Minding the Gap: On the Origins of Probability Weighting and the Description-Experience Gap[*]

Ryan Oprea[†] Ferdinand M. Vieider[‡]

August 20, 2024

Abstract

We provide evidence that "noisy coding" is responsible for both (i) classic probability weighting and (ii) its reversal when the properties of lotteries are learned by sampling rather than by explicit description. Guided by a stylized model of noisy sampling, we show that simply forcing experimental subjects to sample redundant information about the primitives of lotteries causes both probability weighting and the description-experience gap to disappear, replaced with broadly neoclassical behavior. This strongly suggests that these anomalies are a joint outgrowth of decision makers' noisy representations of the primitives of lotteries rather than expressions of true risk preferences.

Keywords: risk taking; noisy coding; probability weighting; decision from experience
JEL codes: C91, D91, G0

1 Introduction

Probability weighting is one of the key anomalies identified by behavioral economists in the last half century. As hundreds of experiments have shown, subjects, when given explicit

[*] We are grateful to Larbi Alaoui, Ido Erev, Cary Frydman, Richard Gonzalez, Ralph Hertwig, Alex Imas, Sebastian Olschewski, to participants at the Economics Psychology Seminar in Basel, the EVRE Workshop in Grenoble, the 2023 RISL$\alpha\beta$ Workshop, the Workshop in Honor of Peter Wakker, and the Summer School on the "Cognitive Foundations of Decision-Making" for helpful comments and suggestions. All errors remain our own. This research was supported by the National Science Foundation under Grant SES-1949366 and by the Research Foundation—Flanders under the project "Causal Determinants of Preferences" (G008021N). It was approved by UC Santa Barbara IRB.

[†] Oprea: Economics Department, University of California, Santa Barbara, roprea@gmail.com.

[‡] Vieider: RISL$\alpha\beta$, Economics Department, Ghent University, Belgium, fvieider@gmail.com.

descriptions of lotteries, tend to overvalue low probabilities and undervalue high probabilities in a manner that suggests that their risk preferences change with the magnitude of probabilities. This violation of expected utility theory has typically been interpreted as an outgrowth of subjects' true preferences for risk and has therefore been enshrined as a centerpiece of alternatives to expected utility theory like prospect theory (Kahneman & Tversky 1979, Tversky & Kahneman 1992).

More recently, researchers have discovered a complication to this interpretation. When subjects are required to discover the properties of lotteries by sampling from them instead of by reading explicit descriptions of their properties, probability weighting reverses: subjects instead tend to *undervalue* low probabilities and *overvalue* high probabilities, exhibiting likelihood dependence that runs opposite to the classical direction (Barron & Erev 2003, Hertwig et al. 2004). "Decision from experience" (DfE) thus produces deviations from expected utility theory that are exactly the reverse of those observed under traditional "decision from description" (DfD) protocols.

This pattern poses a challenge to the literature: in order to understand the nature of probability weighting – and by extension the nature of risk taking – we must first understand why it reverses under DfE. However researchers have found it difficult to empirically close this "description-experience gap" (the systematic differences in behavior in DfD and DfE), meaning both the gap and by extension the nature of probability weighting itself remain an open mystery. Resolving this mystery is crucial to understanding the nature of risky choice in real-world contexts, because real-world decision-making is replete with contexts both in which risks are learned by explicit description and in which they are learned purely via experienced outcomes. Understanding how choice is shaped by these contexts – and why – is central to understanding how social and economic decisions are shaped by risk.

In this paper we offer an explanation for the description-experience gap that also explains what probability weighting really is and where it comes from. Our explanation is rooted in a kind of irony: the decision-experience gap, we argue, is a consequence not only of the fact that DfD and DfE are psychologically different, but also of the fact that they are in an important sense more psychologically similar than has been previously recognized. Drawing on arguments and evidence from neuroscience, we argue that the kind of *explicit* sampling that occurs in DfE also necessarily occurs *implicitly* in the brain when a subject reasons about the properties of fully described lotteries in DfD. The noise that results from this imperfect neuronal sampling can simultaneously generate classic probability weighting and explain why previous efforts in the literature to eliminate the gap have failed. In particular, we show that efforts to close the gap by forcing subjects in DfE to sample more intensively than they naturally would has the unexpected effect of simultaneously removing the kind of coding noise needed for standard probability weighting to occur, leading to behavior broadly consistent with expected utility theory, rather than probability weighting. This means the gap can never be eliminated by forcing subjects in DfE to observe larger samples alone: in order to close the gap, we must also remove coding noise from DfD, and with it probability weighting.

