different Matched-Sampling Conditions.

	Matched Sampling (Exp 2)	Matched Sampling (Exp 7)	Matched Sampling (Exp 3)	Matched Sampling (Exp 4)
Best fit (all gambles)	0.71	0.65	0.66	0.63
Best fit under a linear value function ($\alpha = \beta = 1$)	0.61	0.63	0.63	0.61
Best fit under a equal recency weighting ($\varphi = 1$)	0.67	0.61	0.58	0.59
Best fit under a linear value function and equal recency weighting ($\alpha = \beta = \varphi = 1$)	0.35	0.5	0.46	0.49

However, in terms of the best fitting parameter values the interpretation is much less clear than in the context of the PT models. Due to the functional form of the model there is more variance and we do not have the same big-sized plateaus as in the PT estimations. Nevertheless, there are still sets of parameter values that provide optimal fit. I therefore used contour plots again to illustrate the results (see Figure 6.8).

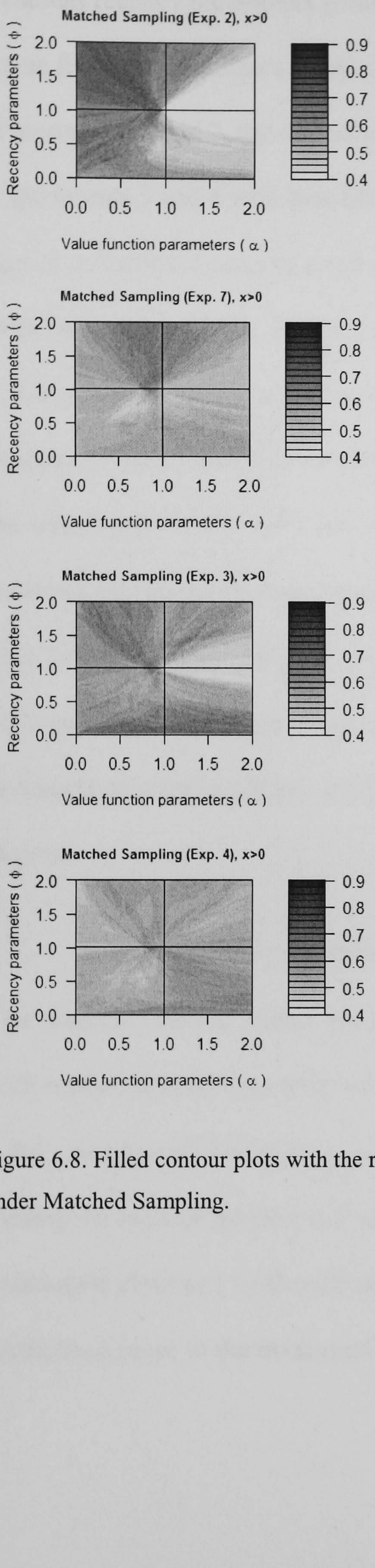


Figure 6.8. Filled contour plots with the rates of correct predictions for the value-updating model under Matched Sampling.

Whereas recency parameters greater than 1, indicating recency weighting, and value function parameters smaller than 1 provide the best fit for the data from Experiments 2 and 7, the opposite pattern was found for the data from Experiments 3 and 4 with best fitting parameter values in quadrant 4. However, as shown in Table 6.6 rates of correct predictions close to the optimum were still observed within all four data sets under the assumption of a recency parameter of 1, which would also be more consistent with the actual findings of the recency analysis presented in the experimental chapters. Setting both parameters to 1, on the other hand, would result in a significant drop in predictive power which indicates the importance of a nonlinear value function in order to describe this data. More consistency than in the PT optimisations was found with regard to the differences between the separate estimations for gains (left side) and the estimations for losses (right side) within each data set, which were much more aligned.

6.3.3 The value-updating model under restricted sampling order

The observed rates of correct predictions for the Matched-Sampling Conditions with restricted sampling order were again slightly lower with a mean of 58% (*SD* = 4%). In terms of the optimum parameters, on the other hand, there was more overlap. In most of the data sets, the best fit was obtained under recency parameters close to 1 or slightly above. This is also reflected in rates of correct predictions close to the maximum for linear recency in Table 6.7.

TABLE 6.7

Maximum fits within the Matched-Sampling Conditions with restricted sampling order.

	40_40 (Exp 5)	40_40 (Exp 7)	1_1 (Exp 7)	5_5 (Exp 7)
Best fit (all gambles)	0.56	0.58	0.64	0.55
Best fit under a linear value function ($\alpha = \beta = 1$)	0.56	0.58	0.64	0.55
Best fit under a equal recency weighting ($\varphi = 1$)	0.55	0.58	0.63	0.55
Best fit under a linear value function and equal recency weighting ($\alpha = \beta = \varphi = 1$)	0.5	0.5	0.57	0.51

This is also more in line with the experimental findings confirming the absence of any form of recency weighting. The best fitting value function parameters, on the other hand, were generally found to be slightly below 1 (see also Figure 6.9). This applies to both gains and losses, and would imply an inverse S-shaped value function as it is usually assumed under prospect theory. Unlike in the Matched-Sampling Conditions with unrestricted sampling order the nonlinearity of the value function is not that important. As the summary in Table 6.7 shows, rates of correct predictions close to the maximum can still be observed under a linear value transformation without recency weighting ($\varphi = 1$).

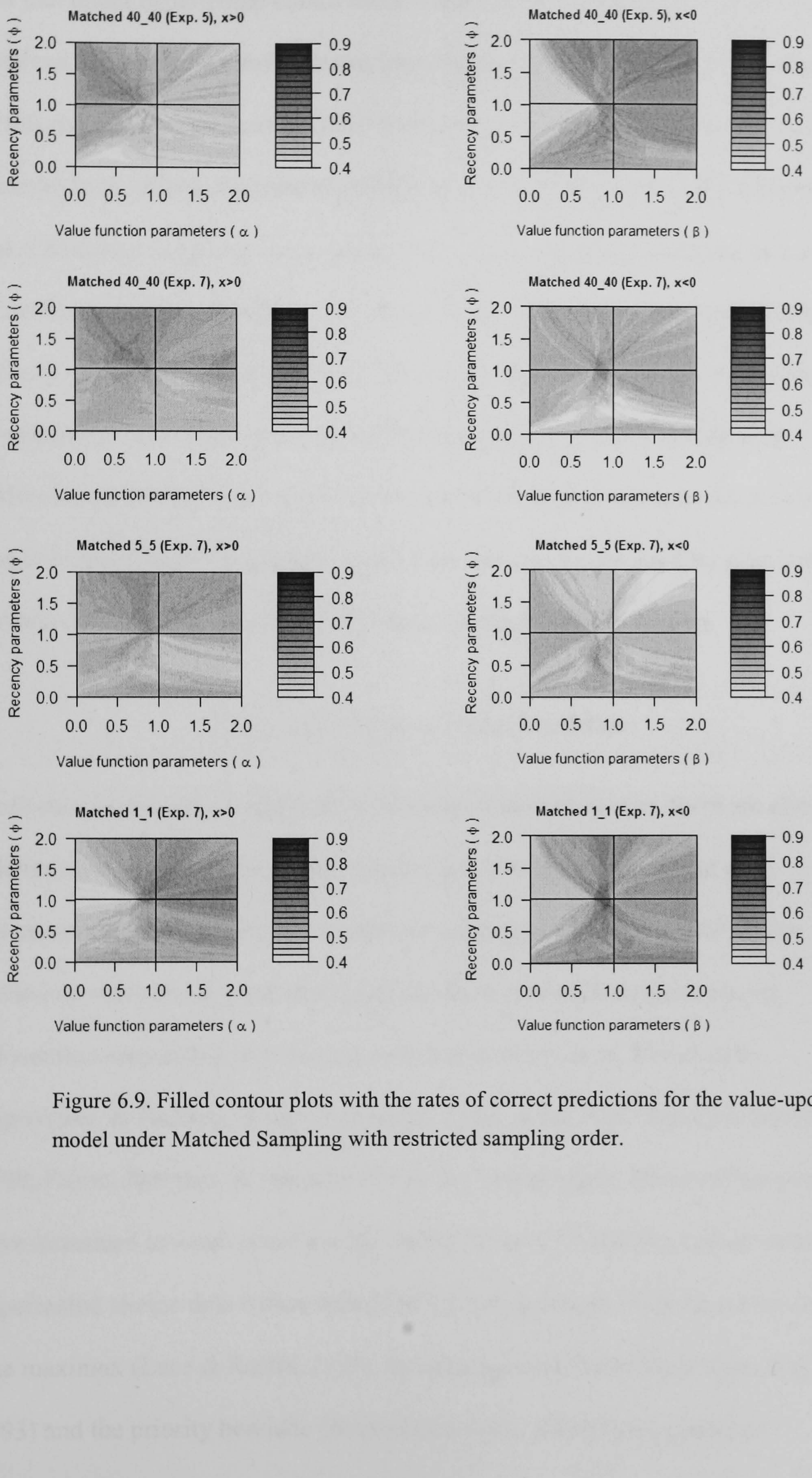


Figure 6.9. Filled contour plots with the rates of correct predictions for the value-updating model under Matched Sampling with restricted sampling order.

