

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

Ilke Aydogan

Erasmus School of Economics

Erasmus University Rotterdam

Yu Gao

Department of Management, Economics and Industrial Engineering

Polytechnic University of Milan

December, 2016

Abstract

This paper investigates the often-discussed over – and under – weighting of rare and extreme events – so called “black swans” – in decisions from experience (DFE). We first resolve the problem of lack of control over experienced probabilities by adjusting the common sampling paradigm of DFE. Our experimental design also controls for utility and uncertainty of experienced probabilities (ambiguity). This enables us to exactly identify the deviations from Expected Utility due to over – or under – weighting of probabilities under risk. Our results confirm the well-known gap between DFE and traditional decisions from description (DFD) but do not provide evidence for underweighting of small probabilities in DFE. We found that experience leads to less pronounced overweighting of small probabilities, and less pronounced underweighting of large probabilities. Thus, our findings suggest a clear de-biasing effect of sampling experience: it attenuates – rather than reverses – the commonly found inverse S-shaped probability weighting in DFD.

Key Words: decisions from experience; decisions under risk; probability weighting; rare outcomes

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

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

Since Barron & Erev (2003) and Hertwig et al. (2004), an intriguing discrepancy between the two decision paradigms, which is called the DFE-DFD gap, has received plenty of attention. The common view in the DFE literature is that people make decisions from experience as if they are underweighting rare and extreme events, so called “black swans”, which are often overweighted under the DFD paradigm (for a review, see Hertwig & Erev, 2009). This pattern implies a reversal of the inverse S-shaped probability weighting that has been documented by many empirical studies under DFD (Abdellaoui 2000; Bleichrodt & Pinto 2000; Bruhin et al. 2010; Booij et al. 2010; Fehr-Duda et al. 2006; Gonzalez & Wu 1999; Tversky & Kahneman 1992; Wu & Gonzalez 1996)¹.

The DFE literature has suggested that the DFE-DFD gap is a robust empirical phenomenon. Although the under-sampling of rare events due to reliance on small samples (sampling error) partly explains the early findings of the gap (Fox & Hadar 2006; Hadar & Fox

¹ See Wakker (2010), section 7.1 for a survey. A comprehensive list of DFD literature supporting inverse S is also provided in the web appendix.

2009, Hertwig et al. 2004), later studies have shown that it does not provide a complete account (Barron & Ursino 2013; Camilleri & Newell 2009; Hau et al. 2008; Hau et al. 2010; Ungemach et al. 2009). Importantly, unlike risk with known probabilities in DFD, the ambiguity in DFE stemming from unknown outcome probabilities – and even from unknown set of possible outcomes – is another cause of the gap (Abdellaoui et al. 2011; Glockner et al. 2016; Kemel & Travers 2016).

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

The second confound concerns the role of utilities. Early studies in the DFE literature argue about the underweighting of rare outcomes in an “as-if” sense. Specifically, the underweighting is typically inferred from a preference for sure gains over expected-value-equivalent lotteries involving unlikely gains (for example, a preference for a sure \$1 over a lottery with 10% chance of winning \$10 and \$0 otherwise). However, the absolute weighting of probabilities stays unclear as the aversion to unlikely gains may as well be due to concave utility (possibly coupled with an unbiased probability weighting) as it may be due to an underweighting of unlikely events. Later studies controlled for utilities by estimating them

together with probability weighting functions using a parametric approach. Nevertheless, one concern about simultaneous parametric estimations is the potential interactions between the parameters of utility and probability weighting functions (Gonzalez & Wu 1999 pp.152; Scheibehenne & Pachur 2015 pp. 403-404; Stott 2006 pp. 112; Zeisberger et al. 2012).

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

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

Deviations from EU due to Probability Weighting

We restrict our attention to probability-contingent binary prospects in the gain domain. A binary prospect of winning α with probability p and β otherwise is denoted $\alpha_p\beta$. Under rank

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

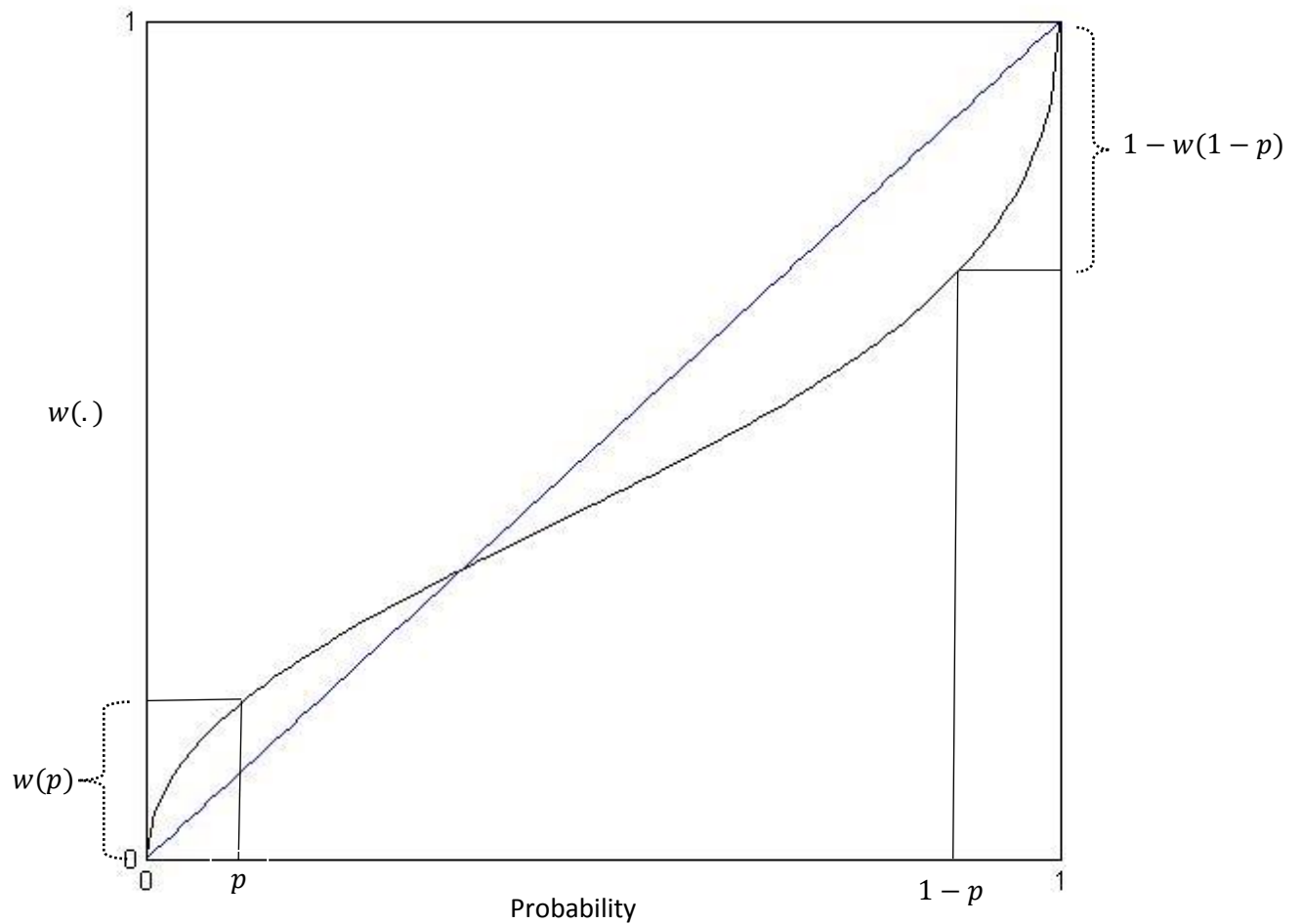
RDU deviates from EU when $w(\cdot)$ is not the identity. Thus, the risk attitude of a decision maker depends not only on the utility curvature as in EU but also on probability weighting. The common finding with the DFD paradigm is an inverse S-shaped (first concave and overweighting, then convex and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

The pattern of inverse S-shaped probability weighting is commonly interpreted as the reflection of both cognitive and motivational deviations from EU (Gonzalez & Wu 1999). On the one hand, the simultaneous overweighting and underweighting of extreme probabilities

² For evidence against inverse S, see Qiu & Steiger (2011), van de Kuilen & Wakker (2011) and Krawczyk (2015). A more complete list of evidence against inverse S in the DFD literature is provided in the web appendix.

implies insufficient sensitivity to intermediate probabilities. This effect is called likelihood insensitivity, and points to cognitive limitations in discriminating different levels of uncertainty. On the other hand, underweighting of moderate probabilities (such as, $w(0.5) < 0.5$) suggests a pessimistic attitude towards risk in the major part of the probability domain. This effect points to motivational deviations from EU.