We provide strong experimental evidence for this "noisy coding" explanation of the gap, and in the process produce very direct evidence that probability weighting is driven, not by true preferences that deviate from expected utility theory, but instead by the way the brain manages imprecise neural representations of information.[1] Following the distinctive predictions of our model, we show that simply by forcing subjects in DfD to sample fully redundant information from lotteries, we are able to *eliminate probability weighting.*[2] As a result, our experiment both (i) closes the description-experience gap and (ii) shows that, when cognitive frictions described by our model are experimentally removed, probability weighting vanishes and subjects make mildly risk averse lottery choices (in both DfD and DfE) that broadly comply with standard expected utility theory.

In the first step of our investigation (Section 2), we report a baseline experiment that replicates probability weighting and the description-experience gap. Subjects in Experiment 1 make binary choices between sure payments and a series of lotteries that vary the probability of earning a non-zero payment between 0.1 and 0.9. In the DfD (decision from description) treatment, we describe the payoffs and probabilities in these lotteries explicitly to subjects on their screens. In the DfE (decision from experience) treatment, by contrast, subjects are instead given buttons that, when clicked, draw a value from one of the two lottery options, but are otherwise told nothing about the lotteries. The subjects are able to learn about the lotteries by sampling as many times as they like from each lottery. We replicate the typical finding in the literature: contra expected utility theory (and most other

[1] There is growing evidence using multiple types of approaches that noisy cognition is responsible for probability weighting (Enke & Graeber 2023, Frydman & Jin 2023, Khaw et al. 2023, Oprea 2022, Vieider 2024*b*). Our contribution is to offer a particularly direct type of evidence for this hypothesis, and to show that it accounts for the description-experience gap.

[2] Here and throughout when we call sampled information "redundant" in DfD experiments, we mean it is redundant in the objective sense that the subject already has access to all of the relevant information. Our hypothesis is that this information is, however, not *subjectively redundant* in the sense that it influences the precision of the subject's belief.

preference-based models), subjects' apparent risk aversion *increases* with the probability of the non-zero payment in DfD but instead *decreases* in DfE, producing a systematic gap in behavior across the two settings.

In the second part of the paper (Section 3), we identify two sources of bias in DfE that may be responsible for the gap, using a stylized model. The first is *inference bias*: because subjects necessarily draw finite samples in DfE, the probabilistic beliefs they form are necessarily noisy. Responding to this noise in a Bayesian way, rational decision makers will combine this information with their prior beliefs about the properties of lotteries, producing systematic distortions in beliefs. The second is *sampling bias*: unless subjects draw very large samples, they will form biased beliefs due to unrepresentative samples, producing severe biases especially at extreme probabilities. The literature has discussed sampling bias as a potential source of the gap, but to our knowledge has not yet identified inference bias as a second influence. However, we show that the two biases are intimately linked: sampling bias is made more likely and worsened by inference bias, given that the latter is a major driver of the decision on when to stop sampling. Examining data from Experiment 1, we find evidence that subjects indeed under-sample and what's more that this under-sampling is linked to evidence of noisy beliefs, suggesting that it is rooted in inference bias as predicted by our model.

A key implication of this model is that both sampling and inference bias should disappear in DfE simply by forcing subjects to draw sufficiently large, representative samples. To the extent that these biases in DfE are responsible for the gap, such forced sampling should cause the gap to disappear too. In Experiment 2, we therefore introduce the DfE+forced treatment, which replicates the environment of Experiment 1 but *forces* subjects to draw a balanced, representative large sample from each option in DfE before making a choice. We find that this indeed alters behavior, removing likelihood dependence (and reverse probability weighting) from DfE as our model predicts. However, as in previous experiments (Ungemach et al. 2009, Aydogan & Gao 2020, Cubitt et al. 2022), we find that removing biases from DfE in this way does not eliminate the decision-experience gap.[3]

In the third step (Section 4), we test our key hypothesis: that the reason eliminating bias in

[3] Forced sampling is not the only strategy that has been deployed in the DfE literature to try to close the gap. Alternative attempts have e.g. used a matching technique whereby probabilities in DfD were modeled after actually observed samples in DfE (Hau et al. 2010). Obtaining the original data of the universe of DfE studies using the sampling paradigm, Wulff et al. (2018) examined the problem only using the subset of data in which probabilities were sampled correctly. However, none of these studies managed to completely close the gap. See Wulff et al. (2018), from p. 151, for a review and discussion of such attempts to close the gap.