Altogether the results from this section have provided additional support for the fact that under experiential choice tasks a similar level of predictive power can be obtained using simple reinforcement learning models. Contrary to previous tests of this model in the context of DfXP under Free Sampling (Hau et al., in press; Hertwig et al., 2006), the parameters that obtain the optimal level of performance under Matched Sampling do generally not make it necessary to assume any form of nonlinear recency weighing. Nevertheless, although the performance comes close to the benchmark set by the PT model it would still be desirable to have an even better model. More complex extensions of these learning models with additional parameters have already been proposed in other contexts. However, it remains open whether a similar performance can also be obtained by even simpler models. This question will be the examined in the following section.

6.4 Predictions of choice heuristics

Until now I have only considered parametric models. However, there are also alternative approaches, for example heuristics, which can do without any parameters. Heuristics claim to model the actual processes behind the choice behaviour more closely and have been shown to predict descriptive choice phenomena reasonably well despite their frugal nature (e.g., Brandstätter, Gigerenzer, & Hertwig, 2006; Gigerenzer, Todd, & the ABC Research Group, 1999; Payne, Bettman, & Johnson, 1993). In a recent paper, Hau et al. (in press) have examined to what extent a wide variety of choice heuristics can account for experiential choice data within their Free Sampling design. They found heuristics like maximax (Luce & Raiffa, 1957), the lexicographic heuristic (Payne et al., 1993) and the priority heuristic (Brandstätter et al., 2006) to be candidate

strategies for DfXP equivalent to PT-based models. The highest fits obtained within their analysis was 69% for maximax, followed by the lexicographic heuristic and the natural mean heuristic, a process model version of EV maximization, with 68% correct predictions. These levels of predictive power seem to be compatible with the ones obtained here, which make these heuristics interesting candidate models in the context of the Matched Sampling paradigm as well.

In order to test whether they can actually perform equally well, I compared their rate of correct predictions with the ones reported in the earlier section in this chapter. Instead of the whole range of available strategies, I only selected a subset of six heuristics from the most successful candidates according to Hau et al's analysis for which it was possible to derive clear predictions within all six choice problems. A detailed description of the selected heuristics and their underlying processes is provided in Table 6.8.

The natural-mean heuristic was not included as it has already been discussed in the context of the fits for PT models with linear value and weighting function parameters.

TABLE 6.8

Descriptions of the chosen heuristics according to Hau et al. (in press)

Heuristic	Choice Policy Steps
Equiprobable	Step 1: Calculate the arithmetic mean of all experienced outcomes within a deck. Step 2: Choose the deck with the higher mean.
Equal Weight	Step 1: Calculate the sum of all experienced outcomes within a deck. Step 2: Choose the deck with the higher sum.
Maximax	Step 1: Choose the deck with the higher experienced maximum outcome.
Better than average	Step 1: Calculate the grand average of all experienced outcomes from all decks. Step 2: For each deck, count the number of outcomes equal to or above the grand average. Step 3: Choose the deck with the highest number of such outcomes.
Lexicographic	Step 1: Determine the most frequently experienced outcome of each deck. Step 2a: Choose the deck with the highest most frequent outcome. Step 2b: If both are equal, determine the second most frequent outcome of each deck, and select the deck with the highest (second most frequent) outcome. Proceed until a decision is reached.
Priority	Step 1: Examine the minimum gains experienced in the samples of outcomes from decks distributions A and B, respectively. If they differ by 1/10 (or more) of the maximum gain experienced (in both samples), stop examination and choose the deck distribution with the more attractive minimum gain; otherwise go to Step 2. Step 2: Examine the sample probabilities of the minimum gains. If the probabilities differ by 1/10 (or more) of the probability scale, stop examination and choose the deck distribution with the more attractive probability; otherwise go to Step 3. Step 3: Examine the maximum gain experienced in each deck distribution. Choose the deck distribution with the more attractive gain.

To calculate the rate of correct predictions, I first generated the predictions of the different models within each of the six decision problems under Matched Sampling (see Table 6.9). As the predictions of the equiprobable heuristic and the equal weight heuristic turned out to be identical in the context of the six choice problems, they were combined.

TABLE 6.9

Predictions of the different choice heuristics within the six choice problems used

	Options		Preferences predicted by the different heuristics				
Decision Problem	H	L	Equi-probable / Equal weight	Maxi-max	Better than average	Lexico-graphic	Priority
1	4, .8	3, 1.0	H	H	L	H	L
2	4, .2	3, .25	H	H	L	H	H
3	-3, 1.0	-32, .1	H	L	H	L	H
4	-3, 1.0	-4, .8	H	L	L	H	H
5	32, .1	3, 1.0	H	H	H	L	L
6	32, .025	3, .25	H	H	H	H	L

As the heuristics only take information into account which, as a result of the matching process, is identical for each participant, the predictions within the different decision problems are uniform for all participants. With a mean rate of 50% (SD = 6%) the heuristics did not perform better than chance. The highest mean rate was observed for the lexicographic heuristic (54%) followed by the priority heuristic with 53%. A more detailed summary of their performance within the different choice problems is given in Table 6.10.

TABLE 6.10

Performance of the choice heuristics across the different data sets

	Matched Sampling				40_40		1_1	5_5
	Exp. 2	Exp. 3	Exp. 4	Exp. 7	Exp. 5	Exp. 7	Exp. 7	Exp. 7
Equiprobable / Equal Weight	0.35	0.46	0.49	0.5	0.5	0.5	0.57	0.51
Maximax	0.49	0.46	0.54	0.49	0.46	0.5	0.55	0.51
Better than average	0.39	0.42	0.41	0.39	0.47	0.42	0.37	0.47
Lexicographic	0.54	0.5	0.56	0.59	0.48	0.57	0.6	0.5
Priority Heuristic	0.55	0.55	0.5	0.51	0.53	0.56	0.54	0.51

The success of the heuristics to account for experiential choice behaviour under Free Sampling could not be confirmed for the Matched Sampling paradigm. However, it is interesting to see that the low performance applies equally to strategies that incorporate probabilities, like the priority heuristics and strategies that do not use probabilities at all, for example the maximax heuristic. Problematic seems the uniform character of their predictions and the fact that strategies do not take into account any characteristics of the individual sequences. Instead, alternative strategies providing simple choice rules on the basis of sequential properties including patterns or other complex structures of the individual experiences should be explored.

6.5 An alternative model looking inside the sequence

I have shown that the performances of the most prominent decision making models still leave room for improvement. Furthermore, there seems to be some indication that one potential source of information that can be utilised for alternative models lies within the individual sequences. The value-updating model has provided one way of capturing this information. A less successful approach, the incorporation of the observed switching behaviour, has already been tested experimentally in Chapter 4. However, there are still other ways of modelling the representation of the sampled information. I therefore want to use the last section of this chapter to briefly explore a completely different approach that is unrelated to any of the models discussed so far. This idea was prompted by comments from a participant who described looking for patterns or specific sub-sequences of outcomes within the experienced sequence to increase the performance within the tasks. At first this looks like an unusual way of formulating the problem. However, similar observations have also been made in the context of probability learning experiments which led to the development of mathematical models to formalise the usage of runs and patterns in these tasks (e.g., Edwards, 1956; Goodnow, 1955; Nicks, 1959). As pointed out by Jones and Myers (1966), the usage of runs and patterns as strategies can actually give participants an advantage in contexts where the underlying sampling process is assumed not to be random. Furthermore, it has been shown that divergent beliefs regarding the randomness of experimental procedures may persist, despite being instructed otherwise (Braveman & Fischer, 1968; McCracken, Osterhout, & Voss, 1962).

The underlying rational of these models is based on one of the oldest known judgement biases which can be traced as far back as to the work of Laplace (1796/1951): the *gambler's fallacy*. It is also referred to as *negative recency* and describes the tendency of predicting the non-reinforced option based on the belief that a run without successes will be balanced out by following future successes. As I have already used the concept of recency in the context of the experiential chapters, it is important to distinguish the two. While the recency definition used earlier was based on the idea that particular parts of the sequence receive more weight, either as a strategy or due to cognitive limitations and memory effects, negative recency is defined with regard to reinforcement of the last outcome. Negative recency does therefore not imply that some information was not retrieved or ignored, the opposite, it derives alternative predictions based on all previously experiences sequences, giving each outcome a similar weight. Negative recency would therefore not have been detected by a recency tests in the form of the sequence-split analyses presented earlier.