Figure 1. Inverse S-shaped probability weighting function



The DFE-DFD Gap

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

All three paradigms lead to similar behavioral patterns with an apparent underweighting of rare and extreme outcomes, which contradicts the common empirical findings from DFD. Although the empirical findings with all three paradigms are alike, the two feedback paradigms are inherently different from the sampling paradigm (for an empirical comparison of three DFE paradigms, see Camilleri & Newell 2011b, also see the theoretical discussion of Gonzalez & Dutt 2011). In particular, repeated choices in the two feedback paradigms, as opposed to single decisions in the sampling paradigm, induce long-run payoff considerations due to accumulating income (Wulff et al. 2015). This predicts more expected value maximization in repeated choices by the law of large numbers (Keren & Wagenaar 1987; Lopes 1982; Tversky & Bar Hillel 1983). Furthermore, distinct psychological factors, such as reinforcement learning, and the hot stove effect³, also play a role in repeated decisions with feedback (March 1996; March

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

& Denrell 2001). Erev & Barron (2005) reviews the effects that lead to deviations from expected value maximization in repeated choice paradigms. The sampling paradigm, on the other hand, is more comparable with the DFD paradigm as both involve single decisions. Therefore, the intriguing gap between the sampling paradigm and DFD has received most attention in the DFE literature. The current paper also focuses on the sampling paradigm of DFE.

The Information Asymmetry Account and the Sampling Error

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

Later studies test this information asymmetry account of the DFE-DFD gap by reducing or completely eliminating the sampling error. Several papers demonstrated that the gap is

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

DFE and DFD: Two Different Sources of Uncertainty

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

In a design that is intermediate between DFE and DFD, Abdellaoui et al. (2011b) and Kemel & Travers (2016) finds inverse-S pattern in DFE with more pronounced pessimism than in DFD. This result reflects ambiguity aversion. Kellen & Pachur (2006) and Glockner et al. (2016) find even more pronounced likelihood insensitivity in DFE. These findings are consistent with previous ambiguity literature (Abdellaoui et al. 2011a; Fox & Tversky 1998; Tversky & Fox 1995; Wakker 2004).

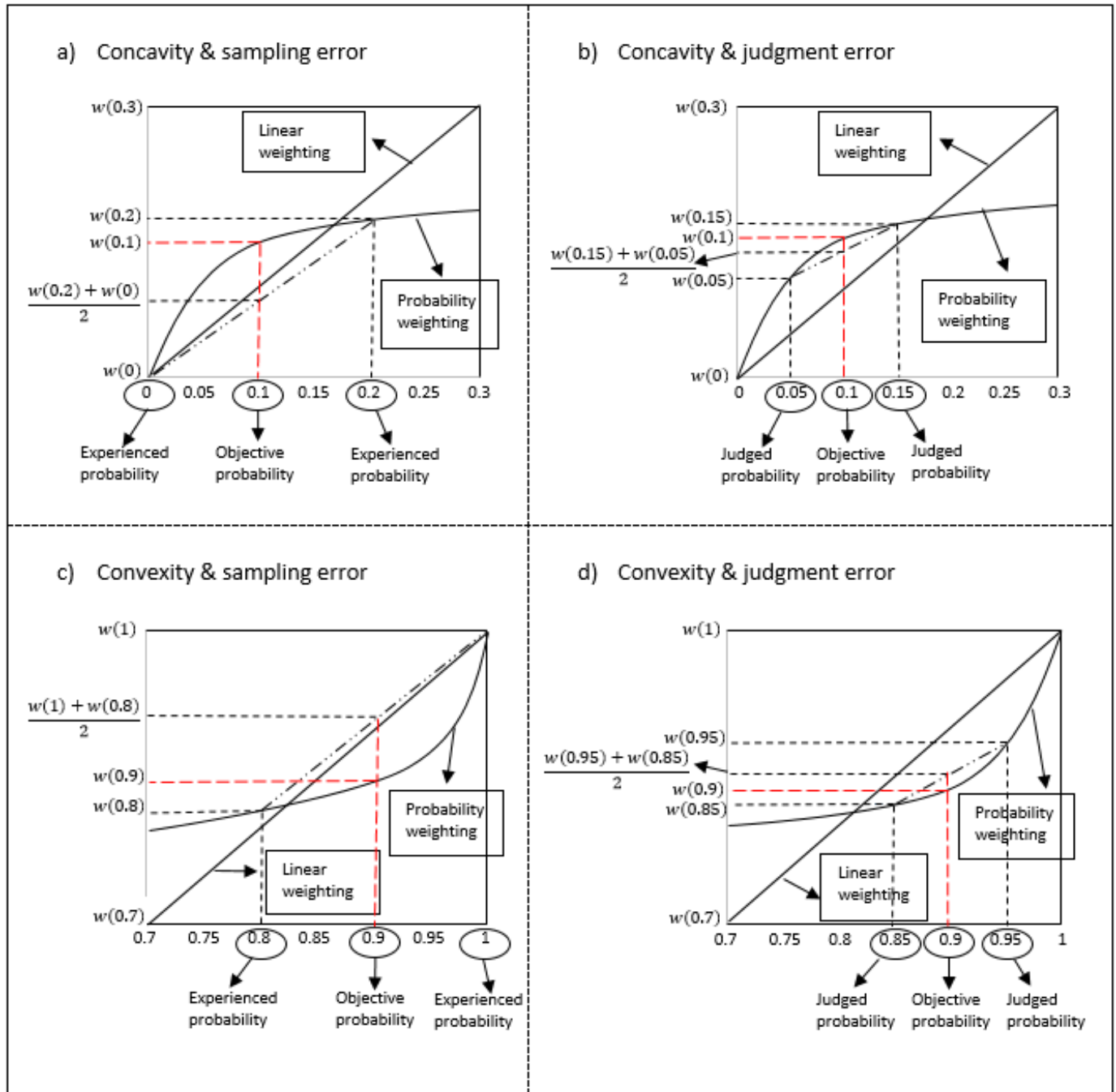
Problem of Aggregation in the Sampling Paradigm

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

To illustrate, assume that all subjects in DFE and DFD have the same probability weighting function depicted in figure 2a, which is concave and overweight 10% probability of a rare and favorable outcome. For the sake of the example, also assume that each subject in DFE draws only 5 times, in which half of the subjects never observe the rare outcome, and the other half observe it once. Therefore, assuming that the subjects do not commit a judgment error, the experienced probabilities will be either 0% or 20%. In this case, aggregating choices over all subjects' amounts to averaging the weightings of 0% and 20% rather than weighting the average 10%. This makes the aggregate choice appear as if 10% is underweighted due to concavity whereas in reality it is overweighted (see figure 2a).

The same effect, although probably smaller in size, also applies when there is no sampling error but only judgment error. Figure 2b illustrates the case where the subjects in DFE accurately observe 10% probability, however, half of them underestimate it as 5% whereas the other half overestimate it as 15%. As a result, the aggregate choice appears as if 10% is weighted less in DFE than in DFD (see figure 2b).

Figure 2. Distortions due to aggregation



By the dual effect, convex probability weighting for large probabilities moves aggregate choices in the direction of overweighting (see figures 2c and 2d). Together with the concavity for small probabilities, this implies a reversed or attenuated inverse S at the aggregate

level, which is what the DFE-DFD gap also suggests. This theoretical conjecture is indeed indirectly supported by the findings of Rakow et al. (2008). In their yoked design, each subject in the DFE treatment is matched with a subject in the DFD treatment who receives the same sampling information in description format. Thus, equating the heterogeneity of the sampling information across the two treatments, they observe that the DFE-DFD gap is almost completely eliminated (also see the discussion of Hau et al. 2010 on the amplification effect in yoked design).

Underweighting or not?

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

Later studies report quantitative estimations of probability weighting under DFE, also by controlling the role of utilities. However, the present evidence on the shape of probability weighting functions is mixed. Hau et al. (2008) and Ungemach et al. (2009) document linear weighting and underweighting respectively, based on the same set of problems used by Hertwig et al. (2004). Among those studies that used larger problem sets, Abdellaoui et al. (2011b),

Kemel & Travers (2016), and Kopsacheilis (2016) report less pronounced overweighting whereas Barron & Ursino (2013) and Frey et al. (2015) report underweighting. Other recent studies by Glockner et al. (2016) and Kellen et al. (2016) reports even more pronounced overweighting under DFE. Differences in methodologies and in manipulations of the sampling paradigm to cope with sampling error are possible sources of the discrepancy. For a further discussion of these discrepant results, see a recent meta-analysis by Wulff et al. (2016).

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

Method

Our experimental procedure consists of two stages. In the first stage, the utility function of each subject is elicited using the trade-off (TO) method of Wakker and Deneffe (1996). The TO method is a well-established method that has been commonly used in studies investigating probability weighting (Abdellaoui 2000; Abdellaoui et al. 2005; Bleichrodt & Pinto 2000; Etchart-Vincent 2004, 2009; Qiu & Steiger 2011). The method entails the elicitation of a standard sequence of outcomes that are equally spaced in utility units. The elicitation procedure consists of a series of adaptive indifference relations. For two fixed gauge outcomes G and g , and a selected starting outcome x_0 with $x_0 > G > g$, $x_1 > x_0$ is elicited such that the subject is indifferent between prospects $x_{1p}g$ and $x_{0p}G$. Then, x_1 is used as an input to elicit $x_2 > x_1$ such that the subject is indifferent between $x_{2p}g$ and $x_{1p}G$. This procedure is repeated n times in order to obtain the standard sequence (x_0, \dots, x_n) with indifferences $x_{i+1p}g \sim x_{ip}G$ for $0 \leq i \leq n - 1$. Under RDU, these indifferences result in $U(x_1) - U(x_0) = U(x_2) - U(x_1) = \dots =$

$U(x_{n-1}) - U(x_n)$ (for the derivation, see Appendix A). A remarkable feature of the TO method is that it elicits these equalities irrespective of what the probability weighting is. Therefore, it is robust against most distortions due to non-expected utility maximization.