DfE doesn't close the gap is because *similar biases* distort choice in standard DfD settings as well. Drawing on a long line of evidence from neuroscience, we argue that even fully described probabilities must be represented in the brain using a finite neuronal architecture. Because of this, beliefs in DfD are afflicted by a similar sort of imprecision as beliefs in DfE (and for a similar reason), producing parallel inference bias: combining noisily represented probabilities with prior beliefs causes a Bayesian distortion of lottery valuations, leading the decision-maker to over-weight small and under-weight large probabilities. However, because probabilities are explicitly described in DfD, DfD beliefs are not subject to the compensating effects of sampling bias. We show that because of this, inference bias will produce a pattern of insensitivity that generates probability weighting in DfD but not DfE. This not only provides a hypothesis about the source of probability weighting (i.e., as an expression of inference bias), it also explains why forced sampling does not close the gap in DfE: forced sampling *simultaneously* removes sampling *and* inference bias in DfE, preventing probability weighting from ever emerging in DfE, and thus preventing DfE and DfD behaviors from converging.

Crucially, this joint explanation for probability weighting and the persistence of the gap also provides a recipe for removing both. By forcing subjects to sample *completely redundant* information in DfD, our model suggests we can cause subjects' beliefs to become more precise, eliminating inference bias and with it probability weighting, and causing DfD and DfE behaviors to converge. To test this, in Experiment 3 we introduce the DfD+forced treatment in which we force subjects in DfD to *also* sample repeatedly from lotteries that have already been fully described to them. Providing this redundant information causes probability weighting and likelihood dependence to entirely disappear, producing DfD behavior that is identical to DfE+forced behavior and thereby entirely closing the decision-experience gap. Indeed, increasing the precision of beliefs in this way causes behavior in both settings to become remarkably neoclassical: under forced sampling, subjects in both settings exhibit mild, uniform risk aversion that is broadly consistent with standard expected utility theory.

Structurally estimating our model, and using additional data on inconsistencies in choices in repeated tasks, we show that subjects indeed hold highly noisy beliefs in *both* our DfE and DfD environments, a key premise of our explanation for probability weighting and the gap. We also show that forced sampling causes a dramatic improvement in the precision of these beliefs in both treatments, accounting for our treatment effects. What's more, as our model predicts, we find that it is precisely the subjects with the noisiest initial beliefs whose behavior is most impacted by forced sampling in DfD.

Finally, we introduce a robustness treatment that tests two additional distinctive predictions of our explanation and model. In the DfD+free treatment, subjects are allowed to randomly sample from fully described lotteries, but are not forced to. Remarkably, we find that subjects in this treatment choose to *voluntarily* sample information from already fully described lotteries – evidence that subjects, in an important sense, have residual uncertainty about the properties of lotteries that have been described to them, a key ingredient of our explanation for the gap. Next, we show that because of this residual uncertainty, voluntary sampling introduces sampling bias in DfD, another strong indication that subjects are uncertain about the properties of these fully described lotteries. In fact, this introduction of sampling bias to DfD causes DfD behavior to converge to DfE behavior, closing the decision-experience gap in a second, complementary way. Both of these findings seem to strongly reinforce our finding that subjects suffer from significant cognitive imprecisions in classic DfD environments, the key premise of our model.

Our paper contributes to several literatures. First is a long running literature on probability weighting and related anomalies, going back to Preston & Baratta (1948). Probability weighting became a key component of prospect theory (Kahneman & Tversky 1979, Tversky & Kahneman 1992, Tversky & Wakker 1995, Wakker 2010), and is the mechanism by which that theory accounts for phenomena like the coexistence of lottery play and insurance uptake and the Allais paradoxes. Numerous empirical studies have documented systematic increases of relative risk aversion in the probability of winning a prize (e.g., Hershey et al. 1982, Wu & Gonzalez 1996, Gonzalez & Wu 1999, Abdellaoui 2000, Bruhin et al. 2010, L'Haridon & Vieider 2019), a key signature of probability weighting.

