Running Head: ARE PROBABILITIES OVERWEIGHTED OR UNDERWEIGHTED

Are probabilities overweighted or underweighted, when rare outcomes are experienced (rarely)?

Christoph Ungemach

University of Warwick, England

Nick Chater

University College London, England

Neil Stewart

University of Warwick, England

c.ungemach@warwick.ac.uk

+44 (0) 24 7652 8386

Department of Psychology

University of Warwick

Coventry

CV4 7AL

England

Ungemach, C., Chater, N., & Stewart, N. (2009). Are probabilities overweighted or underweighted, when rare outcomes are experienced (rarely)? *Psychological Science, 20,* 473-479.

Abstract

People often make decisions with risky outcomes. When presented with verbal descriptions of outcomes and their associated probabilities, people behave as if they overweight small probabilities. When the same outcomes are instead experienced in a series of samples, people behave as if they underweight small probabilities. We present two experiments showing that the reversal from overweighting to underweighting occurs even if the existing explanations are controlled for: (a) by using representative samples of events, underweighting cannot be attributed to undersampling of the small probabilities; (b) because earlier samples predicted decisions just as well as later samples, underweighting cannot be attributed to recency weighting; (c) and because frequency judgments were accurate, underweighting cannot be attributed to judgment error. Furthermore, we show that this reversal is also reflected in the best-fitting parameter values when applying Prospect Theory, the dominant model of risky choice.

Are probabilities overweighted or underweighted, when rare outcomes are experienced (rarely)?

In the last several decades of decision making research in experimental psychology, the predominant experimental paradigm has been to present summary symbolic descriptions of simple lotteries. For example, a gamble might be described as "an 80% chance of winning £4 otherwise a 20% chance of winning £0." It is widely accepted that choice data from descriptive problems imply that people overweight low probabilities and underweight high probabilities. In Prospect Theory (PT) (Kahneman & Tversky, 1979), the dominant model of descriptive choice, objective probabilities and amounts are transformed into their subjective counterparts and then combined multiplicatively. The transform of the values is an S-shaped function and the transform of probability into the subjective decision weight is an inverse S-shaped function, overweighting small probabilities and underweighting probabilities near 1. The shape of this weighting function has also been tested directly (e.g., Bleichrodt, 2001; Gonzalez & Wu, 1999; Tversky & Kahneman, 1992; Wu & Gonzalez, 1996).

Recently, researchers have revisited choice problems in which objective information regarding the outcomes and probabilities is not initially known, but instead must be inferred from a sample of possible outcomes (e.g., Barkan, Zohar, & Erev, 1998; Barron & Erev, 2003; Hertwig, Barron, Weber & Erev, 2004; Weber, Shafir, & Blais, 2004). Most of these studies (e.g., Barron & Erev, 2003; Hertwig et al., 2004; Weber, Shafir, & Blais, 2004) report choice behavior that differs sharply from that observed in descriptive choice. Crucially, choices appear to indicate an *underweighting* of small probabilities.

Hertwig et al. (2004) argue that reliance on small samples and overemphasis of outcomes from recent samples may explain the apparent underweighting of rare events in decision from experience. Their "decisions from experience" paradigm is based on an initial sampling phase in

which participants freely explore a pair of lotteries without cost by drawing samples without replacement from their 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). After this sampling phase, participants decide which lottery to play once for real.

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 generally 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 the rare "0 points" outcome fewer than twice, and will thus underestimate the 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. 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. Fox and Hadar (2006) go further and argue that what seems to be evidence for underweighting of probability in decisions from experience is not really evidence for underweighting, and, instead, is entirely consistent with overweighting of probabilities, as also occurs in decisions from description. That is, people are experiencing a smaller probability than intended because of the skew from small samples, but then overweight the smaller experienced probability (as in Prospect Theory) resulting in a net effect of slight underweighting. According to this explanation, apparent underweighting is not a psychological phenomenon at all, but results from the statistical properties of small samples.