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

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

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

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

$$x_k x_j < s_q \Leftrightarrow w(q) < q \text{ (underweighting)} \quad (2)$$

$$x_k x_j \sim s_q \Leftrightarrow w(q) = q \text{ (EU)} \quad (3)$$

$$x_{k_q}x_j > s_q \Leftrightarrow w(q) > q \text{ (overweighting)} \quad (4)$$

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

Barron & Ursino (2013) also investigates the DFE-DFD gap under risk (their experiment 1) similar to our study by using a different two-stage experimental procedure. Their procedure replicates the well-known DFE-DFD gap. However, it does not make inferences about the actual over- or under- weighting of rare outcomes under DFE and DFD⁴.

The Experiment

Subjects and Incentives

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

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

Procedure

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

Stimuli

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

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

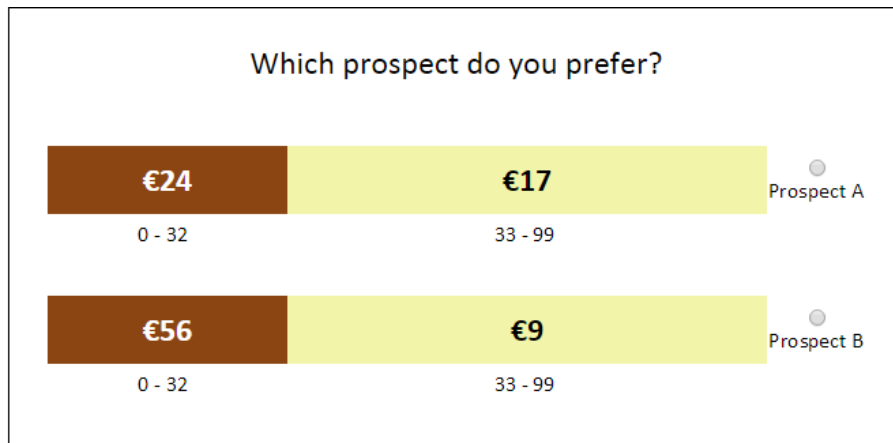
Our choice of the fixed parameters p, G, g, x_0 was fine-tuned based on a pilot session so that the elicitation yields a well-spaced outcome sequence giving reliable certainty equivalent values of s_q in equation 1.

Indifferences were obtained by a bisection method requiring 7 iterations for each x_i . In addition, the last iteration of one randomly chosen x_i was repeated at the end of stage 1, in order to test the reliability of the indifferences. Hence, subjects answered a total of 36 questions

in this part. The bisection iteration procedure is described in Appendix B. The prospects were presented on screen as in Figure 3.

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

Figure 3. Choice situation in the TO part



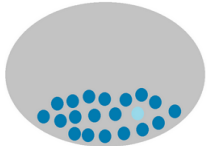
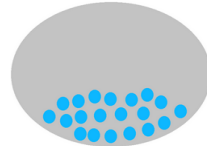
Stage 2: DFD and DFE. Before the start of the second part, each subject was randomly assigned to one of the two treatments: DFE or DFD. Subjects in both treatments answered 7 subject-specific binary choice questions. Each question entailed a choice between a risky prospect $x_{5q}x_1$ and the safe prospect s_q as described in the method section. Note that both x_1 and x_5 were endogenously determined, and varied between subjects.⁵ Values of s_q were

⁵ We used the elicited x_1 as the minimum outcome of the risky prospects to avoid problems related to the extreme behavior of power utility near its origin (Wakker 2008), i.e. x_0 in our design. In particular, for $\alpha < 1$,

always rounded to the nearest integer. The seven probabilities used for the investigation of probability weighting were 0.05, 0.10, 0.20, 0.50, 0.80, 0.90 *and* 0.95. Within each treatment, the orders of the seven questions were counterbalanced. The position of the risky prospect and the safe prospect were also randomized in each question.

Prospects were represented by Ellsberg-type urns containing 20 balls with different monetary values attached to them. This means that all the aforementioned probabilities were fractions of 20; i.e. 5% is 1 out of 20, 10% is 2 out of 20, etc. The two treatments differed from each other in terms of how the contents of the urns were learnt. In the DFD treatment, the contents of the urns were explicitly described to the subject. Figure 4 shows a screen shot of a choice situation for DFD.

Figure 4. Choice situation in DFD

<p>20 balls in total</p>  <p>19 balls, each with value of €96 1 ball with value of €24</p>	<p>20 balls in total</p>  <p>All 20 balls, each with value of €92</p>
<p>Please choose the urn you prefer to draw one ball from:</p> <div style="display: flex; justify-content: space-around; margin-top: 10px;"> <div style="text-align: center;"> <p>Left</p> <input type="radio"/> </div> <div style="text-align: center;"> <p>Right</p> <input type="radio"/> </div> </div>	

Subjects in the DFE treatment were initially given no information about the contents of the urns except the total number of balls. They could only learn about the outcome

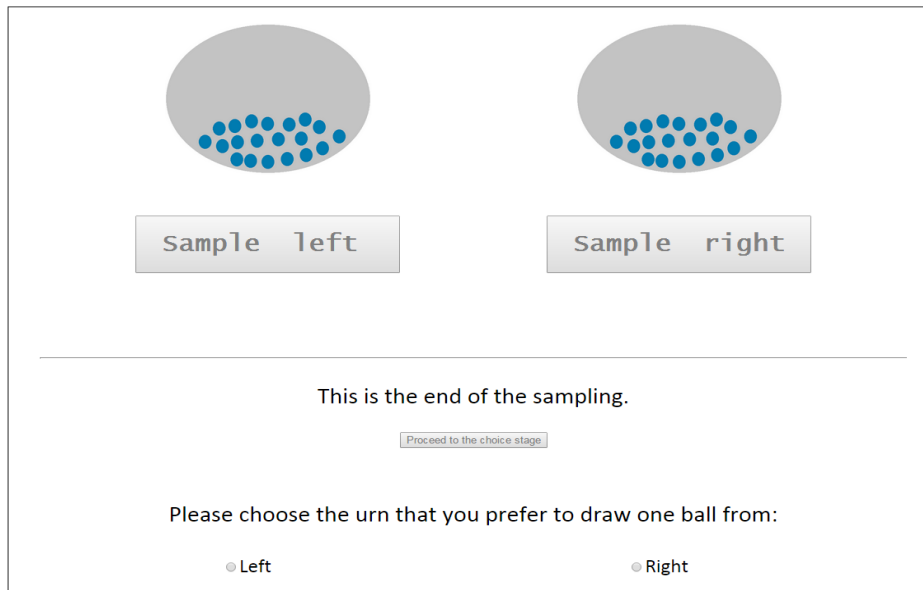
the slope of the power utility converges to infinity as x tends to the origin. This implies extreme risk aversion near the origin. Similarly, $\alpha > 1$ implies extreme risk seeking near the origin.

compositions of the urns by sampling each and every ball one-by-one without replacement, and observing the monetary values attached. Figure 5 shows a screen shot of the sampling phase in the DFE treatment. Subjects sampled balls from urns by clicking “Sample left” or “Sample right” on the screen. Each time, the monetary outcome attached to the ball sampled was shown to the subject for 1.5 seconds, and then disappeared. Subject could sample in their own speed, in whichever order they preferred, and switch as many times as they wanted, but they could only proceed to the choice stage after sampling all the balls in both urns.

Figure 5. Sampling stage in DFE



Figure 6 shows the screen shot of the choice stage in DFE. In case a question in this part was drawn for the payment at the end of the experiment, the experimenters physically created the relevant urn seen on the screen by filling an opaque urn with 20 ping-pong balls painted to dark blue or light blue, each associated with the payoffs in question (see Figure 4). Then, the subject drew a ball from the urn, which determined her payoffs.

Figure 6. Choice stage in DFE

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

Results

Reliability and Consistency of Utility Elicitation

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

repeated indifference values did not differ from original elicitations ($p=0.44$, Wilcoxon sign-rank).

In our data, one subject reached the possible lower bound of x_i 's in all 5 cases. Consequently, her standard sequence was not well spaced enough for the estimations of s_q with Equation (1).⁶ We excluded this subject from the following analysis. The analysis with this subject included does not alter our conclusions. The same problem was not observed with any other subject.

Utility Functions

Table 1 gives the descriptive statistics for the elicited outcome sequence. The parameter α of the power utility $u(x) = x^\alpha$ was estimated at the individual level by ordinary least squares regression. The average R^2 over all individual utility estimations was 0.985 which indicated that our estimations fit the data well.

Table 1. Descriptive statistics of the elicited outcome sequence (N=88)

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

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

α	1.05	0.36	0.41	0.99	2.65
----------	------	------	------	------	------

The summary statistics for the mean and median α are reported in the last row of Table 1. The aggregate data did not deviate from linearity ($p=0.92$, Wilcoxon sign-rank). Although the mean alpha suggested slight convexity, this was due to the outliers in our data. Three subjects exhibited extreme convexity with $\alpha > 2$, and the Skewness/Kurtosis test rejected the normality of the distribution of α 's ($p=0.00$). Utilities did not differ across the two treatments ($p=0.84$, Wilcoxon rank-sum).