The second is a literature documenting the gap between decisions from description and decisions from experience (Barron & Erev 2003, Hertwig et al. 2004). Sampling bias was proposed as an early explanation for the systematic gap observed between DfE and DfD (Fox & Hadar 2006). However, subsequent investigations showed that, although sampling bias is an important contributor to the gap, interventions including (i) eliminating sampling bias by matching probabilities in DfD to DfE, (ii) increasing the samples by offering higher stakes, and (iii) forcing people to sample the complete urn in DfE fail to eliminate the gap (Ungemach et al. 2009, Hau et al. 2010, Hertwig & Pleskac 2010, Wulff et al. 2018). Because of this, the underlying causes of the gap have largely remained a mystery—see Hertwig & Erev (2009) and de Palma et al. (2014) for narrative reviews, and Wulff et al. (2018) for a systematic meta-analysis of the decision-experience gap and possible factors contributing to it.

The third is a growing literature documenting the role noisy cognition plays in behavioral anomalies (Natenzon 2019, Khaw et al. 2021, Frydman & Jin 2022). Most closely related is a line of research examining how cognitive noise (and efficient ways the brain deals with such noise) contributes to probability distortions (Zhang & Maloney 2012, Steiner & Stewart 2016, Zhang et al. 2020, Netzer et al. 2021, Frydman & Jin 2023, Herold & Netzer 2023, Khaw et al. 2023, Vieider 2024*b*). More broadly, our work is related to a literature documenting the role cognitive frictions play in decision-making under risk (Oprea 2022, Enke & Graeber 2023, Bohren et al. 2024) and, broader still, the way cognitive constraints and the brain's response to these constraints explain a wide class of anomalies in decision making (Simon 1959, Robson 2001*a*,*b*, Netzer 2009, Robson & Samuelson 2011).[4]

2 The Description-Experience Gap

Suppose a decision maker (DM) has to make a choice between two lotteries:

- **Lottery S (safe)**: pays c with probability 1.
- **Lottery R (risky)**: pays $x > c$ with probability p (and $y < c$ otherwise).[5]

For about 70 years, researchers have been studying this choice problem in what has come to be the standard way: decision makers are explicitly told how many outcomes each lottery can produce, the payoffs each outcome results in and the probabilities of each outcome. The DM uses this information to choose the lottery she prefers. Call this standard paradigm "decision from description", or *DfD*.

One of the key regularities researchers have found in DfD experiments is *probability weighting*: experimental subjects tend to treat low probability outcomes as if they are more likely, and high probability outcomes as if they are less likely than they really are. This produces systematic differences in the severity of subjects' apparent risk aversion at low relative to

[4] A recent, contemporaneous paper, Bohren et al. (2024), documents and decomposes a complementary description-experience gap that operates in richer environments than the one we (and the previous description-experience literature) study. In evaluating realistic lotteries with many potential outcomes (e.g., eleven outcomes), they show that subjects' behavior tends to be constrained by memory limitations in DfE, while it tends to be constrained by attentional limitations in DfD. This leads to systematic differences in lottery choices in DfE and DfD environments – a gap that can be eliminated with aids to attention and memory.

[5] Although we will focus on binary lotteries in our exposition, our framework easily extends to multi-outcome lotteries via an N-dimensional generalization; see the Online Appendix for details.

high probabilities, violating standard expected utility theory (EUT) because it implies that the measured concavity of the subject's utility function increases in the probability used to obtain the measurement (Hershey et al. 1982). Observed risk attitudes thus exhibit *likelihood-insensitivity*, changing most for very small and very large probabilities near 0 and 1, but reacting insufficiently to changes in probabilities at interior probabilities (Tversky & Wakker 1995, Wu & Gonzalez 1996, Prelec 1998).

To illustrate, consider the popular linear in log odds (LLO) probability weighting function from Gonzalez & Wu (1999):

$$ln\left(\frac{w(p)}{1-w(p)}\right) = ln(\delta) + \gamma\, ln\left(\frac{p}{1-p}\right), \tag{1}$$

where $w(p)$ designates the probability weighting function.[6] In this parameterization, the intercept parameter δ governs the elevation of the function, producing (in conjunction with utility curvature) average risk aversion, while γ compresses evaluations: $\gamma < 1$ inflates odds smaller than 1, and deflates odds larger than 1, generating the pattern of likelihood-insensitivity widely found in the PT literature (Tversky & Kahneman 1992, Wu & Gonzalez 1996, Gonzalez & Wu 1999, Bruhin et al. 2010, L'Haridon & Vieider 2019). Previewing results from our experiment, panel A of Figure 1 illustrates the typical pattern using the median parameters from our DfD data.