Fox and Hadar (2006) also raise the possibility of a distortion through judgment error,

which might arise in the mapping from experienced frequencies to probabilities. This could occur at the first stage of the two-stage model of decision under uncertainty (Tversky & Fox, 1995; Fox & Tversky, 1998), in which probabilities of uncertain events are subjectively assessed before being further transformed by the weighting function. To evaluate this possibility, Fox and Hadar (2006) added an explicit probability judgment task to the design, finding that probability judgments were well calibrated, and thus cannot underlie the apparent underweighting of rare events. Moreover, using PT, they successfully applied value- and weighting-function parameters reported by Tversky and Kahneman (1992), originally fitted to descriptive problems, to individual probability judgments, and found a good fit with the observed choices.

The two experiments reported here test the impact of sampling error directly, by eliminating it. The second experiment also further examines the potential impact of judgment error.

Experiment 1. The Matched Sampling Design

If the apparent underweighting of small probabilities is explained by the undersampling of rare outcomes, a possibility raised, in different ways, by both Hertwig et al. (2004) and Fox and Hadar (2006), then it should be eliminated and reversed to match overweighting in decision from description if participants experience perfectly representative samples. This prediction is tested in Experiment 1.

Method. 75 students at the University of Warwick received £2 for the completion of six choice tasks in the laboratory. The six decision problems were taken from Hertwig et al. (2004) and are summarized in Table 1. Participants were randomly assigned to one of three experimental conditions, a Free-Sampling Condition, a Matched-Sampling Condition, and a Description Condition. The first two conditions involved 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. 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 determined randomly for each participant. In the crucial Matched-Sampling Condition, however, participants had to sample 40 outcomes from each option, in any order, before proceeding to the decision phase; and the frequencies of outcomes precisely matched the underlying probabilities, with the order of outcomes decided randomly for each participant. Thus, a decision problem with an underlying probability of .8 of 4 points, and .2 of nothing, would be realized as exactly 32 trials with 4 points, and 8 trials with 0 points, in random order. The Description Condition involved presenting summaries of the same lotteries in the format:

80% chance to win 4 points;

20% chance to win 0 points.

Participants were instructed to attempt to obtain as many points as possible across the six choice problems. At the end of the experiment, the lotteries were played randomly for each participant and they were then informed of their points total.

Results and Discussion. With a median number of 19 draws per choice problem (both buttons together) in the Free-Sampling Condition, participants showed an information search pattern typical for decisions from experience formats. Across all lotteries, the rare event was encountered less frequently than expected in 50% of the sequences, showing less undersampling than reported by Hertwig et al. (2004). Nonetheless, in one third of all cases the rare event was not encountered at all, leaving participants completely ignorant regarding the existence of the

rare event—the extreme case of undersampling.

Table 1 shows the raw choice proportions in the direction of overweighting of small probabilities, for the different formats within the 6 individual choice problems, together with the *z*-values for the tests on the differences between the proportions for the Description Condition and the Sampling Conditions. Not only are the choice proportions in the direction of overweighting lower in the Sampling Conditions; the proportions also differ across the 50% line, which implies an actual change in modal choice. This is the case for problems 1, 3, 4 and 6 in the Free-Sampling Condition and for problems 3 and 6 in the Matched-Sampling Condition. Following Hertwig et al. (2004), the differences in the raw proportions and the change in modal choice can be interpreted as actual reversals of choice in the direction of over- rather than under-weighting of small probabilities.

Averaging over the six choice problems, there was a significant effect of task format on the mean choice proportions in the direction of overweighting, $F(2, 72) = 18.6$, $p < .0001$, $p_{rep} > .999$, $\omega = .65$. Planned contrasts revealed that sampling from the options instead of receiving a description led to significantly smaller proportion of choices in the direction of overweighting of small probabilities, independent of sampling method, $t(72) = -5.59$, $p < .0001$, $p_{rep} > .999$, two-sided, $r = .55$ and that the decrease was stronger in the Free-Sampling Condition than in the Matched-Sampling Condition, $t(72) = -2.45$, $p = .017$, $p_{rep} = .933$, two-sided, $r = .28$. This means that in the Matched-Sampling Condition, where sample size is controlled and sample frequencies precisely match the underlying probabilities, the effect is reduced but crucially is not eliminated. Thus, sampling error seems to explain only part of the difference between decision by experience and decision by description.