This concept can be applied to the appearance of runs within the sequences under DfXP as it provides a number of appealing properties that could explain some of the observed choice phenomena. For example, the tendency to choose options with rare but high payoffs more often under DfXP could be explained by negative recency as participants might anticipate a long run of the alternative non-rare outcome to be balanced out again. In addition, as sequential properties like the experienced numbers of runs and run lengths differ considerably between subjects due to the sampling process, this approach allows the modelling of variations of choice behaviour across participants. I will therefore briefly outline a particularly suitable run-based model before, given

restrictions of time and space, presenting one of many possible applications of the model within the matched sampling paradigm.

6.5.1 The application of a run-based model

The approach I have selected to test in the context of DfXP is based on a model by Restle (1961), which was originally proposed to account for negative recency effects observed in the probability learning experiments and which seems to be derived from earlier ideas by Goodnow, Rubinstein and Lubin (1960). It provides a clearly formalised mechanism which facilitates the transfer to a different domain. In Restle's (1961) original theory of patterns in guessing the anticipation of a binary outcome *a* or *b* is determined by matching the currently experienced sequence of events (for example *abbbb*) with sequences experienced earlier from memory. The greater the number of previous encounters with sequences of run lengths greater than the current length (*k*), the greater the probability of predicting the same outcome again (e.g., b). Conversely, if run lengths higher than *k* have not been encountered before, a shift in the response pattern (e.g., a) will follow. The expectancy of outcome *b* after trial *t+1* does therefore not primarily depend on the actual probability of *b* but is instead determined by the probability of the schemata with run length $k + 1$. It is obvious that the probability of an outcome and the distribution of the expected run lengths are not completely independent. However, there are conditions under which both make different predictions.

Restle's (1961) generalisation of this model has the following form:

$$P(B \mid k) = \frac{\sum_{j=k+1}^{\infty} W_j}{\sum_{i=k}^{\infty} W_i} \tag{6.6}$$

where the probability of response B after k runs of b is equal to the ratio of the weighted sums of schemata of run length $>k$ and schemata with run length k. The weights W are added to model the saliency of different run lengths in memory and are set to k which implies that longer runs are assumed to be more salient than shorter runs.

6.5.2 Test of a run-based model

In order to apply such a model in the context of DfXP, it has to be reformulated and adapted to the different structure of the task. First, predictions are only made for the trial that would follow after the learning phase ($t = 40+1$) and which represents the last draw from the option to determine the outcome that is actually added to the final score. Second, instead of two outcomes a and b that appear with probability p and $1-p$ we have actually two such pairs, one within each choice option. Consequently, it has to be determined separately for each option which of the two outcomes is more likely to appear in the last trial. Furthermore, a choice rule has to be introduced to model the actual decision between the two pairs of prospects. Following Restle's approach the model was divided into the following steps:

1. Extract the last encountered schemata from the sampled sequence separately for both choice options (A and B).
2. Determine the run length of these schemata (k_A and k_B) and the event included in the run (x_A and x_B).

3. Calculation of the probabilities $P(x_A \mid k_A)$ and $P(x_B \mid k_B)$ according to formula 6.6.
4. Determine the most likely outcome for the next trial for each option: If $P(x_A \mid k_A) > .5$ then x_A is predicted to appear in the final trial. Otherwise the alternative event y_A is predicted to be the outcome of the final trial.
5. Choice rule: The option with the highest predicted outcome is chosen.

The addition of steps 4 and 5 makes the model deterministic and allow the calculation of rates of correct predictions similar to the previous model tests. Evaluations of this model were conducted on the basis of the data from all the Matched-Sampling Conditions, including experiments with one and all six decision problems and experiments with predetermined sampling order. For the cases in which participants had more than one choice problem the frequencies of the different run length were assumed not to be affected by the run lengths experienced in previous choice problems. As all the components of the model emerge from the actual properties of the individual sequences, it was not necessary to estimate any parameters.

The mean rates of correct predictions obtained from this analysis, however, did not exceed chance level (51%). The predictive power of the model across the different data sets has been summarised in Table 6.11.

TABLE 6.11

Rates of correct predictions based on the run-based model for the Matched Sampling data

	Free sampling order				40_40 Order		1_1	5_5
	Exp. 2	Exp. 3	Exp. 4	Exp. 7	Exp. 5	Exp. 7	Exp. 7	Exp. 7
Rate of correct predictions	0.55	0.52	0.47	0.51	0.43	0.51	0.51	0.54

A version of the model without any salience weighting was also tested but did not offer a significant improvement of the rate of correct predictions. Although the model does not provide the same level of predictive power as the other approaches it has shown that there are alternative properties of the sequence that can be explored and utilised for the modelling of the cognitive processes driving decisions from experience.

6.6 Discussion

The first part of this chapter has provided model fits for PT which offered further insights regarding the shapes of the transformations of the model that would be able to account for DfXP. Furthermore, from these shapes it was possible to infer the apparent underweighting of small probabilities which brings together the different experimental findings and emphasises once more that PT cannot be generalised across DfXP without assuming different parameters for DfD and DfXP. Moreover, the estimations for the two-stage model have confirmed the results of the Matched-Sampling Conditions suggesting that such an adaptation of the parameter values cannot be avoided by resorting to subjective probability estimates.

Despite the rather low benchmark, set by the predictive power of the PT-based approaches, the model tests in the second part of the chapter have shown

that it is difficult to outperform them. Nevertheless, the application of the value-updating model has revealed that it is possible to obtain identical fit under a model that does not require any assumptions regarding the weighting of probabilities. The summary of the maximum rate of correct predictions for the tested models in Figure 6.10 illustrates also that the application of simple choice heuristics in the context if the Matched Sampling design was not as fruitful as suggested by the results of Hau et al (in press).

The same has to be said for the run-based model. Although unsuccessful in predicting the observed preferences within DfXP the application of this model has shown that there are a range of alternative approaches that should be considered and tested, either by being integrated within existing models or separately.

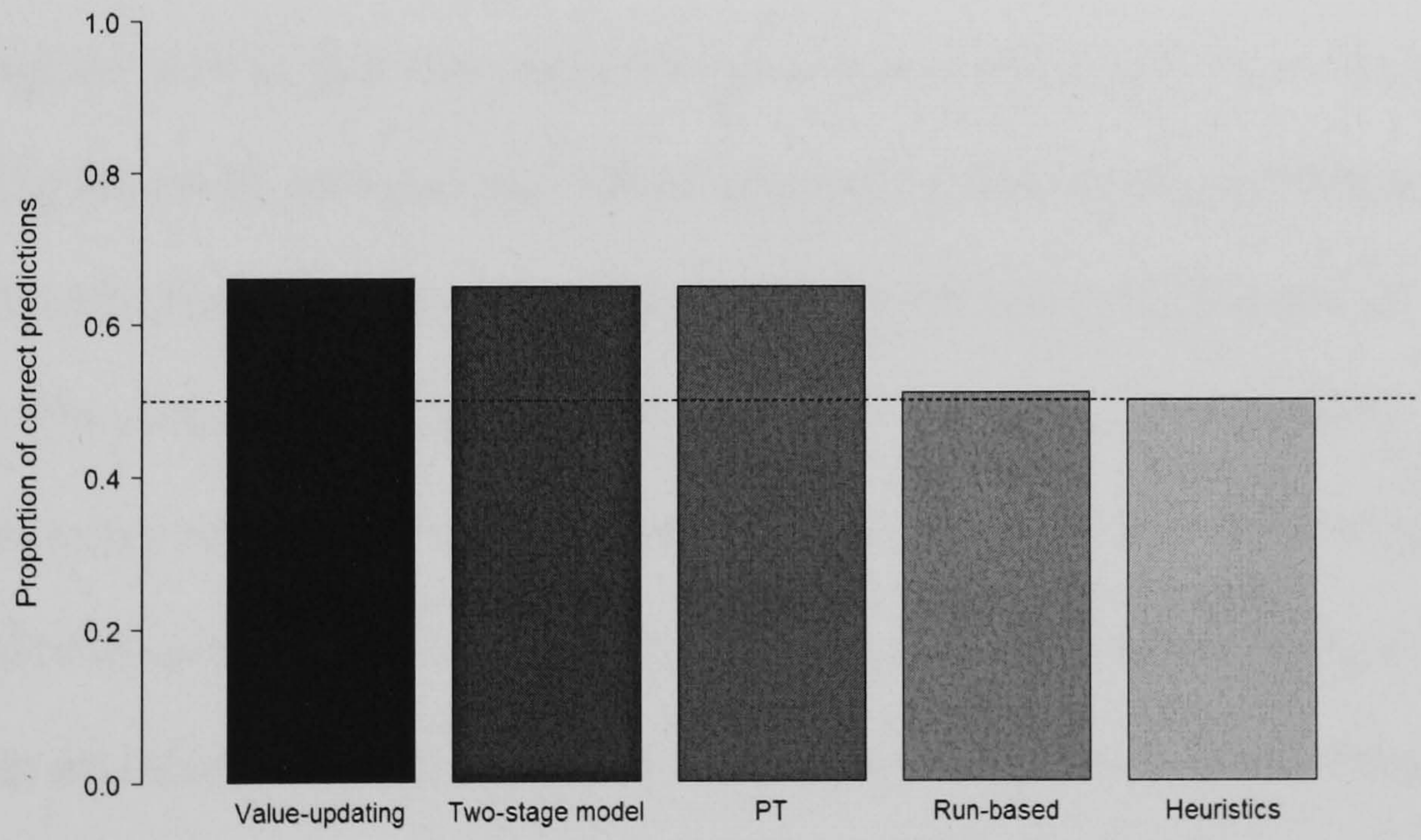


Figure 6.10. Summary of the mean proportions of correct predictions for the tested models. The dotted line indicates chance performance.