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

Probability Weighting: DFE vs. DFD

Aggregate data. In this section, we report the aggregate choices in the direction of overweighting and underweighting according to (2) and (4) in the Method section. The proportions of overweighting and underweighting of small and large probabilities are given in Figures 7 and 8 respectively.

The aggregate choices replicated the common DFE-DFD gap at the extreme probabilities. The gap was significant at 0.95 ($p=0.02$, χ^2); and marginally significant at 0.10, and 0.90 ($p=0.06$, and $p=0.07$ respectively, χ^2). It was always in the expected direction, alleviating the overweighting of small probabilities and the underweighting of large

probabilities under DFE. The gap at probability 0.05 was not significant ($p=0.20$, χ^2), although the trend suggested reduced overweighting in DFE. There was also no apparent DFE-DFD gap in the middle range, $0.20 \leq q \leq 0.80$ ($p=0.35$, $p=0.92$, and $p=0.37$ for $q = 0.20, 0.50$, and 0.80 respectively, χ^2).

In what follows, we focus on absolute overweighting and underweighting of probabilities under DFD and DFE. We first test the deviations from unbiased weighting in either directions with the classical two-sided binomial tests for proportions. In addition, to interpret the relative evidence for overweighting and underweighting, we report Bayes factors for the null hypothesis of overweighting against the alternative hypothesis of underweighting. Bayes factors state the relative evidence for the null hypothesis. For instance, a Bayes factor of 10 indicates that overweighting is 10 times more likely than underweighting for the given probability. Following Jeffreys (1961), a Bayes factor between 3 and 10 is interpreted as “some evidence”, a Bayes factor between 10 and 30 is interpreted as “strong evidence”, and a Bayes factor larger than 30 is interpreted as “very strong evidence” for the null of overweighting. Similarly, Bayes factors between 0.1 and 0.33, between 0.1 and 0.03, and less than 0.03 are interpreted respectively as “some evidence”, “strong evidence”, and “very strong evidence” for the alternative hypothesis of underweighting.⁷

As shown in Figure 7, for small probabilities, we found a marginally significant deviation from unbiased weighting at 0.05 ($p=0.07$) under DFD. Interpreting from Bayes factors, there was strong evidence of overweighting 0.05 ($BF=28.04$), some evidence of overweighting 0.1 ($BF=8.54$) and some evidence of underweighting 0.2 ($BF=0.02$). Under DFE, we only found a significantly biased weighting at 0.2 ($p=0.03$). Interpreting from Bayes

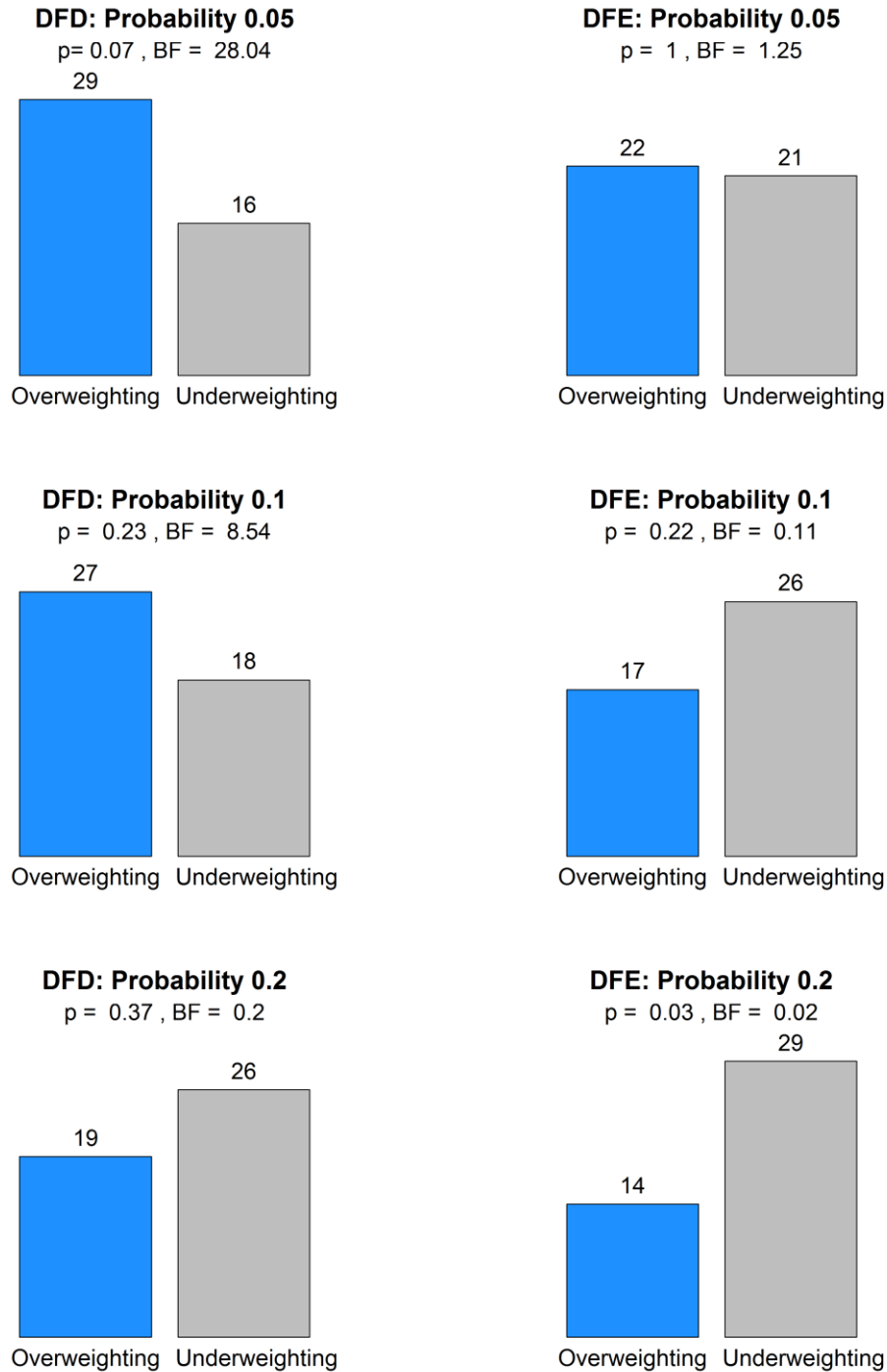
⁷ Bayes factors are computed with the package BayesFactor in R (Morey, Rouder, Jamil & R Core Team, 2015)

factors, there was strong evidence of underweighting 0.2 ($BF=0.02$) and some evidence of underweighting 0.1 ($BF=0.11$). There was no evidence for the underweighting or the overweighting of 0.05 ($BF=0.11$ and $BF=1.25$ respectively).

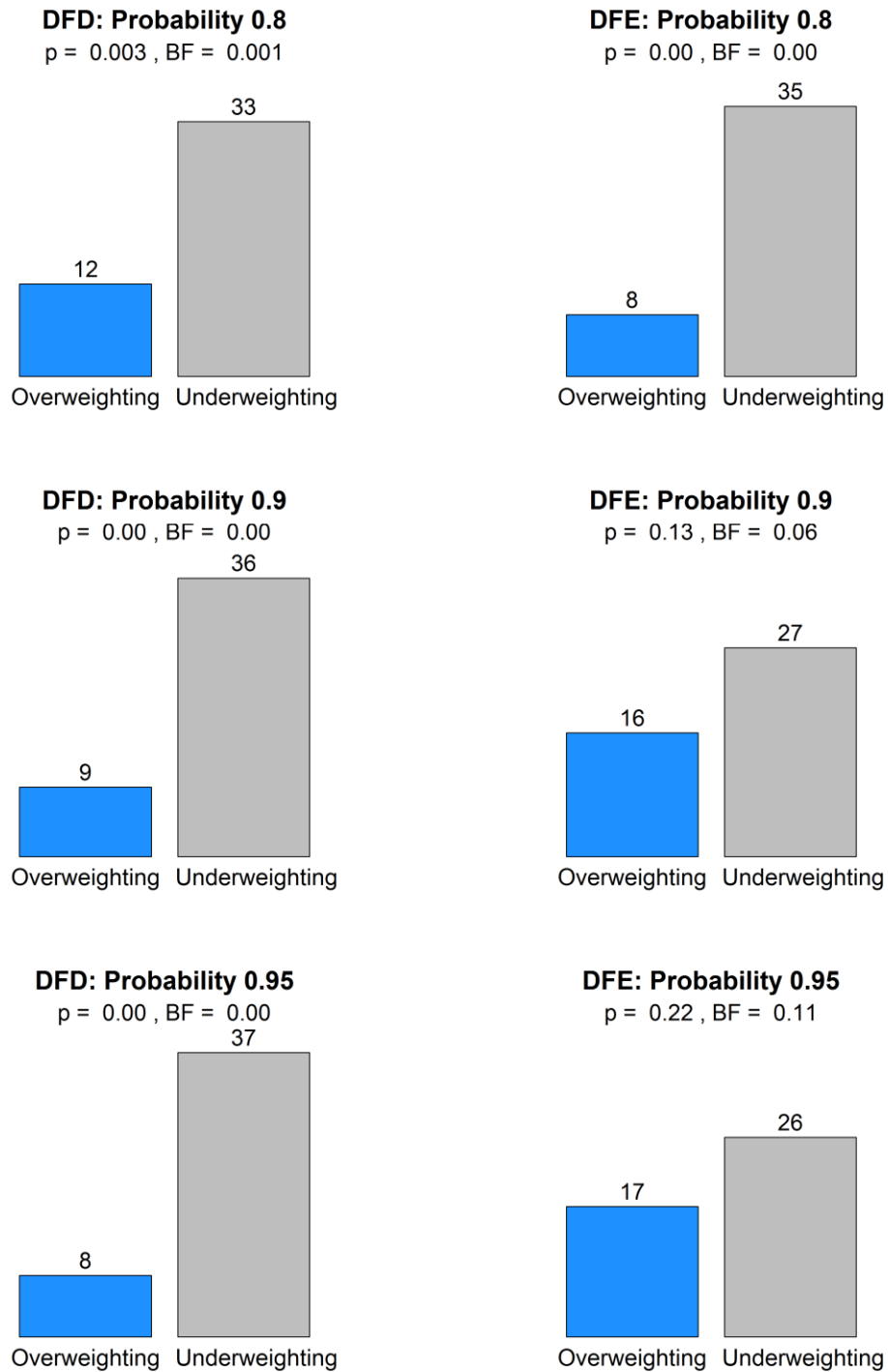
For large probabilities as shown in Figure 8, under DFD, we found significant biases in weighting of probabilities 0.8, 0.9 and 0.95 significant ($p=0.00$ for all). The Bayes factors indicated very strong evidence for underweighting of 0.8, 0.9 and 0.95 ($BF=0.00$ for all). Under DFE, we found significant bias only at 0.8 ($p=0.00$). The Bayes factors suggested very strong evidence of underweighting of 0.8 ($BF=0.00$), strong evidence of underweighting 0.9 ($BF=0.06$) and some evidence of underweighting 0.95 ($BF=0.11$).