One possible explanation for why the effect is maintained, even with matched, equal samples is that people's "mental samples" are smaller than the actual samples. If so, we should expect most recently sampled outcomes to be overrepresented because they are more available in memory (e.g., Atkinson and Shiffrin, 1968), and hence more strongly determine the final choices. But such recency effects are not observed. Following Hertwig et al. (2004), using only the expected value of the second half of the sample to predict people's choices is no more reliable than using the first half, neither in the Free-Sampling Condition (69% vs. 65%, $t(24) = 0.54$, $p = .596$, $p_{rep} = .43$, two-sided) nor in the Matched-Sampling Condition (48% vs. 42%, $t(24) = 1.12$, $p = .272$, $p_{rep} = .67$, two-sided).

To investigate the extent to which established models can account for these results we calculated the rate of correct predictions based on PT. Fox and Hadar (2006) found that 63% of choices conformed to PT when applying the median value- and weighting-function parameters from Tversky and Kahneman (1992) to probabilities judged by the participants. Instead of applying a specific set of parameters implying overweighting of small probabilities, we tested the performance of PT across a range of values between 0 and 2 in steps of .01 which also covers potential shapes for the weighting function that imply actual underweighting of small probabilities. The choices from different subjects were pooled and parameters estimated across all choices, separately for decision problems involving either only gains (parameters α and γ) or losses (parameters β and δ). Figure 1 shows the proportion of correct predictions as a function of the value- and weighting function parameters.

[Insert Figure 1]

In the Free-Sampling Condition, the highest rate of correct predictions obtained under PT

was 81%. The contour plots of the top row of Figure 1 show that the regions with the highest fit, represented by the darkest colors, lie predominantly within the top half of the plots, which are marked by weighting function parameters > 1. Functions based on such parameter values imply underweighting of small probabilities. Conversely, the brightest areas with the lowest fit are mostly found in the lower half of the plots. In the Matched-Sampling Condition the maximum rate of correct predictions for PT was much lower, at 64%. The contour plots for this data (second row of Figure 1) show that this degree of fit is obtained across a wide range of weighting function parameters including values implying both over- and underweighting of small probabilities.

In summary, this experiment demonstrates the robustness of the underweighting of small probabilities in decisions from experience, even when there is no sampling error and no recency effect. Furthermore, underweighting appears also to be reflected in the best fitting parameter values for the PT model. However, this experiment does not rule out the possibility that people systematically misjudge probabilities from their sample.

Experiment 2. Matched Sampling with the Assessment of Judgment Error

In a second experiment, we repeated the first study, adding a probability judgment task to test directly for any systematic judgment biases.

Method. This Web-based experiment was completed by a total of 197 participants, including both students and the general population. The design was similar to the Matched-Sampling Condition in Experiment 1, with 40 outcomes per button matching the underlying probabilities and an additional judgment task after the decision phase. Participants estimated the number of times they had seen the rare event, for both of the options from which they had sampled. To prevent participants from using a counting strategy in subsequent problems, each

participant received only one of the six choice problems used previously.

Results and Discussion. The differences between choice proportions within the six individual choice problems are provided in Table 2. In 5 out of 6 decision problems the proportions again differ across the 50% line, indicating reversals in modal choice in the direction of underweighting of small probabilities under decisions from experience. Although sampling error was eliminated, significant differences were again observed when comparing the mean choice proportions in the direction of overweighting across all six problems between the Matched-Sampling Condition and the Description Condition of the first experiment, $t(220) = 4.22$, $p < .0001$, $p_{rep} > .999$, two-sided.