However, there are also obvious limitations of the presented parameter estimations and their comparisons. First, it has to be pointed out that the rather low levels of maximum fits might be due to an increased level of noise which may be partly circumvented by using bigger sets of problems. This would also allow

for an estimation of individual parameters, rather than fitting at an aggregated level with data that is pooled across participants, and would get rid of the variance due to individual differences

Furthermore, it has to be assumed that the generally lower upper limits are also a result of the usage of the actual rate of correct predictions as a criterion instead of the agreement on the level of predicted choice proportions, as it is usually reported in descriptive choice tasks. However, this would also apply to the previous parameter estimations that have been conducted in this context (e.g., Fox & Hadar, 2006; Hau et al., in press). With regard to the evaluation of the performance of the models it is important to note, that there is an inevitable "upper limit" for every model that predicts modal choice without somehow taking into account individual differences. An estimation of this upper limit that might be expected for such models was calculated from the average proportion of people choosing the modal option in the choices reported by Kahneman and Tversky (1979), taken from the summary table provided by Stewart and Simpson (in press). The average is 77.4%, which means that even a model that correctly predicts every modal choice (i.e., higher subjective value for the preferred option) it would only score 77.4% correct across people's choices. Without being able to account for individual differences, one cannot get beyond this number. This applies to the PT models and the heuristics and suggests that the performance might not be as bad as indicated by the distance to the 100%-line. However, this does not apply to the value-updating model and the run-based model, which try to incorporate properties that are unique to the individual sequence.

More generally, a side effect of the functional form of the PT model is that the same level of optimal fit can always be obtained under different parameter

combinations. This means that the shape of both functions can be traded off against each other which makes it very difficult to determine which shape to prefer over the other (see also Stott (2006) for a further discussion of this issue). This problem even holds for non-parametric approaches (for details see Gonzalez & Wu, 1999). Nonetheless, given the converging evidence from the choice proportions and the constraints of the decision biases it was possible to refine the results at least slightly further.

Finally, it is important to emphasise that this was by no means intended to be a comprehensive test of all candidate models. There is a range of alternatives that have been suggested in the literature but which do not lend themselves to a direct application within the Matched Sampling paradigm, especially models that have been developed to account for feedback based decisions or bandit problems with inherent trade-offs between exploration and exploitation of an option (e.g., Denrell, 2007; Erev & Barron, 2005). An adoptable approach that should be considered in addition to the above mentioned models is the recently proposed primed-sampler model (Erev et al., 2008). This is an extension of the value-updating model which incorporates the idea that people evaluate choices by drawing an even smaller mental sample from the input they have received. However, this could conflict with the observation presented here, that participants are well aware of the overall frequencies within experienced samples. Another dynamic model that does not require a value- or weighting function is the decisions by sampling model (Stewart et al., 2006) which has also been shown to provide a good predictions under descriptive choice. However, with the extension of the range of models considered it becomes also important to find new experimental conditions that allow distinguishing the predictions of these models.

Potential implications and suggestions emerging from the presented work regarding the design of such tests will be discussed in the concluding chapter of this thesis.

CHAPTER 7

GENERAL DISCUSSION AND CONCLUSIONS

7.1 Summary of empirical findings and their contributions

The research presented in the thesis investigated whether the phenomenon of underweighting of small probabilities, as it has been observed in decisions from experience (e.g., Hertwig et al., 2004; Weber et al., 2004), can be explained by the accounts that have been put forward in the literature thus far. One of these explanations has been the reliance on small samples (Hertwig et al., 2004) which can result in systematic underrepresentation of the actual frequencies of rare events. Another explanation, recency weighting, is based on the assumption that more recent outcomes, which are also more likely to be the non-rare events, receive more weight than earlier outcomes (Barron & Erev, 2003; Hertwig et al., 2004). Alternatively, Fox and Hadar (2006) proposed that the effect could also be a result of judgement error, the difference between the actually experienced probability and the estimated subjective probability. Yet, their experimental results did not confirm this claim. Instead, they found evidence suggesting that the phenomenon is a result of sampling error, the difference between the experienced probability and the actual objective probability. Furthermore, Fox and Hadar (2006) argue that the phenomenon can therefore simply be accounted for by applying prospect theory in the form of the two-stage model, predicting choices on the basis of subjective probabilities estimates rather than objective probabilities, thereby implying equivalence in terms of the underlying processes involved in both decisions from experience and decisions from description.

However, the crucial experimental test of this hypothesis within a design that allows controlling for sampling error is yet to be delivered.

The first strand of this research was intended to investigate these experimental conditions with varying degrees of inherent sampling error. The results obtained, however, cast serious doubt upon the claimed equivalence of the two different choice formats and the cognitive processes involved. Instead, the apparent underweighting of small probabilities appeared to be a robust phenomenon of experiential choice tasks even under the absence of sampling error. Consequently, it was the endeavour of the second part of this research to explore alternative explanations for this phenomenon. This was done through both additional experimental work and, more theoretically, through the application of various candidate models. The empirical findings from the individual chapters can be summarised as follows.

Chapters 2 provided a first experiment which included both a replication of the initial experiment by Hertwig et al. (2004) and an extension of the design in the form of a Comprehensive-Sampling Condition. This meant that it was possible to test whether the effect is limited to conditions under which people only sample a few items resulting in underrepresentation of the probabilities of the rare events. Although sampling error was significantly reduced under Comprehensive-Sampling, the apparent underweighting of small probabilities typical for DfXP was replicated under both conditions. In addition, there was no evidence for recency weighting under any of the conditions. This confirmed the robustness of the apparent underweighting of small probabilities and raised questions regarding the validity of the sampling error explanation put forward by Fox and Hadar (2006). However, the design did not allow eliminating sampling error.

In Chapter 3, I therefore introduced the Matched-Sampling design as a new experimental paradigm that permitted the investigation of decisions from experience without the distortion of sampling error. The first experiment, employing this design together with a Descriptive-Choice Condition and a Free-Sampling Condition, provided initial evidence that choice behaviour different to decisions from description can be observed even without the presence of sampling error. A comparison with the Free-Sampling Condition revealed that the effect under Matched-Sampling was attenuated but still significant. Sampling error has therefore been found to moderate the underweighting of small probabilities, but it is not the complete explanation of the phenomenon. Furthermore, frequency and probability estimations collected in the context of two additional Matched-Sampling experiments could show that there was no systematic underestimation of rare events. Rather, the estimates were found to be well-adjusted with a slight tendency to overestimate small probabilities in accordance with the results typically reported in the judgement literature. It can therefore be concluded that estimation error is not an explanation either. Further, with the equivalence of the estimates and the actual probabilities, the two-stage model makes the same predications as prospect theory and can therefore not improve its performance.

Given this elimination of all existing explanations, Chapter 4 was dedicated to exploring potential alternatives. Following an observation of extensive switching between options within the previous experimental work, the chapter explored the impact of different sampling orders and the resulting partitioning of the experienced sequences. However, even when the two options could only be evaluated separately without switching, which is more equivalent to a descriptive choice task, the apparent underweighting was still observed. In

addition, the design in Experiment 6 and 7 also allowed for an initial investigation of the reversal of choice patterns between the DfD and DfXP within participants. The analysis of the proportions of DfXP and Non-DfXP reversals though did not provide conclusive results.

A re-analysis of the data in the context of classical decision biases was provided in Chapter 5. Instead of the usual common ratio effect observed under DfD, I found evidence for a reversed common ratio effect under Free-Sampling Conditions and an attenuated intermediate effect under Matched-Sampling Conditions, echoing the findings on the difference in terms of the probability weighting function within DfXP throughout the experimental work presented. A new discovery was the finding of a reversed reflection effect in the context of DfXP, indicating differences with regard to the value transformation.

Furthermore, from this analysis a few constraints could be derived regarding the shape of both transformations that would allow prospect theory to account for the data. This is a value function that is concave for losses and convex for gains to incorporate the reversed reflection effect and a weighting function that is either slightly steeper between .2 and .25 than between .8 and 1.0., or has equal slopes between both points, accommodating the apparent underweighting of small probabilities and the reversed or attenuated common ratio effect.