Lastly, we examined the weighting of the moderate 0.5 probability. 38 out of 45 subjects in the DFD treatment and 36 out of 43 subjects in the DFE treatment underweighted 0.5. Hence, the deviations from unbiased weighting was highly significant at 0.5 in both treatments ($p=0.00$ for both treatments, two-sided binomial tests). The Bayes factors also indicated very strong evidence in favor of underweighting at 0.5 ($BF<0.03$ for both treatments).

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

Figure 7. Weighting of Small Probabilities

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

Figure 8. Weighting of Large Probabilities

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

Individual data. Next, we examine the shape of probability weighting functions at the individual level. We classify each subject's probability weighting function as inverse S-shaped, S-shaped, pessimistic or optimistic based on the number of over – and under – weightings at three small and three large probabilities examined in Figures 7 and 8. Specifically, a probability weighting function is inverse S-shaped if it overweights at least two out of three small probabilities and underweights at least two out of three large probabilities, at the same time. An S-shaped probability weighting function is implied by the opposite pattern. Similarly, a pessimistic probability weighting function underweights at least two small and two large probabilities at the same time, and the opposite pattern implies an optimistic probability weighting function.

The classification results are in Table 2. The probability weighting functions were mainly classified as inverse S-shaped, S-shaped or pessimistic while the proportion of optimistic weighting functions was negligible in both treatments. Among the three main types, the majority of the probability weighting functions was inverse S-shaped in the DFD treatment ($p=0.00$, one-sided binomial, H_0 : Proportion of inverse S is $\frac{1}{3}$ among inverse S, S and pessimistic types). The inverse S-shape was also the most frequent type in the DFE treatment but it was not the majority ($p=0.13$, one-sided binomial, H_0 : Proportion of inverse S is $\frac{1}{3}$ among inverse S, S and pessimistic types).

The comparison across the two treatments indicated that the proportion of S-shaped probability weighting functions was higher in the DFE treatment, although the difference was only marginally significant ($p=0.08$, two-sided Fisher's exact test). There was no significant difference in the proportions of inverse S-shaped, pessimistic and optimistic probability weighting functions across the two treatments.

Table 2. Type of Probability Weighting Functions

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

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

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

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

We estimated Goldstein & Einhorn's (1987) weighting function given by $w(q) = \frac{\beta q^\alpha}{\beta q^\alpha + (1-q)^\alpha}$. The parameter α determines the curvature and captures the sensitivity towards changes in probabilities. Here, $\alpha < 1$ indicates inverse S-shape and likelihood insensitivity, and $\alpha > 1$ indicates S-shape and likelihood sensitivity. The parameter β determines the

elevation, and captures the degree of pessimism. For $\beta = 1$, we have $w(0.5) = 0.5$. Lower (higher) values of β indicates less (more) elevation and more (less) pessimism. Following Kruschke (2011), we evaluate the credibility of likelihood insensitivity and pessimism based on the ranges of 95% intervals from posterior distribution of parameters. The details on estimation procedures are in Appendix C.

We report the estimated group level mean parameters and corresponding 95% credibility intervals in Table 3. Figure 9 shows the estimated probability weighting functions. The estimated parameters indicated credible likelihood insensitivity and pessimism in both treatments as $\alpha = 1$ and $\beta = 1$ fell on the right side of 95% credibility intervals. The DFE-DFD gap in terms of likelihood insensitivity and pessimism was not credible, although the difference in likelihood insensitivity was suggestive. Hence, we observed a less pronounced inverse S-shape in the DFE weighting function, while the elevation was comparable across the two treatments (black curves in Figure 9).

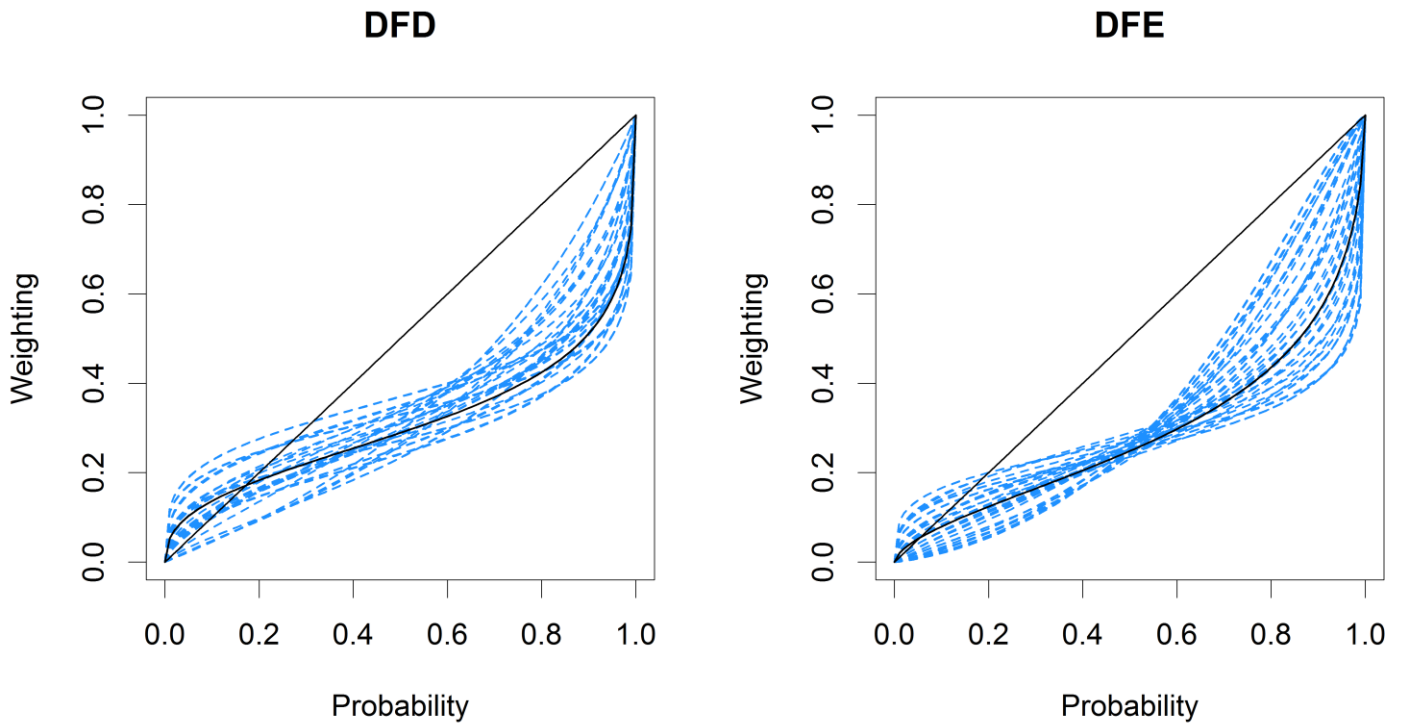
Table 3. Group level mean parameters

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

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

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

Figure 9. Probability weighting functions



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

Discussion

The Impact of Sampling Experience

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

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

As regards the second factor, the DFE-DFD gap can signify other internal biases due to memory limitations and/or inattention (Camilleri & Newell 2011a). To avoid these potential confounds in our experiment, we provided our subjects with paper and pen and reminded them that they can keep track of the outcomes during the sampling stage in DFE. We observed that

more than half of the subjects in the DFE treatment took notes. Hence, our results were less likely to be driven by misremembering the past observations.