If judgment error is responsible for the reversed choice behavior, one would expect systematic underestimation of low frequencies, in contrast to the general finding of overestimation of low frequencies reported in the frequency judgment literature (e.g., Zacks & Hasher, 2002). The judgments observed here were well-calibrated with a high correlation between judged and actual frequencies, $r\ (370) = .98$, $p < .0001$, $p_{rep} > .999$ and a mean absolute difference of 1.57 ($SD = 2.37$). Figure 2 shows the distribution of observed deviations between actual and estimated frequencies.

[Insert Figure 2]

An examination of the mean estimation errors showed small deviations in the direction of overestimation of low frequencies and underestimation of high frequencies (from low to high frequency: $t(34) = 1.46$, $p = 0.154$, $p_{rep} = .76$; $t(61) = 5.16$, $p < 0.001$, $p_{rep} > .999$; $t(90) = 3.18$, $p = 0.002$, $p_{rep} = .98$; $t(64) = 2.71$, $p = 0.009$, $p_{rep} = .95$; $t(118) = 2.89$, $p = 0.005$, $p_{rep} = .97$, all two-sided).

Recency weighting in the form of a superior rate of correct predictions based on outcomes from the second half of the sampled sequences was again not found (51% and 47% respectively, χ^2 McNemar (1) = 0.006, $p = 0.938$, $p_{rep} = .14$).

Using the frequency judgments, it was possible to calculate model fits for both PT and the two-stage model (Tversky & Fox, 1995; Fox & Tversky, 1998). As can be seen from Figure 3, the highest rate of correct predictions for both models was obtained with weighting function parameters > 1 (upper half of the contour plots), giving both models a weighting function with a shape that incorporates underweighting of small probabilities. The highest prediction rate for the PT model with parameters estimated on the basis of the experienced probabilities was 62%. For the two-stage model a similar maximum fit of 65% was obtained using the judged probabilities.

[Insert Figure 3]

General Discussion

The matched sampling design used here provides additional support for the robustness of the apparent underweighting of small probabilities in decisions from experience: underweighting remains even when the influence of statistical sampling error, due to small sample sizes, is eliminated. Sampling error can therefore not be the sole explanation for the phenomenon. Both experiments also provide evidence for the occurrence of underweighting in the absence of recency weighting. Moreover, in the light of well-adjusted frequency estimations (with a slight tendency to overweight small frequencies), judgment error in the form of underestimation of small probabilities from frequency data can also be excluded as an explanation. Finally, the best fits of PT to this data involve probability weighting function parameters that also imply underweighting of small probabilities, with shapes inverse to those established for decisions from description. The results of the two-stage model suggest that the same results are obtained

independent of sampling error.

With all the candidate explanations disqualified, it remains unclear what exactly causes the reversal. In order to gain a deeper understanding of the cognitive processes involved in the phenomenon, it might be important to extend the range of the models considered. One alternative class of models that seems to be able to describe behavior in experiential choice tasks quite well stems from the reinforcement learning literature (e.g., March, 1996). Other approaches could be derived from findings within probability learning tasks, (e.g., Goodnow, 1955; Nicks, 1959; Restle, 1961), which show that people make decisions on the basis of smaller units of a sequence such as runs or recurring patterns. There is also evidence suggesting that lotteries are evaluated relative to one another and that the preferences can be dependent on the set of options available (e.g., Stewart, Chater, Stott & Reimers, 2003; Wedell, 1991). This could also apply to the evaluation of options in decisions from experience. An examination of the sampling process shows that majority of participants frequently switch between options (mean number of switches = 10.77, *SD* = 13.84), which could facilitate the evaluation of the options relative to each other throughout the exploration. Depending on the switching pattern, the resulting sub-samples that might be compared will have outcome sequences that no longer represent the objective probabilities of the original options. This perspective suggests a different view of the nature of the problem and extends the range of plausible strategies that have to be considered in order to explain the observed behavior.

References

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation* (Vol. 8). London: Academic Press.