Finally, Chapter 6 provided additional and more comprehensive model tests which could accommodate most of the findings in the previous chapters regarding the suitability of prospect theory. The overall fit was lower than usually observed under descriptive choice. Also, the inferences from the optimal parameters for both transformations under a valid prospect theory framework matched the constraints derived from the decision biases analysis and the apparent

underweighting observed for the differences in choice proportions between DfXP and DfD. The same results were also obtained for the two-stage model, again validating earlier observations. Crucially, as predicted from the combined findings, the optimal parameters under DfXP differed noticeably from parameters obtained for the available descriptive choice data.

Apart from the analysis of PT, it could also be shown that a simple reinforcement model like the value-updating model by Hertwig et al. (2006), which does not make any assumptions about probability weighting, can also explain the data just as well as PT. Simple heuristics and an application of a run-based model, on the other hand did not seem to provide predictive power beyond chance level.

7.2 Conclusions

With the results summarised in the previous section, this thesis has provided wide-ranging evidence for the robustness of the differences between decisions made on the basis of simple lottery descriptions and decisions from experience, where information regarding the outcomes and probabilities of the choice alternatives are not explicitly provided but have to be inferred from an experienced sequence of outcomes. The different behaviour observed under decisions from experience seems to imply underweighting of small probabilities.

As the current literature is dominated by a discussion around the sampling error hypothesis of Fox and Hadar, the introduction of the Matched Sampling paradigm and the resulting rejection of sampling error as a complete explanation is a significant contribution to the field. Indeed, this will shift the focus of future research. Furthermore, this design should be embraced as a replacement for the

Free-Sampling design and will hopefully inspire the development of additional extensions facilitating the identification of the relevant properties and the underlying processes.

The absence of any form of recency weighting throughout the seven experiments presented here can be seen as strong evidence against any involvement of it in the phenomenon. However, there are of course other ways of measuring recency weighting. In the analyses reported here, I have been following the original approach of Hertwig et al. (2004) by predicting choices on the basis of the expected value of earlier and later parts of the sequences. Alternatively, one could use different models to generate the predictions, for example prospect theory or one of the choice heuristics described in Chapter 6. Nevertheless, the findings reported here seem to echo the results of other very recent studies by Hau et al. (in press) and Rakow, Demes and Newell (in press) who also report similar effects without the coexistence of recency effects. Finally, there was also converging evidence from the estimations for the value-updating model, which verify that close to optimal performance of the model in most of the data sets can be obtained without any form of recency weighting. The consistent observation of well-adjusted frequency and probability estimates is also an important finding which is in line with results by Fox and Hadar (2006). Again, this emphasises the convergence of the PT and the two-stage model under DfXP.

In summary, the first four experiments of this thesis showed that all the explanations put forward in the literature thus far are not sufficient to account for the choice pattern under decisions from experience. It was therefore the challenge of the second part to identify alternative accounts. Although the investigation of the sampling order effects did not lead to an ultimate explanation of the effect, the

deduction of its design from the observation of the switching behaviour will hopefully direct attention towards the structural differences of the experiential tasks and facilitate a further investigation of the specific properties of the sequential accumulation of the incorporated information over time.

In the context of the analysis of the decision biases in Chapter 5, another significant contribution was the examination of the reflection effect under DfXP. The observation of a reversed reflection effect under Matched Sampling suggests that the underlying differences are not necessarily restricted to a difference in probability weighting, as it is discussed in the current literature, but are based on variations in the transformations of both values and probabilities. This was also confirmed by the optimal parameters found for the PT and the two-stage model in Chapter 6 which offered the first comprehensive parameter estimation on experiential choice data that could actually match the observed deviations from the PT predictions with congruent parameter values. Furthermore, as the optimal parameters for the available descriptive choice data could confirm the properties usually observed for PT, this thesis has provided the first comprehensive evaluation of the two formats that could track the differences between them from raw choice proportions over the differences in decision biases to the level of contrasting sets of optimal parameters values.

Taken together, the results indicate that the functional form of PT-based models can, to some extent, account for experiential choice data. However, in order to do so drastic adjustments of the parameters are required. As the implications of these changes are also diametrical to the proposed mechanisms under descriptive choice it seems necessary to extend these models. Without a plausible rationale for the shift in parameters when applied to experiential choice

tasks the PT model seems to maintain its descriptive quality but loses its explanatory quality. Therefore, the derivation of simple reinforcement models in the context of DfXP, which have been proven to account for the data as well as PT-based models, seem more intuitive. Further, the learning models seem to be flexible and extendable enough to accommodate the different properties of experiential tasks. As the data and designs presented here do not allow any further conclusions regarding their validity, it will be left to future research to provide additional model comparisons under conditions that demand distinguishable predictions from the different models.

In terms of the experimental methods employed, including laboratory and Web-based experiments, it could be shown that the difference between DfD and DfXP, although attenuated, can be replicated across a wider range of demographics outside the laboratory. However, it appeared to be easier to replicate the experiential choice proportions than the classical choice patterns under descriptive choice tasks within the Web-based experiments. As the gamble descriptions have been studied extensively in the lab, it could be argued that this is most likely due to special characteristics of the chosen samples. This is in line with observations by Birnbaum (1999), who demonstrated in a series of decision making experiments on the Internet, involving lottery descriptions, that student samples were generally more biased than Internet samples.

7.3 Limitations and Future directions

Nevertheless, besides these notable contributions, it is also important to point out a few limitations of the research that should be addressed by future investigations. The first limitation relates to the range of choice problems that have been

employed so far. As the research presented here was seeking to replicate the effect under Matched Sampling resembling the original conditions as closely as possible, I decided to use the same choice problems used by Hertwig et al. (2004), Fox and Hadar (2006), and Barron and Erev (2003). Within future experiments, however, it would be advantageous to see results from a wider range of choice problems, including mixed gambles and a more comprehensive range of probabilities. This would allow a more systematic and fine grained investigation of the transformation of probabilities in experiential tasks. A first step in this direction seems to have been made in the most recent studies (Erev et al., 2008; Hau et al., in press; Rakow et al., in press). It would also be important to test how far these results can be generalised to decisions involving more than two outcomes.

An area that could only be touched briefly was the investigation of alternative representations of the task. Although the exploration of sub-sequences (Chapter 4) and the analysis of runs (Chapter 6) were not especially fruitful, it could be shown that the explored sequences do provide a rich source of alternative properties that can be utilised in order to model choice behaviour in decisions from experience. This includes observations of the information search and the use of individual stopping rules, which might also help to explain some of the differences found between Free Sampling, marked by a self-determined end of the exploration, and Matched Sampling.

However, findings by Rakow, Demes and Newell (in press) seem to suggest that neither sequential information nor the actual experience itself can explain the phenomenon. They observed the same choice behaviour of underweighting of small probabilities under conditions in which participants were

merely provided with the raw frequencies from another participant's observations (e.g., (1) 4 points on 7 out of 10 occasions, 0 points on 3 out of 10 occasions, versus (2) 3 points on 5 out of 5). This, in turn, contradicts the findings of Simonson, Karlson, Loewenstein and Ariely (2008) who showed, in the context of the repeated play of games like the Weak-link and the Prisoner's Dilemma, that experienced information has more impact on actual behaviour than information, identical in content, format and relevance that was merely observed. It also goes against the explanation put forward by Erev et al. (2008) claiming that rare events are only underweighted when they are neglected and not made explicit. The presentation of the frequencies makes them explicit but it does not reduce the underweighting.

It is also interesting to note that the results presented here seem to contradict the "frequentist hypothesis" (Cosmides & Tooby, 1996), which contends that the presentation in a frequency format eliminates biases usually observed under the presentation of probabilities. Instead of unbiased transformations under DfXP, we actually find a reversal of the biases normally observed when providing explicit probabilities in descriptive choice tasks. It would therefore be interesting to examine whether both biases can be compensated when combining the two formats. The results from Experiment 4, in which participants had to give probability judgements before choosing their preferred option, seem to suggest that this does not necessarily eliminate underweighting. It would be helpful to explore this more systematically, including alternative ways of combining the formats.

An issue that has not yet been considered in the literature are the prior assumptions that participants might have regarding the sampling processes behind

the two buttons. These could be assumptions that they bring into the experiments or assumptions that are triggered by the device metaphor used within the experiment. Some of the participants' comments after the completion of the tasks seem to indicate that despite the instructions, participants believed in the presence of patterns or other forms of dependencies between buttons. Computer tasks like the button design of the DfXP paradigm could be the reason for such alternative assumptions. In order to clarify whether this has any impact on the observed choice behaviour, it would be vital to examine the phenomena using tasks that trigger more natural representations of the stochastic properties of the sampling process. This could be either draws from shuffled decks of cards or draws from an urn with a known proportion of differently coloured balls.