Our conclusions on the impact of sampling experience on probability weighting are consistent with some previous findings. Gottlieb et al. (2007), Hilbig & Glöckner (2011), and Humphrey (2006) also report reduced probability weighting with different variants of the sampling paradigm. Erev et al. (2015, problems 1,2,7 – 11), Jessup et al. (2008), van de Kuilen & Wakker (2006), and van de Kuilen (2009) report significant convergence to EU maximization under risk in repeated choice settings, when immediate feedback after each choice is available but not when it is unavailable. These results also suggest the distinct impact of experience in repeated choice settings (also see Lejarraga & Gonzalez 2011 on strong impact of experience).

Two-Stage Design, Non-parametric and Parametric Analysis

Our two stage experimental design avoided potential interdependencies between utility and probability weighting components of RDU, which was reported by previous studies. Moreover, it enabled a reliable non-parametric analysis of probability weighting functions without relying on specific functional forms, which can be subject to distortions. Our parametric estimations were consistent with our nonparametric analysis. To further test the descriptive adequacy of the parametric Bayesian estimations, we compared posterior predictions of the estimated model with the actual data observed (see Appendix C, Figure C2). The model was accurate in predicting choices.

Despite the aforementioned advantages of the nonparametric approach, one might still have concerns about our two-stage design. One concern is the error propagation in the chained procedure. In particular, the stimulus for the measurement of probability weighting in the

second stage is determined based on the utilities elicited in the first stage. Thus, any error in calculation of s_q from the first stage may result in a bias in probability weighting measurements. However, studies investigating this point have shown that this problem is indeed negligible (Abdellaoui et al. 2005; Bleichrodt and Pinto 2000). Moreover, high goodness of fit in estimations of utility functions, and the replication of common qualitative patterns of probability weighting under DFD confirm the validity of our procedure.

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

Lastly, our experimental design makes an implicit assumption that the sampling experience has an impact on the probability domain but not on utilities. This assumption was supported by some previous studies (Abdellaoui et al. 2011b; Glockner et al. 2016; Kopsacheilis 2016; but also see Kellen et al. 2016), and enabled us to measure utilities under the more efficient DFD paradigm in the first stage.

Conclusion

This paper clarifies the controversy about the DFE-DFD gap. Our strictly controlled sampling paradigm isolates the impact of the sampling experience from other confounds, and

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

Appendix A: Derivation of the Standard Sequence of Outcomes in TO Method

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

Appendix B: Bisection Procedure

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

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

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

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

The fourth term on the left hand side (+16) ensured the outcome monotonicity (see point 2).

One concern for the TO method and the bisection iteration process is the incentive compatibility due to the adaptive design. A subject who is fully aware of the adaptive design can strategically drive the value x_i upwards by pretending to be extremely risk averse in the bisection questions. In this way, he or she can increase the expected values of prospects in the subsequent questions for the elicitation of x_{i+1} . To make it more difficult for our subjects to

fully grasp the process, we included two filler questions in the iteration process of each x_i . The two filler choices were after the first and the third choice questions for every x_i . In these questions, x_i was changed in the direction that is opposite to the changes described in point 2 above. These questions had no further impact on the flow of the procedure.

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

Appendix C: Bayesian Hierarchical Estimation Procedure

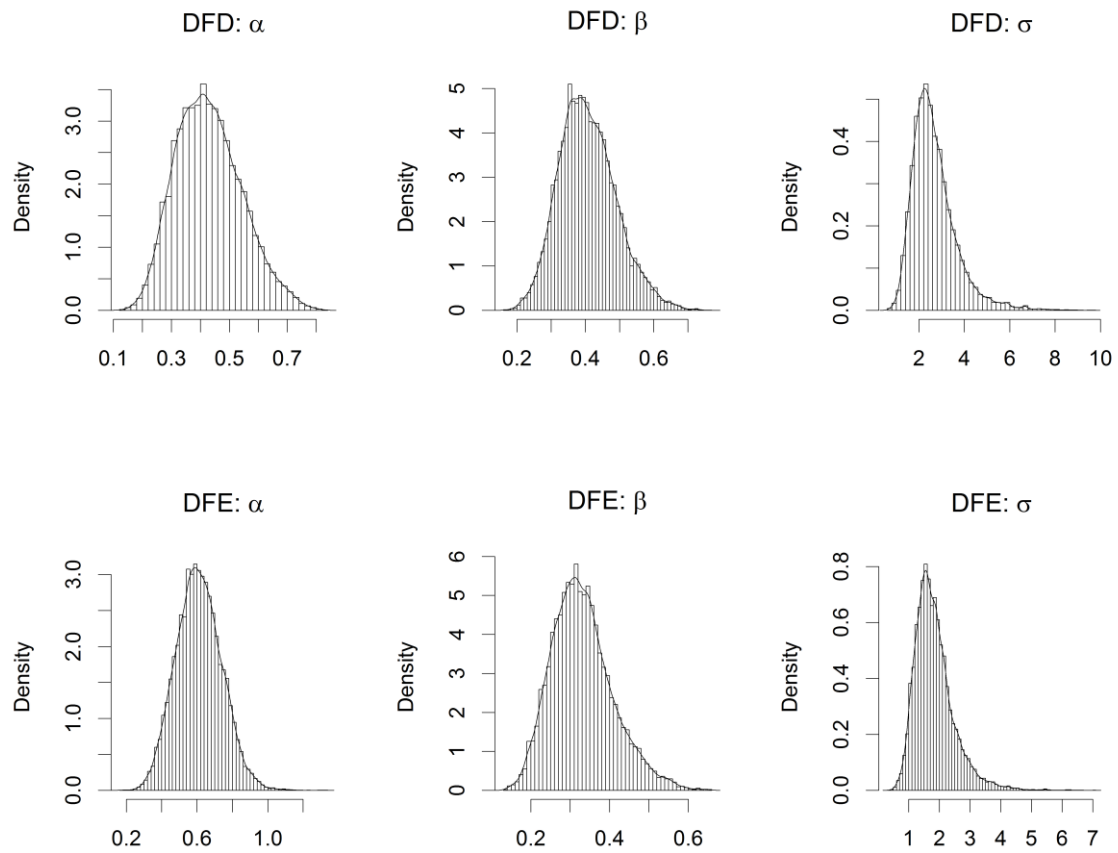
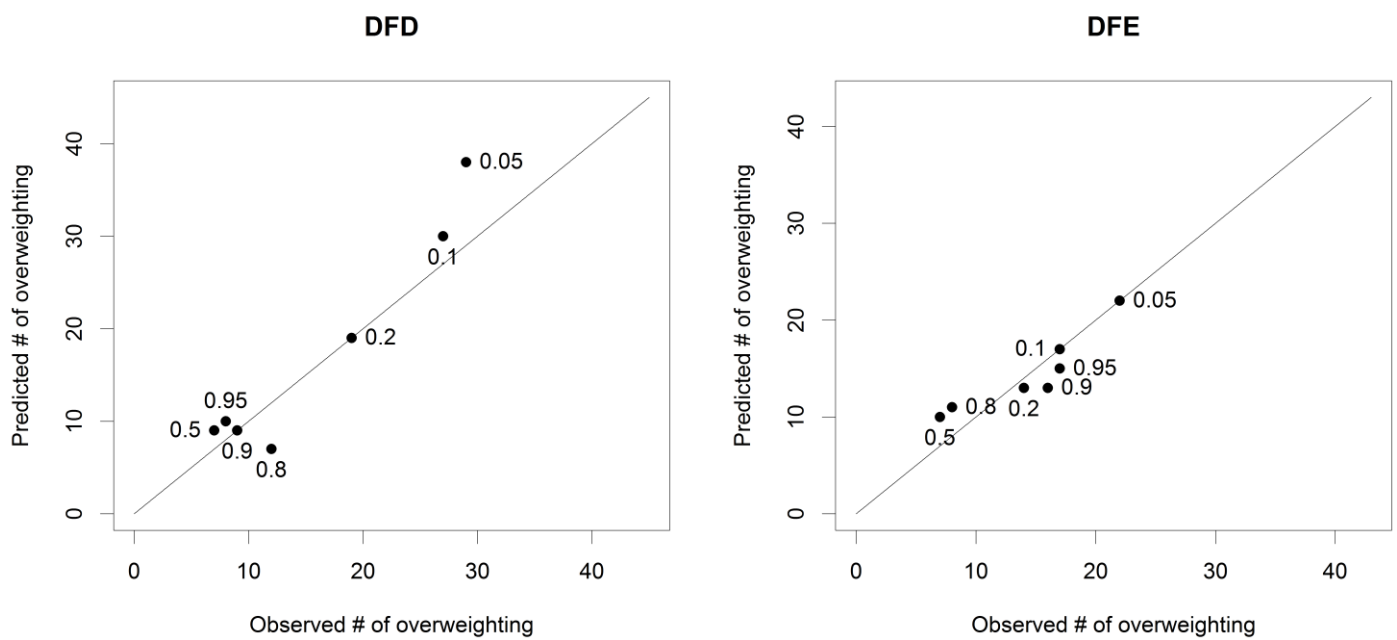
We implemented Bayesian hierarchical estimation procedure as follows. The Goldstein & Einhorn's (1987) probability weighting function is $w(q) = \frac{\beta q^\alpha}{\beta q^\alpha + (1-q)^\alpha}$. The probability of choosing the risky prospect was calculated using Luce's (1959) stochastic choice function,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

In the estimations, prior distributions of individual level parameters α_i and β_i were set to uniform distributions. The ranges of the uniform distributions were from 0.1 to 2 for α_i and from 0.1 to 1.5 for β_i . The range chosen for α_i allows a wide array of curvatures ranging from strong inverse S-shape to strong S-shape. The range chosen for β_i implies that $w(0.5)$ is between $\frac{1}{11}$ and $\frac{3}{5}$, which is considered as a reasonable range given the previous findings in the literature and our nonparametric results suggesting strong underweighting at 0.5. Group level parameters were linked to individual level parameters through probit transformations and linear linkages (see Nilsson et al. 2011; Scheibehenne & Pachur 2015). Group level means were assumed to follow standard normal distribution so that the individual level parameters had uniform distributions. Group level standard deviations were uniformly distributed ranging from 0 to 10. The individual level noise parameters σ_i were assumed to come from a lognormal distribution. The log-normal group mean for σ was assumed to be uniformly distributed ranging from 0.1 to 10. The lognormal standard deviation for σ was uniformly distributed

ranging from 0 to 1.33. The upper bound of 1.33 was the standard deviation of the transformed uniform distribution $U(\ln(0.1), \ln(10))$, following Nilsson et al. (2011, pg. 88).