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.

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.

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

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.

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

Goodnow, J. J. (1955). Determinants of Choice-Distribution in Two-Choice Situations. *The American Journal of Psychology, 68*, 106-116.

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.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*, 263-291.

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

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

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

Stewart, N., Chater, N., Stott, H. P., & Reimers, S. (2003). Prospect relativity: How choice options influence decision under risk. *Journal of Experimental Psychology: General, 132,* 23-46.

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

Tversky, A., & Fox, C. R. (1995). Weighting risk and uncertainty. *Psychological Review, 102,* 269-283.

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.

Wu, G., & Gonzalez, R. (1996). Curvature of the Probability Weighting Function. *Management Science, 42*, 1676-1690.

Wedell, D. H. (1991). Distinguishing among models of contextually induced preference reversals. *Journal of Experimental Psychology-Learning Memory and Cognition, 17*, 767-778.

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.

Author Note

Christoph Ungemach, Department of Psychology, University of Warwick; Nick Chater, Department of Psychology and Centre for Economic Learning and Social Evolution, University College London; Neil Stewart, Department of Psychology, University of Warwick.

This research was funded in part by ESRC grant RES-062-23-0952. We thank the members of the London Judgment and Decision Making Group for the insightful discussion. Nick Chater is supported by a Leverhulme Trust Major Research Fellowship.

Correspondence concerning this article should be addressed to Christoph Ungemach, Department of Psychology, University of Warwick, Coventry, CV4 7AL, UK. E-mail: c.ungemach@warwick.ac.uk.

Table 1

Summary of the observed choice proportions within the individual choice problems for the three experimental conditions in Experiment 1 including the differences between the Description Condition and the two Sampling Conditions (Free- and Matched-Sampling).

			Proportions of choices in the direction of overweighting of small probabilities (in %)				
Choice problem	Option A	Option B	Description (n = 25)	Free - Sampling (n=25)	Difference between groups	Matched - Sampling (n=25)	Difference between groups
1	4, .8; **0, .2,**	3, 1.0	64	36	+38 (z = 1.98, p =. 024)	52	+12 (z = .86, p =.195)
2	**4, .2;** 0, .8	3, .25; 0, .75	72	56	+16 (z = 1.18 , p = .119)	60	+16 (z = .9, p =.185)
3	-3, 1.0	**-32, 0.1;** 0, .9	64	16	+48 (z = 3.46, p < .001)	28	+36 (z = 2.55, p = .005)
4	-3, 1.0	-4, 0.8; **0, .2,**	64	32	+32 (z = 2.26, p = .012)	68	-4 (z = .3, p =.383)
5	**32,.1;** 0, .9	3, 1.0	48	8	+40 (z = 3.15 , p < .001)	16	+32 (z = 2.43, p =.008)
6	**32, .025;** 0, .975	3, .25; 0, .75	52	28	+24 (z = 1.73 p = .042)	28	+24 (z = 1.73 , p =.042)

The rare event within each option has been highlighted in bold. The column on the right side presents the z-statistic testing whether the difference between the sample proportions is significantly different from zero.

Table 2

Summary of the observed choice proportions within the individual choice problems for the Matched-Sampling Condition in Experiment 2, the Description Condition in Experiment 1 and the differences between them.

			Proportions of choices in the direction of overweighting of small probabilities (in %)		
Choice problem	Option A	Option B	Description (n = 25)	Matched - Sampling	Difference between groups
1	4, .8; **0, .2,**	3, 1.0	64	32 (10/31)	+32 (z = 2.39, p =.008)
2	**4, .2;** 0, .8	3, .25; 0, .75	72	39 (12/31)	+33 (z = 2.46, p =.007)
3	-3, 1.0	**-32, 0.1;** 0, .9	64	42 (13/31)	+22 (z = 1.64, p = 051)
4	-3, 1.0	-4, 0.8; **0, .2,**	64	45 (17/38)	+19 (z = 1.48, p =.07)
5	**32,.1;** 0, .9	3, 1.0	48	45 (14/31)	+3 (z = .22 , p =.411)
6	**32, .025;** 0, .975	3, .25; 0, .75	52	26 (9/35)	+26 (z = 2.06, p =.02)

The rare event within each option has been highlighted in bold. The column on the right side presents the z-statistic testing whether the difference between the sample proportions is significantly different from zero.