Finally, an interesting series of findings that could also have implications for decisions from experience has been reported in research on decisions within human movement planning tasks involving rewards and penalties. A typical design for such tasks (e.g., Trommershäuser, Maloney, & Landy, 2003b) requires participants to repeatedly hit a small circular reward region which overlaps with an equally sized penalty region on a touch screen. Due to a time limit, movements cannot be executed with full accuracy resulting in residual motor variability around the selected target point. The selection of the point that maximises hitting the reward region and minimises hitting the penalty region under the constraints of individual motor uncertainty provides therefore conditions that are mathematically equivalent to choices among lotteries in decision making under risk (Maloney, Trommershäuser, & Landy, 2007). However, it could be shown that unlike in decision making under risk, the performance of participants within these motor planning tasks show a startling ability to maximise expected value

across a long series of trials (e.g., Trommershäuser, Gepshtein, Maloney, Landy, & Banks, 2005; Trommershäuser, Landy, & Maloney, 2006; Trommershäuser, Maloney, & Landy, 2003a; Trommershäuser et al., 2003b). As the task is repeated and the uncertainty implicit, it shares a lot of similarities with decisions from experience. Even more intriguing in the context of investigation of the differences between DfD and DfXP is an extension of the design introducing explicit stochastic rewards and penalties (similar to lottery descriptions) under which performance could be shown to suddenly drop below optimal (Maloney et al., 2007). A further investigation of these problems using an extended range of probabilities and configurations, including non-mixed lotteries and rare events, would allow testing of whether there is also evidence for underweighting of small probabilities comparable with the result reported in DfXP and whether the differences found for implicitly experienced forms of uncertainty can be generalised across a wider range of modalities.

7.4 Applications

In everyday life we often lack explicit objective information regarding the risks that we are exposed to. Instead, people have to rely on personal experiences. Consequently, a better understanding of the cognitive processes dominating the behaviour in experiential contexts might also enhance the understanding of human choice in general. As suggested by the title of this thesis and the results reported throughout, the reliance on experience might come with the burden of a different bias implying underweighting of small probability events. What this means for an individual decision depends of course on the context. However, as rare events are often also events which have an proportionally high impact (Taleb, 2007), it is

important to investigate potential implications of decisions from experience in the context of everyday live, like the usage of safety devices or the consideration of health risks, to name but a few.

Initial applications have been provided in different domains, including the perceived risks constituted by global warming (Weber, 2006), risk assessment of industrial accidents (Barkan, Zohar, & Erev, 1998), the usage of safety devices (Yechiam, Erev, & Barron, forthcoming), and the fear of terrorist attacks (Yechiam, Barron, & Erev, 2005). These examples indicate the far-ranging implications of this new strand of decision making research. We may reasonably hope that further applications will follow.

REFERENCES

Abdellaoui, M. (2000). Parameter-Free Elicitation of Utility and Probability Weighting Functions. *Management Science, 46*, 1497-1512.

Adobe.com. (n.d.). *Flash Player Penetration.* Retrieved November 4, 2007, from http://www.adobe.com/products/player_census/flashplayer/

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

Anderson, N. H. (1960). Effect of first-order conditional probability in a two-choice learning situation. *Journal of Experimental Psychology: General, 59*, 73-93.

Anderson, N. H., & Whalen, R. E. (1960). Likelihood judgments and sequential effects in a two-choice probability learning situation. *Journal of Experimental Psychology: General, 60*, 111-120.

Barkan, R., Zohar, D., & Erev, I. (1998). Accidents and decision making under uncertainty: A comparison of four models. *Organizational Behavior and Human Decision Processes, 74*, 118-144.

Barlow, R. E. (1972). *Statistical inference under order restrictions*. London: Wiley.

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

Battalio, R. C., Kagel, J. H., & MacDonald, D. N. (1985). Animals' choices over uncertain outcomes: Some initial experimental results. *The American Economic Review, 75*, 597-613.

Bernoulli, D. (1954). Exposition of a new theory on the measurement of risk. *Econometrica, 22*, 23-36.

Berry, D. A., & Fristedt, B. (1985). *Bandit problems sequential allocation of experiments*. London; New York: Chapman and Hall.

Binmore, K. (1999). Why Experiment in Economics? *The Economic Journal, 109*, 16-24.

Birnbaum, M. H. (1999). Testing critical properties of decision making on the Internet. *Psychological Science, 10*, 399.

Birnbaum, M. H. (2000). *Psychological experiments on the Internet*. San Diego, CA: Academic Press.

Birnbaum, M. H. (2001). *Introduction to behavioral research on the Internet*. Upper Saddle River, N.J.: Prentice Hall.

Birnbaum, M. H. (2004). Human research and data collection via the Internet. *Annual Review of Psychology, 55*, 803-832.

Birnbaum, M. H. (2007). New paradoxes of risky decision making. Unpublished Manuscript submitted for publication.

Birnbaum, M. H., & Martin, T. (2003). Generalization across people, procedures, and predictions: Violations of stochastic dominance and coalescing. In S. L. Schneider & J. Shanteau (Eds.), *Emerging perspectives on decision research* (pp. 84-107). New York: Cambridge University Press.

Birnbaum, M. H., & Wakcher, S. V. (2002). Web-based experiments controlled by JavaScript: An example from probability learning. *Behavior Research Methods Instruments & Computers, 34*, 189-199.

Bleichrodt, H. (2001). Probability weighting in choice under risk: An empirical test. *Journal of Risk and Uncertainty, 23*, 185-198.

Bower, G. H. (1994). A turning point in mathematical learning theory. *Psychological Review, 101*, 290-300.

Bowker, A. H. (1948). A test for symmetry in contingency tables. *Journal of the American Statistical Association, 43*, 572-574.

Brackbill, Y., & Bravos, A. (1962). Supplementary report: The utility of correctly predicting infrequent events. *Journal of Experimental Psychology: General, 64*, 648-649.

Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2006). The priority heuristic: Making choices without trade-offs. *Psychological Review, 113*, 409-432.

Braveman, N. S., & Fischer, G. J. (1968). Instructionally induced strategy and sequential information in probability learning. *Journal of Experimental Psychology: General, 76*, 674-676.

Brehmer, B. (1980). In one word: Not from experience. *Acta Psychologica, 45*, 223-241.

Buchanan, T. (2000). Potential of the Internet for personality research. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 121-140). San Diego: Academic Press.

Buchanan, T., & Smith, J. L. (1999). Using the Internet for psychological research: Personality testing on the World Wide Web. *British Journal of Psychology, 90*, 125.

Busemeyer, J. R. (1985). Decision making under uncertainty: A comparison of simple scalability, fixed-sample, and sequential-sampling models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*, 538-564.

Bush, R. R., & Mosteller, F. (1955). *Stochastic models for learning*. New York: Wiley.

Camerer, C. F. (1997). Rules for experimenting in psychology and economics, and why they differ. In W. Albers, W. Guth, P. Hammerstein, B. Moldovanu & E. Van Damme (Eds.), *Understanding strategic interaction: Essays in honor of Reinhard Selten* (pp. 313-327). Berlin: Springer-Verlag.

Christensen-Szalanski, J. J., & Beach, L. R. (1982). Experience and the base-rate fallacy. *Organizational behavior and human performance, 29*, 270-278.

Chu, Y. P., & Chu, R. L. (1990). The subsidence of preference reversals in simplified and marketlike experimental settings - a Note. *American Economic Review, 80*, 902-911.

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

Dayton, C. M., & Schafer, W. D. (1973). Extended tables of t and chi square for bonferroni tests with unequal error allocation. *Journal of the American Statistical Association, 68*, 78-83.

Denrell, J. (2007). Adaptive learning and risk taking. *Psychological Review, 114*, 177-187.

Dougherty, M. R., & Franco-Watkins, A. M. (2002). A memory models approach to frequency and probability judgement: Applications of Minerva 2 and Minerva DM. In P. Sedlmeier & T. Betsch (Eds.), *ETC Frequency processing and cognition* (pp. 121-136). New York, NY: Oxford University Press.

Edwards, W. (1956). Reward probability, amount, and information as determiners of sequential two-alternative decisions. *Journal of Experimental Psychology: General, 52*, 177-188.

Edwards, W. (1961). Probability learning in 1000 trials. *Journal of Experimental Psychology: General, 62*, 385-394.

Einhorn, H. J. (1980). Learning from Experience and Suboptimal Rules in Decision Making. In T. S. Wallsten (Ed.), *Cognitive Processes in Choice and Decision Behavior* (pp. 1 - 20). Hillsdale, NJ: Erlbaum.

Einhorn, H. J., & Hogarth, R. M. (1978). Confidence in judgment: Persistence of the illusion of validity. *Psychological Review, 85*, 395-416.

Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *Quarterly Journal of Economics, 75*, 643-669.

Erev, I., & Barron, G. (2005). On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological Review, 112*, 912-931.