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

Figure C1 shows the posterior histograms for the group level mean parameters. Figure C2 shows the predictive performance of the estimations by comparing the median numbers of overweighting predicted by the posterior distributions of group level parameters with the actual numbers of overweighting observed in our data. The model predictions match with the observed data for 0.2 and 0.9 in the DFD treatment, and for 0.05 and 0.1 in the DFE treatment. The predictions for the other probabilities were close to the actual data in the DFE treatment. The predictions for 0.05 and 0.8 in the DFD treatment indicated some misalignment with the actual data, although they performed well in the rest of the probabilities.

Figure C1. Posterior histograms for group level means**Figure C2. Posterior predictions based on group level parameters**

References

- Abdellaoui, M. (2000). Parameter-Free Elicitation of Utility and Probability Weighting Functions. *Management Science*, 46(11), 1497–1512. <http://doi.org/10.1287/mnsc.46.11.1497.12080>
- Abdellaoui, M., Baillon, A., Placido, L., & Wakker, P. P. (2011a). The rich domain of uncertainty: Source functions and their experimental implementation. *The American Economic Review*, 101(2), 695-723.
- Abdellaoui, M., L'Haridon, O., & Paraschiv, C. (2011b). Experienced vs. Described Uncertainty: Do We Need Two Prospect Theory Specifications? *Management Science*, 57(10), 1879–1895. <http://doi.org/10.1287/mnsc.1110.1368>
- Abdellaoui, M., Vossman, F., & Weber, M. (2005). Choice-Based Elicitation and Decomposition of Decision Weights for Gains and Losses under Uncertainty. *Management Science*, 51(9), 1384–1399. <http://doi.org/10.1287/mnsc.1050.0388>
- Balcombe, K. & Fraser, I. (2015), Parametric Preference Functionals under Risk in the Gain Domain: A Bayesian Analysis. *Journal of Risk and Uncertainty*, 50 (2) , 161 – 187. <http://doi:10.1007/s11166-015-9213-8>
- Bardsley, N., Cubitt, R., Loomes, G., Moffatt, P., Starmer, C., & Sugden, R. (2010). *Experimental economics: Rethinking the rules*. Princeton University Press.
- Barron, G., & Erev, I. (2003). Small feedback-based decisions and their limited correspondence to description-based decisions. *Journal of Behavioral Decision Making*, 16(3), 215–233. <http://doi.org/10.1002/bdm.443>
- Barron, G., & Ursino, G. (2013). Underweighting rare events in experience based decisions: Beyond sample error. *Journal of Economic Psychology*, 39, 278–286. <http://doi.org/10.1016/j.joep.2013.09.002>

- Barron, G., & Yechiam, E. (2009). The coexistence of overestimation and underweighting of rare events and the contingent recency effect. *Judgment and Decision Making*, 4(6), 447.
- Bleichrodt, H., Cillo, A., & Diecidue, E. (2010). A Quantitative Measurement of Regret Theory. *Management Science*, 56(1), 161–175. <http://doi.org/10.1287/mnsc.1090.1097>
- Bleichrodt, H., & Pinto, J. L. (2000). A Parameter-Free Elicitation of the Probability Weighting Function in Medical Decision Analysis. *Management Science*, 46(11), 1485–1496. <http://doi.org/10.1287/mnsc.46.11.1485.12086>
- Booij, A. S., Praag, B. M. S. van, & Kuilen, G. van de. (2010). A parametric analysis of prospect theory's functionals for the general population. *Theory and Decision*, 68(1-2), 115–148. <http://doi.org/10.1007/s11238-009-9144-4>
- Bruhin, A., Fehr-Duda, H., & Epper, T. (2010). Risk and Rationality: Uncovering Heterogeneity in Probability Distortion. *Econometrica*, 78(4), 1375–1412. <http://doi.org/10.3982/ECTA7139>
- Camilleri, A. R., & Newell, B. R. (2009). The role of representation in experience-based choice. *Judgment and Decision Making*, 4(7), 518.
- Camilleri, A. R., & Newell, B. R. (2011a). Description-and experience-based choice: Does equivalent information equal equivalent choice? *Acta Psychologica*, 136(3), 276–284.
- Camilleri, A. R., & Newell, B. R. (2011b). When and why rare events are underweighted: A direct comparison of the sampling, partial feedback, full feedback and description choice paradigms. *Psychonomic Bulletin & Review*, 18(2), 377–384. <http://doi.org/10.3758/s13423-010-0040-2>
- Denrell, J., & March, J. (2001). Adaptation as Information Restriction: The Hot Stove Effect. *Organization Science*, 12(5), 523–538. <http://doi.org/10.1287/orsc.12.5.523.10092>

- Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *The quarterly journal of economics*, 643-669.
- Erev, I., & Barron, G. (2005). On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological Review*, 112(4), 912.
- Erev, I., Plonsky, O., & Ert, E. (2015). From Anomalies to Forecasts: A Choice Prediction Competition for Decisions under Risk and Ambiguity. *Mimeo*.
- Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, 53(2), 134.
- Estes, W. K. (2002). Traps in the route to models of memory and decision. *Psychonomic Bulletin & Review*, 9(1), 3–25. <http://doi.org/10.3758/BF03196254>
- Etchart-Vincent, N. (2004). Is Probability Weighting Sensitive to the Magnitude of Consequences? An Experimental Investigation on Losses. *Journal of Risk and Uncertainty*, 28(3), 217–235. <http://doi.org/10.1023/B:RISK.0000026096.48985.a3>
- Etchart-Vincent, N. (2009). Probability weighting and the “level” and “spacing” of outcomes: An experimental study over losses. *Journal of Risk and Uncertainty*, 39(1), 45–63. <http://doi.org/10.1007/s11166-009-9066-0>
- Fehr-Duda, H., Gennaro, M. de, & Schubert, R. (2006). Gender, Financial Risk, and Probability Weights. *Theory and Decision*, 60(2-3), 283–313. <http://doi.org/10.1007/s11238-005-4590-0>
- Fox, C. R., & Tversky, A. (1998). A belief-based account of decision under uncertainty. *Management science*, 44(7), 879-895.
- Fox, C. R., & Hadar, L. (2006). “Decisions from experience”= sampling error+ prospect theory: Reconsidering Hertwig, Barron, Weber & Erev (2004). *Judgment and Decision Making*, 1(2), 159.

- Glöckner, A., Fiedler, S., Hochman, G., Ayal, S., & Hilbig, B. (2012). Processing differences between descriptions and experience: A comparative analysis using eye-tracking and physiological measures. *Frontiers in psychology*, 3, 173.
- Glöckner, A., Hilbig, B. E., Henninger, F., & Fiedler, S. (2016). The reversed description-experience gap: Disentangling sources of presentation format effects in risky choice. *Journal of Experimental Psychology: General*, 145(4), 486.
- Goldstein, W. M., & Einhorn, H. J. (1987). Expression theory and the preference reversal phenomena. *Psychological Review*, 94(2), 236.
- Gonzalez, C., & Dutt, V. (2011). Instance-based learning: Integrating sampling and repeated decisions from experience. *Psychological Review*, 118(4), 523.
- Gonzalez, R., & Wu, G. (1999). On the Shape of the Probability Weighting Function. *Cognitive Psychology*, 38(1), 129–166. <http://doi.org/10.1006/cogp.1998.0710>
- Gottlieb, D. A., Weiss, T., & Chapman, G. B. (2007). The Format in Which Uncertainty Information Is Presented Affects Decision Biases. *Psychological Science*, 18(3), 240–246. <http://doi.org/10.1111/j.1467-9280.2007.01883.x>
- Gul, F. (1991). A theory of disappointment aversion. *Econometrica: Journal of the Econometric Society*, 667–686.
- Hadar, L., & Fox, C. R. (2009). Information asymmetry in decision from description versus decision from experience. *Judgment and Decision Making*, 4(4), 317.
- Hau, R., Pleskac, T. J., & Hertwig, R. (2010). Decisions from experience and statistical probabilities: Why they trigger different choices than a priori probabilities. *Journal of Behavioral Decision Making*, 23(1), 48–68. <http://doi.org/10.1002/bdm.665>