Figure Captions

Figure 1. Contour plots for the two Sampling Conditions in Experiment 1 with the percentage of correct predictions calculated for each combination of value- and weighting function parameters between 0 and 2 in steps of .01, separately for the problems involving gains (left) and losses (right). The regions with the darkest color indicate the combinations providing the highest fit.

Figure 2. Deviations of the frequency judgments for the rare events across the actually experienced frequencies. For the sure options participants could only provide estimates for the common event (frequency of 40). The dotted line indicates perfect calibration. The white dots show the spread of the deviations for the different frequencies. The mean estimation errors are indicated by the black dots.

Figure 3. Contour plots for the Matched-Sampling Condition in Experiment 2 with the percentage of correct predictions calculated for each combination of value- and weighting function parameters between 0 and 2 in steps of .01, separately for the problems involving gains (left) and losses (right). The regions with the darkest color indicate the combinations providing the highest fit.

Figure 1

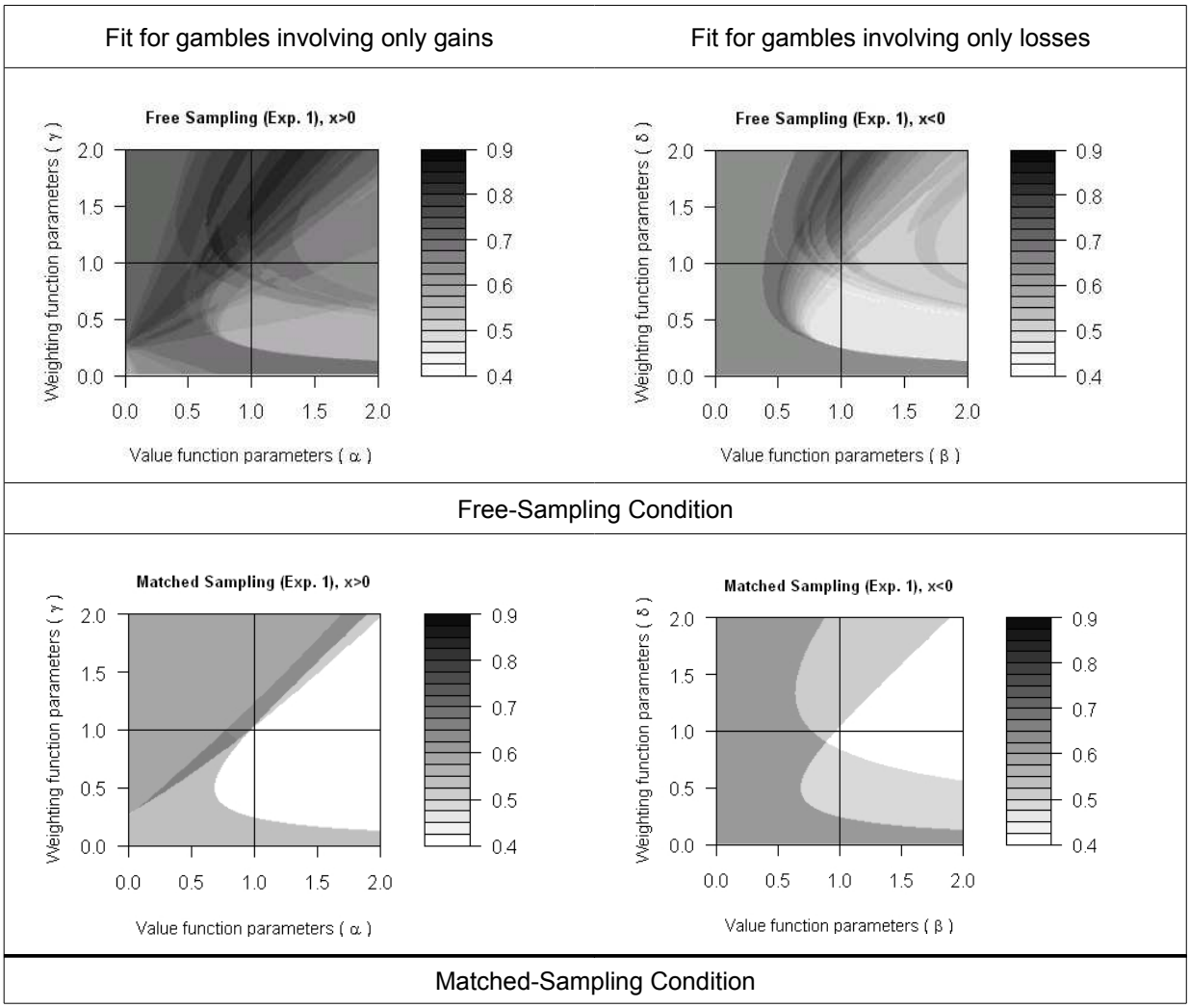

Fit for gambles involving only gains
Fit for gambles involving only losses
Free Sampling (Exp. 1), x>0
Free Sampling (Exp. 1), x<0
Weighting function parameters (γ)
Value function parameters (α)
Value function parameters (β)
Free-Sampling Condition
Matched Sampling (Exp. 1), x>0
Matched Sampling (Exp. 1), x<0
Matched-Sampling Condition