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

Erev, I., & Haruvy, E. (in preparation). Learning and the Economics of Small Decisions. In J. H. Kagel & A. E. Roth (Eds.), *The Handbook of Experimental Economics* (Vol. 2). Princeton: University Press.

Erev, I., & Roth, A. E. (1999). On the role of reinforcement learning in experimental games: The cognitive game-theoretic approach. In D. V. Budescu, I. Erev & R. Zwick (Eds.), *Games and Human Behavior: Essays in Honor of Amnon Rapoport* (pp. 53-78). Mahwah, N.J: Lawrence Erlbaum Associates.

Erlick, D. E. (1964). Absolute judgments of discrete quantities randomly distributed over time. *Journal of Experimental Psychology: General, 67*, 475-482.

Estes, W. K. (1950). Toward a statistical theory of learning. *Psychological Review, 57*, 94-107

Estes, W. K. (1959). The statistical approach to learning theory. In S. Koch (Ed.), *Psychology: a study of a science* (Vol. 2, pp. 380-491). New York: McGraw-Hill.

Estes, W. K. (1976). The cognitive side of probability learning. *Psychological Review, 83*, 37-64.

Estes, W. K., Campbell, J. A., Hatsopoulos, N., & Hurwitz, J. B. (1989). Base-rate effects in category learning: A comparison of parallel network and memory storage-retrieval models. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 15*, 556-571.

Fiorina, M. P. (1971). A note on probability matching and rational choice. *Behavioral Science, 16*, 158-166.

Fleiss, J. L., Levin, B. A., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd ed. / Joseph L. Fleiss, Bruce Levin, Myunghee Cho Paik. ed.). Hoboken, N.J.: Wiley-Interscience.

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

Fox, C. R., & Tversky, A. (1998). A belief-based account of decision under uncertainty. *Management Science, 44*, 879-895.

Fraley, R. C. (2004). *How to conduct behavioral research over the internet : a beginner's guide to HTML and CGI/Perl*. New York ; London: Guilford.

Friedman, D., & Massaro, D. W. (1998). Understanding variability in binary and continuous choice. *Psychonomic Bulletin & Review, 5*, 370-389.

Fudenberg, D., & Levine, D. K. (1998). *The theory of learning in games*. Cambridge, Mass.: MIT Press.

Gigerenzer, G. (1989). *The Empire of chance : How probability changed science and everyday life*. Cambridge: Cambridge University Press.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychological Review, 102*, 684-704.

Gigerenzer, G., & Murray, D. J. (1987). *Cognition as intuitive statistics*. Hillsdale, NJ: Erlbaum.

Gigerenzer, G., Todd, P. M., & the ABC Research Group. (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.

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

Goodnow, J. J. (1955). Response-sequences in a pair of two-choice probability situations. *The American Journal of Psychology, 68*, 624-630.

Goodnow, J. J., Rubinstein, I., & Lubin, A. (1960). Response to changing patterns of events. *The American Journal of Psychology, 73*, 56-67.

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

Hammond, K. R., Summers, D. A., & Deane, D. H. (1973). Negative effects of outcome-feedback in multiple-cue probability learning. *Organizational Behavior and Human Performance, 9*, 30-34.

Hasher, L., & Chromiak, W. (1977). The processing of frequency information: An automatic mechanism? *Journal of Verbal Learning and Verbal Behavior, 16*, 173-184.

Hasher, L., & Zacks, R. T. (1979). Automatic and effortful processes in memory. *Journal of Experimental Psychology: General, 108*, 356-388.

Hasher, L., & Zacks, R. T. (1984). Automatic processing of fundamental information: The case of frequency of occurrence. *American Psychologist, 39*, 1372-1388.

Hau, R. C., Pleskac, T. J., Kiefer, J., & Hertwig, R. (in press). The description-experience gap in risky choice: The role of sample size and experienced probabilities. *Journal of Behavioral Decision Making.*

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*, 534-539.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2006). The Role of Information Sampling in Risky Choice. In K. Fiedler & P. Juslin (Eds.), *Information*

Sampling and Adaptive Cognition (pp. 72-91). New York: Cambridge University Press.

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

Hertwig, R., Pachur, T., & Kurzenhauser, S. (2005). Judgments of risk frequencies: Tests of possible cognitive mechanisms. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 31*, 621-642.

Hintzman, D. L. (1976). Repetition and memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 10, pp. 47-91). New York: Academic Press.

Hogarth, R. M., McKenzie, C. R. M., Gibbs, B. J., & Marquis, M. A. (1991). Learning from feedback: Exactingness and incentives. *Journal of Experimental Psychology: Learning Memory, and Cognition, 17*, 734-752.

Howell, W. C. (1973). Representation of frequency in memory. *Psychological Bulletin, 80*, 44-53.

Jones, M. R., & Myers, J. L. (1966). A comparison of two methods of event randomization in probability learning. *Journal of Experimental Psychology: General, 72*, 909-911.

Kacelnik, A., & Bateson, M. (1996). Risky theories: The effects of variance on foraging decisions. *American Zoologist, 36*, 402-434.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*, 263-292.

Keren, G., & Wagenaar, W. A. (1987). Violation of utility theory in unique and repeated gambles. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 13*, 387-391.

Klayman, J. (1988). On the how and why (not) of learning from outcomes. In B. Brehmer & C. R. B. Joyce (Eds.), *Human judgment: The SJT view* (pp. 115-162). Oxford, England: North-Holland.

Knight, F. H. (1921). *Risk, Uncertainty and Profit*. New York: Houghton Mifflin.

Krantz, J. H. (2001). Stimulus delivery on the web: What can be presented when calibration isn't possible. In U.-D. Reips & M. Bosnjak (Eds.), *Dimensions of Internet Science* (pp. 113-130). Berlin: Pabst Science Publishers.

Krantz, J. H., & Dalal, R. (2000). Validity of Web-based research. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet*. San Diego: Academic Press.

Laplace, P. S., de. (1796/1951). *A philosophical essay on probabilities*. New York: Dover.

Lee, W. (1971). *Decision theory and human behavior*. New York: Wiley.

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., & Combs, B. (1978). Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory, 4*, 551–578.

Loomes, G., & Sugden, R. (1982). Regret theory: An alternative theory of rational choice under uncertainty. *Economic Journal, 92*, 805-824.

Lopes, L. L. (1981). Decision making in the short run. *Journal of Experimental Psychology: Human Learning & Memory, 7*, 377-385.

Lopes, L. L. (1983). Some thoughts on the psychological concept of risk. *Journal of Experimental Psychology: Human Perception and Performance, 9*, 137-144.

Lopes, L. L. (1987). Between hope and fear: The psychology of risk. In L. Berkowitz (Ed.), *Advances in experimental social psychology, Vol 20* (pp. 255-295). San Diego, CA: Academic Press.

Luce, R. D., & Raiffa, H. (1957). *Games and decisions*. New York: Wiley.

MacDonald, D. N., Kagel, J. H., & Battalio, R. C. (1991). Animals' choices over uncertain outcomes: Further experimental results. *The Economic Journal, 101*, 1067-1084.

Maloney, L. T., Trommershäuser, J., & Landy, M. S. (2007). Questions without Words: A Comparison between Decision Making under Risk and Movement Planning under Risk. In W. D. Gray (Ed.), *Integrated Models of Cognitive Systems* (pp. 297-315). New York: Oxford University Press.

March, J. G. (1996). Learning to be risk averse. *Psychological Review, 103*, 309-319.

Marsh, B., & Kacelnik, A. (2002). Framing effects and risky decisions in starlings. *Proceedings of the National Academy of Sciences of the United States of America, 99*, 3352-3355.

McCracken, J., Osterhout, C., & Voss, J. F. (1962). Effects of instructions in probability learning. *Journal of Experimental Psychology: General, 64*, 267-271.

Myers, J. L. (1976). Probability learning and sequence learning. In W. K. Estes (Ed.), *Handbook of learning and cognitive processes: Approaches to*

human learning and motivation (Vol. 3, pp. 171–205). Hillsdale, NJ: Erlbaum.

Nash, H. (1964). The judgment of linear proportions. *The American Journal of Psychology, 77*, 480-484.

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal, 7*, 308-313.

Nicks, D. C. (1959). Prediction of sequential two-choice decisions from event runs. *Journal of Experimental Psychology: General, 57*, 105-114.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge; New York, NY, USA: Cambridge University Press.

Peterson, C. R., & Beach, L. E. (1967). Man as an intuitive statistician. *Psychological Bulletin, 68*, 29–46.

Pitz, G. F. (1966). The sequential judgment of proportion. *Psychonomic Science, 4*, 397-398.

Plott, C. R. (1996). Rational individual behaviour in markets and social choice processes: The discovered preference hypothesis. In K. J. Arrow, E. Colombatto, M. Perlman & C. Schmidt (Eds.), *The Rational Foundations of Economic Behaviour*. Basingstoke: MacMillan.