- Hau, R., Pleskac, T. J., Kiefer, J., & Hertwig, R. (2008). The description–experience gap in risky choice: the role of sample size and experienced probabilities. *Journal of Behavioral Decision Making*, 21(5), 493–518. <http://doi.org/10.1002/bdm.598>
- Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from Experience and the Effect of Rare Events in Risky Choice. *Psychological Science*, 15(8), 534–539. <http://doi.org/10.1111/j.0956-7976.2004.00715.x>
- Hertwig, R., & Erev, I. (2009). The description–experience gap in risky choice. *Trends in Cognitive Sciences*, 13(12), 517–523. <http://doi.org/10.1016/j.tics.2009.09.004>
- Hilbig, B. E., & Glöckner, A. (2011). Yes, they can! Appropriate weighting of small probabilities as a function of information acquisition. *Acta Psychologica*, 138(3), 390–396. <http://doi.org/10.1016/j.actpsy.2011.09.005>
- Hogarth, R. M., & Soyer, E. (2011). Sequentially simulated outcomes: Kind experience versus nontransparent description. *Journal of Experimental Psychology: General*, 140(3), 434.
- Hogarth, R. M., Lejarraga, T., & Soyer, E. (2015). The two settings of kind and wicked learning environments. *Current Directions in Psychological Science*, 24(5), 379–385.
- Humphrey, S. J. (2006). Does Learning Diminish Violations of Independence, Coalescing and Monotonicity? *Theory and Decision*, 61(2), 93–128. <http://doi.org/10.1007/s11238-006-8047-x>
- Jeffreys, H (1961). *Theory of Probability* (3rd ed.). New York: Oxford University Press.
- Jessup, R. K., Bishara, A. J., & Busemeyer, J. R. (2008). Feedback Produces Divergence From Prospect Theory in Descriptive Choice. *Psychological Science*, 19(10), 1015–1022. <http://doi.org/10.1111/j.1467-9280.2008.02193.x>
- Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, 263–291.

- Kellen, D., Pachur, T., & Hertwig, R. (in press). How (in)variant are subjective representations of described and experienced risk and rewards? *Cognition*.
- Kemel, E., & Travers, M. (2015). Comparing attitudes toward time and toward money in experience-based decisions. *Theory and Decision*, 80(1), 71–100.
<http://doi.org/10.1007/s11238-015-9490-3>
- Keren, G., & Wagenaar, W. A. (1987). Violation of utility theory in unique and repeated gambles. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13(3), 387.
- Kopsacheilis, O. (2016, June). Information about risky decisions: decomposing and reinterpreting the Description-Experience gap. *Presented at the FUR conference, University of Warwick; 29, June, 2016; Coventry, UK*
- Krawczyk, M. W. (2015). Probability weighting in different domains: The role of affect, fungibility, and stakes. *Journal of Economic Psychology*. Retrieved from
<http://www.sciencedirect.com/science/article/pii/S0167487015000835>
- Kruschke, J. K. (2011). *Doing Bayesian Data Analysis*. New York: Academic Press
- Lejarraga, T., Pachur, T., Frey, R., & Hertwig, R. (2016). Decisions from Experience: From Monetary to Medical Gambles. *Journal of Behavioral Decision Making*, 29(1), 67–77.
<http://doi.org/10.1002/bdm.1877>
- Lejarraga, T., & Gonzalez, C. (2011). Effects of feedback and complexity on repeated decisions from description. *Organizational Behavior and Human Decision Processes*, 116(2), 286-295.
- Lopes, L. L. (1982). Doing the impossible: A note on induction and the experience of randomness. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 8(6), 626.
- Luce, R. Duncan. (1959). *Individual Choice Behavior*. New York: John Wiley & Sons.

- March, J. G. (1996). Learning to be risk averse. *Psychological Review*, 103(2), 309.
- Nilsson, H., Rieskamp, J., & Wagenmakers, E. J. (2011). Hierarchical Bayesian parameter estimation for cumulative prospect theory. *Journal of Mathematical Psychology*, 55(1), 84-93.
- Morey, R. D., Rouder, J. N., Jamil, T., & R Core Team, (2015). BayesFactor. Vienna, Austria: R Foundation for Statistical Computing.
- Qiu, J., & Steiger, E.-M. (2010). Understanding the Two Components of Risk Attitudes: An Experimental Analysis. *Management Science*, 57(1), 193–199.
<http://doi.org/10.1287/mnsc.1100.1260>
- Rakow, T., Demes, K. A., & Newell, B. R. (2008). Biased samples not mode of presentation: Re-examining the apparent underweighting of rare events in experience-based choice. *Organizational Behavior and Human Decision Processes*, 106(2), 168–179.
<http://doi.org/10.1016/j.obhdp.2008.02.001>
- Rakow, T., & Newell, B. R. (2010). Degrees of uncertainty: An overview and framework for future research on experience-based choice. *Journal of Behavioral Decision Making*, 23(1), 1–14. <http://doi.org/10.1002/bdm.681>
- Scheibehenne, B., & Pachur, T. (2015). Using Bayesian hierarchical parameter estimation to assess the generalizability of cognitive models of choice. *Psychonomic bulletin & review*, 22(2), 391-407.
- Schunk, D., & Betsch, C. (2006). Explaining heterogeneity in utility functions by individual differences in decision modes. *Journal of Economic Psychology*, 27(3), 386–401.
<http://doi.org/10.1016/j.joep.2005.08.003>
- Sidman, M. (1952). A note on functional relations obtained from group data. *Psychological Bulletin*, 49(3), 263.

- Stott, H. P. (2006). Cumulative prospect theory's functional menagerie. *Journal of Risk and Uncertainty*, 32(2), 101–130. <http://doi.org/10.1007/s11166-006-8289-6>
- Trautmann, S. T., Van De Kuilen, G., Keren, G., & Wu, G. (2015). Ambiguity attitudes. *The Wiley Blackwell Handbook of Judgment and Decision Making*, 1, 89-116.
- Tversky, A., & Bar-Hillel, M. (1983). Risk: The long and the short. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 9(4), 713–717. <http://doi.org/10.1037/0278-7393.9.4.713>
- Tversky, A., & Fox, C. R. (1995). Weighing risk and uncertainty. *Psychological review*, 102(2), 269.
- Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4), 297–323.
<http://doi.org/10.1007/BF00122574>
- Tversky, A., Slovic, P., & Kahneman, D. (1990). The causes of preference reversal. *The American Economic Review*, 204–217.
- Ungemach, C., Chater, N., & Stewart, N. (2009). Are Probabilities Overweighted or Underweighted When Rare Outcomes Are Experienced (Rarely)? *Psychological Science*, 20(4), 473–479. <http://doi.org/10.1111/j.1467-9280.2009.02319.x>
- van de Kuilen, G. (2007). Subjective Probability Weighting and the Discovered Preference Hypothesis. *Theory and Decision*, 67(1), 1–22. <http://doi.org/10.1007/s11238-007-9080-0>
- van de Kuilen, G., & Wakker, P. P. (2006). Learning in the Allais paradox. *Journal of Risk and Uncertainty*, 33(3), 155–164. <http://doi.org/10.1007/s11166-006-0390-3>
- van de Kuilen, G., & Wakker, P. P. (2011). The Midweight Method to Measure Attitudes Toward Risk and Ambiguity. *Management Science*, 57(3), 582–598.
<http://doi.org/10.1287/mnsc.1100.1282>

- Wakker, P., & Deneffe, D. (1996). Eliciting von Neumann-Morgenstern Utilities When Probabilities Are Distorted or Unknown. *Management Science*, 42(8), 1131–1150. <http://doi.org/10.1287/mnsc.42.8.1131>
- Wakker, P., Erev, I., & Weber, E. U. (1994). Comonotonic independence: The critical test between classical and rank-dependent utility theories. *Journal of Risk and Uncertainty*, 9(3), 195–230. <http://doi.org/10.1007/BF01064200>
- Wakker, P. P. (2004). On the composition of risk preference and belief. *Psychological review*, 111(1), 236.
- Wakker, P. P. (2008). Explaining the characteristics of the power (CRRA) utility family. *Health Economics*, 17(12), 1329–1344. <http://doi.org/10.1002/hec.1331>
- Wakker, P. P. (2010). *Prospect Theory: For Risk and Ambiguity*. Cambridge University Press.
- Weber, E. U., Shafir, S., & Blais, A.-R. (2004). Predicting Risk Sensitivity in Humans and Lower Animals: Risk as Variance or Coefficient of Variation. *Psychological Review*, 111(2), 430–445. <http://doi.org/10.1037/0033-295X.111.2.430>
- Wu, G., & Gonzalez, R. (1996). Curvature of the Probability Weighting Function. *Management Science*, 42(12), 1676–1690. <http://doi.org/10.1287/mnsc.42.12.1676>
- Wulff, D. U., Hills, T. T., & Hertwig, R. (2015). How short-and long-run aspirations impact search and choice in decisions from experience. *Cognition*, 144, 29-37.
- Wulff, D. U., Canseco, M. M., & Hertwig, R. (2016). A Meta-Analytic Review of Two Modes of Learning and the Description-Experience Gap. *Under Submission*
- Yechiam, E., & Busemeyer, J. R. (2006). The effect of foregone payoffs on underweighting small probability events. *Journal of Behavioral Decision Making*, 19(1), 1–16. <http://doi.org/10.1002/bdm.509>

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