Figure 2

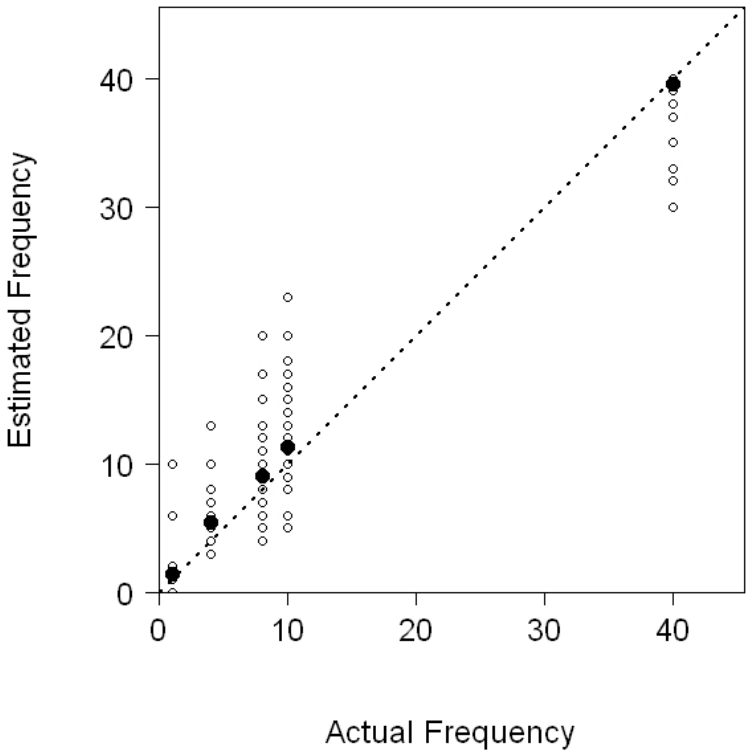
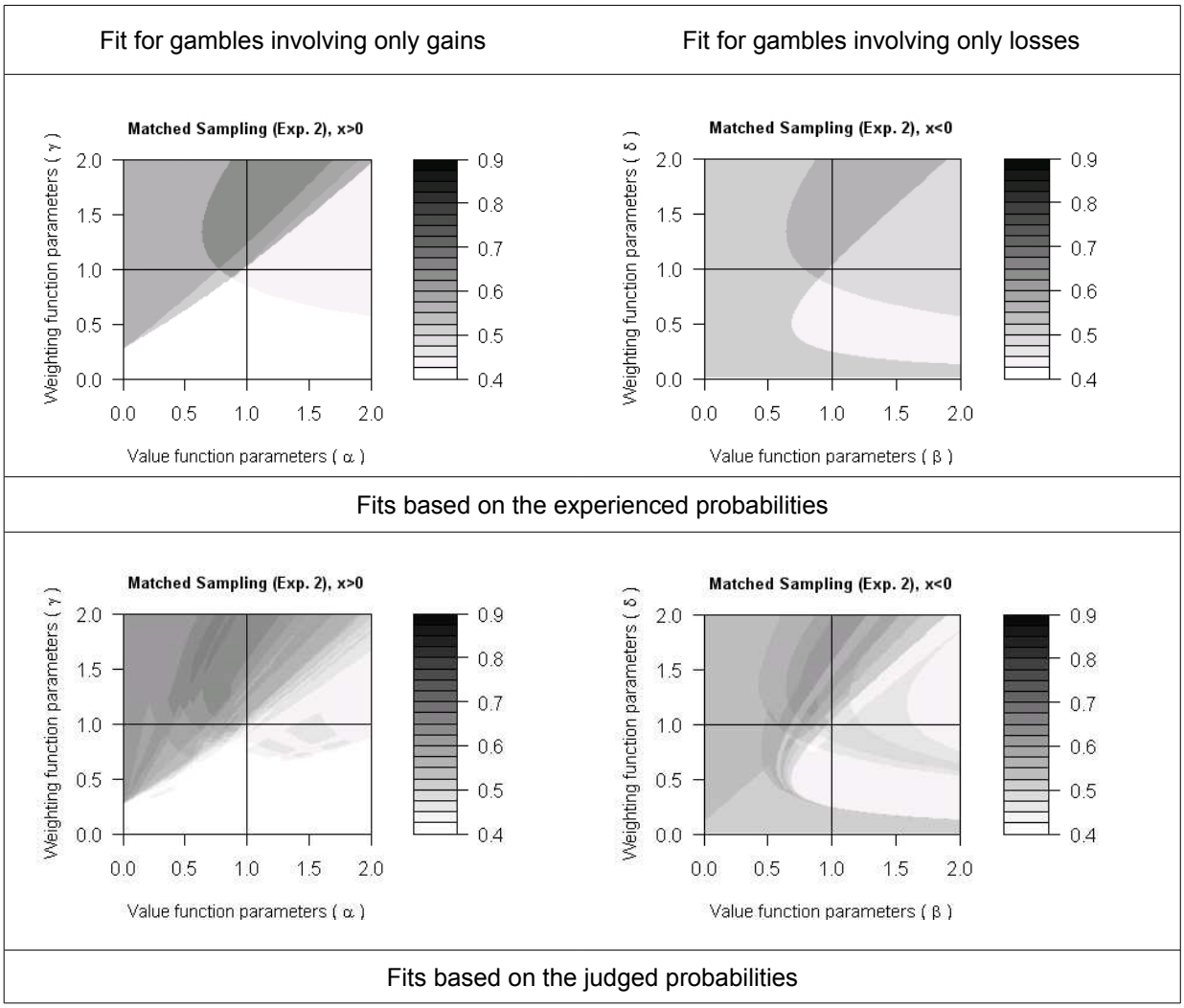


Figure 3


Fit for gambles involving only gains
Fit for gambles involving only losses
Matched Sampling (Exp. 2), x>0
Matched Sampling (Exp. 2), x<0
Weighting function parameters (γ)
Weighting function parameters (δ)
Value function parameters (α)
Value function parameters (β)
0.0
0.5
1.0
1.5
2.0
0.4
0.5
0.6
0.7
0.8
0.9
Fits based on the experienced probabilities
Fits based on the judged probabilities