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

Quiggin, J. (1993). *Generalized expected utility theory : the rank-dependent model*. Boston: Kluwer Academic Publishers.

Rakow, T., Demes, K., & Newell, B. (in press). Biased samples not mode of presentation: Re-examining the apparent underweighting of rare events in

experience-based choice. *Organizational Behavior and Human Decision Processes*.

Real, L. A. (1991). Animal choice behavior and the evolution of cognitive architecture. *Science, 253*, 980-986.

Real, L. A. (1996). Paradox, performance, and the architecture of decision-making in animals. *American Zoologist 36*, 518-529.

Reimers, S., & Maylor, E. A. (2005). Task switching across the life span: Effects of age on general and specific switch costs. *Developmental Psychology, 41*, 661-671.

Reimers, S., & Stewart, N. (2007). Adobe Flash as a medium for online experimentation: A test of RT measurement capabilities. *Behavior Research Methods, 39*, 365-370.

Reips, U.-D. (2000). The Web experiment method: Advantages, disadvantages, and solutions. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 89-117). San Diego, CA: Academic Press.

Reips, U. D. (2001). The Web experimental psychology lab: Five years of data collection on the Internet. *Behavior Research Methods, Instruments, & Computers, 33*, 201-211.

Restle, F. (1961). *Psychology of judgment and choice; a theoretical essay*. New York: Wiley.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society, 58*, 527-535.

Roth, A. E., & Erev, I. (1995). Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. *Games and Economic Behavior, 8*, 164-212.

Sarin, R., & Vahid, F. (2001). Predicting how people play games: A simple dynamic model of choice. *Games and Economic Behavior, 34*, 104-122.

Savage, L. J. (1954). *The foundations of statistics*. New York,: Wiley.

Schmidt, W. C. (1997). World-wide web survey research: Benefits, potential problems, and solutions. *Behavior Research Methods, Instruments & Computers, 29*, 274-279.

Schmidt, W. C. (2001). Presentation accuracy of Web animation methods. *Behavior Research Methods, Instruments, & Computers, 33*, 187-200.

Schoemaker, P. J. H. (1990). Are risk-attitudes related across domains and response modes? *Management Science, 36*, 1451-1463.

Schwarz, N., & Wänke, M. (2002). Experiential and Contextual Heuristics in Frequency Judgment: Ease of Recall and Response Scales. In P. Sedlmeier & T. Betsch (Eds.), *Etc.: Frequency Processing and Cognition* (pp. 89-108). New York: Oxford University Press.

Sedlmeier, P. (1999). *Improving statistical reasoning: Theoretical models and practical implications*. Mahwah, NJ: Lawrence Erlbaum.

Shafir, S. (1994). Intransitivity of preferences in honey bees: support for 'comparative' evaluation of foraging options. *Animal Behaviour, 48*, 55-67.

Shanks, D. R. (1990). Connectionism and the learning of probabilistic concepts. *The Quarterly Journal of Experimental Psychology Section A, 42*, 209 - 237.

Shanks, D. R. (1995). *The psychology of associative learning*. Cambridge, UK: Cambridge University Press.

Shanks, D. R., Tunney, R. J., & McCarthy, J. D. (2002). A re-examination of probability matching and rational choice. *Journal of Behavioral Decision Making, 15*, 233-250.

Shanteau, J. (1978). When does a response error become a judgmental bias? Commentary on "Judged frequency of lethal events." *Journal of Experimental Psychology: Human Learning & Memory, 4*, 579-581.

Shuford, E. H. (1961). Percentage estimation of proportion as a function of element type, exposure time, and task. *Journal of Experimental Psychology, 61*, 430-436.

Siegel, S. (1959). Theoretical models of choice and strategy behavior: Stable state behavior in the two-choice uncertain outcome situation. *Psychometrika, 24*, 303-316.

Siegel, S., & Goldstein, D. A. (1959). Decision-making behavior in a two-choice uncertain outcome situation. *Journal of Experimental Psychology: General, 57*, 37-42.

Simonsohn, U., Karlsson, N., Loewenstein, G., & Ariely, D. (2008). The tree of experience in the forest of information: Overweighing experienced relative to observed information. *Games and Economic Behavior, 62*, 263-286.

Simpson, W., & Voss, J. F. (1961). Psychophysical judgments of probabilistic stimulus sequences. *Journal of Experimental Psychology, 62*, 416-422.

Skitka, L. J., & Sargis, E. G. (2005). Social psychological research and the Internet: the promise and peril of a new methodological frontier. In Y. Amichai-Hamburger (Ed.), *The Social Net: The Social Psychology of the Internet* (pp. 1-25). New York: Cambridge University Press.

Skitka, L. J., & Sargis, E. G. (2006). The Internet as psychological laboratory. *Annual Review of Psychology, 57*, 529-555.

Starmer, C. (2000). Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of Economic Literature, 38*, 332-382.

Stevens, S. S., & Galanter, E. H. (1957). Ratio scales and category scales for a dozen perceptual continua. *Journal of Experimental Psychology: General, 54*, 377-411.

Stewart, N., Chater, N., & Brown, G. D. A. (2006). Decision by sampling. *Cognitive Psychology, 53*, 1-26.

Stewart, N., & Simpson, K. (in press). A decision-by-sampling account of decision under risk. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind: Prospects for Bayesian cognitive science*. Oxford: Oxford University Press.

Stott, H. (2006). Cumulative prospect theory's functional menagerie. *Journal of Risk and Uncertainty, 32*, 101-130.

Taleb, N. (2007). *The black swan : The impact of the highly improbable*. New York: Random House.

Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *Psychological Review Monograph Supplement 2*.

Trommershäuser, J., Gepshtein, S., Maloney, L. T., Landy, M. S., & Banks, M. S. (2005). Optimal compensation for changes in task-relevant movement variability. *Journal of Neuroscience, 25*, 7169-7178.

Trommershäuser, J., Landy, M. S., & Maloney, L. T. (2006). Humans rapidly estimate expected gain in movement planning. *Psychological Science, 17*, 981-988.

Trommershäuser, J., Maloney, L. T., & Landy, M. S. (2003a). Statistical decision theory and the selection of rapid, goal-directed movements. *Journal of the Optical Society of America A, 20*.

Trommershäuser, J., Maloney, L. T., & Landy, M. S. (2003b). Statistical decision theory and trade-offs in the control of motor response. *Spatial Vision, 16*, 255-275.

Tversky, A., & Fox, C. R. (1995). Weighting risk under uncertainty. *Psychological Review, 102*, 269-283.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5*, 207-232.

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

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review, 101*, 547-567.

Von Neumann, J., & Morgenstern, O. (1947). *Theory of games and economic behavior*. Princeton: Princeton University Press.

Vulkan, N. (2000). An economist's perspective on probability matching. *Journal of Economic Surveys, 14*, 101.

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

Waters, E., Weinstein, N., Colditz, G., & Emmons, K. (2006). Formats for improving risk communication in medical tradeoff decisions. *Journal of Health Communication, 11*, 167-182.

Weber, E. (2006). Experience-based and description-based perceptions of long-term risk: Why global warming does not scare us (yet), *Climatic Change* (Vol. 77, pp. 103-120).

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*, 430-445.

Wedell, D. H., & Bockenholt, U. (1990). Moderation of preference reversals in the long run. *Journal of Experimental Psychology: Human Perception & Performance, 16*, 429-438.

Wu, G., & Gonzalez, R. (1996). Curvature of the probability weighting function. *Management Science, 42*, 1676-1690.

Wu, G., & Gonzalez, R. (1999). Nonlinear decision weights in choice under uncertainty. *Management Science, 45*, 74-85.

Wu, G., Zhang, J., & Gonzalez, R. (2004). Decison under risk. In D. Koehler & N. Harvey (Eds.), *The Blackwell Handbook of Judgment and Decision Making* (pp. 399-423). Oxford: Oxford University Press.

Yechiam, E., Barron, G., & Erev, I. (2005). The role of personal experience in contributing to different patterns of response to rare terrorist attacks. *Journal of Conflict Resolution, 49*, 430-439.

Yechiam, E., & Busemeyer, J. R. (2005). Comparison of basic assumptions embedded in learning models for experience-based decision making. *Psychonomic Bulletin & Review, 12*, 387-402.

Yechiam, E., Erev, I., & Barron, G. (forthcoming). The effect of experience on using a safety device. *Safety Science*.

Zacks, R. T., & Hasher, L. (2002). Frequency processing: A twenty-five year perspective. In P. Sedlmeier & T. Betsch (Eds.), *ETC Frequency processing and cognition* (pp. 21-36). New York, NY: Oxford University Press.

Zwick, R., Neuhoff, V., Marascuilo, L. A., & Levin, J. R. (1982). Statistical tests for correlated proportions: Some extensions. *Psychological Bulletin, 92*, 258-271.