More recently, researchers have studied an alternative paradigm to DfD for studying lottery choice, that gives us an important potential clue as to the nature of probability weighting. In "decisions from experience" (DfE) experiments (Barron & Erev 2003, Hertwig et al. 2004), subjects are told nothing about lotteries R and S but must learn all of their properties, including the number of distinct outcomes they can produce, entirely by *sampling* each of the lotteries. In standard DfE experiments (under the so-called "sampling paradigm"), subjects choose how many times to sample each lottery and use the information gleaned from these samples to make their decision.

Perhaps the most important finding from DfE experiments is that they produce a *reversal* of standard probability weighting: subjects tend to treat low probability outcomes as if they are *less likely* than they are, and high probability outcomes as if they are *more likely* than they really are. Again previewing results from our experiment, Panel B in figure 1

[6] On the probability scale this function takes the form $w(p) = \frac{\delta p^{\gamma}}{\delta p^{\gamma} + (1-p)^{\gamma}}$. Prelec (1998) presents an alternative 2-parameter function that is also often used in the literature. The two functions produce virtually identical predictions except at extreme probabilities, which are rarely tested for practical reasons.

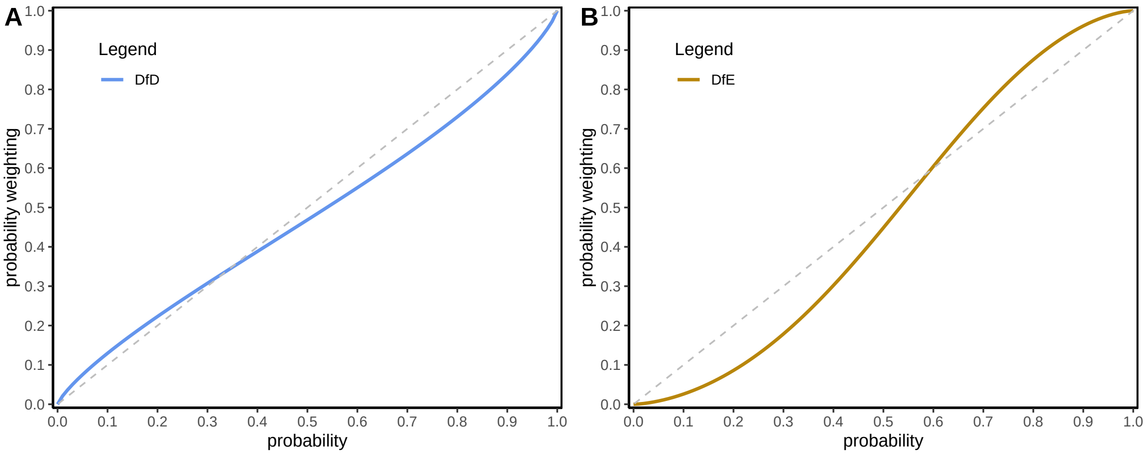


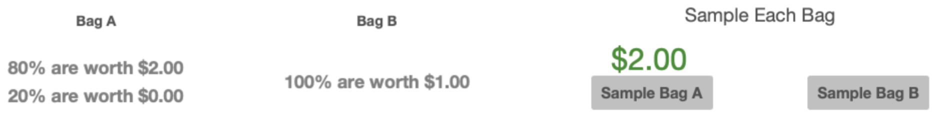
Figure 1: Probability weighting functions in DfD and DfE

The figure shows probability weighting functions based on the median parameters of a linear in log-odds function estimated from our experiments. Panel A shows the weighting function obtained from description-based choices, producing the standard pattern of likelihood-insensitivity (parameters are $\delta = 0.88$ and $\gamma = 0.81$). Panel B shows the weighting function obtained from the DfE data, which exhibits the opposite pattern of likelihood-over-sensitivity (the parameters are $\delta = 0.81$ and $\gamma = 1.56$).

illustrates the resulting LLO function using the median parameters from our experiment's DfE data (discussed in detail below).[7] In this data we find the typical pattern in DfE: small probabilities are treated as though they were less likely than they truly are, whereas large probabilities are treated as being more likely—a reversal of standard DfD probability weighting.

The inverted responses in DfE and DfD produce what the literature has called the "decision-experience gap" (hereafter, simply the GAP) in lottery choice – an open mystery in the literature. Understanding the source of this gap is crucial to understanding not only the way description and experience differentially shape choice, but also the true nature of the classical phenomenon of probability weighting itself. Despite many attempts, researchers have failed to close this gap, and therefore have failed to fully uncover why behaviors in these two settings differ as dramatically as they do. In this section, we will begin by reporting an experiment that replicates this GAP, illustrating its features.

[7] The estimate is based on a "naive PT estimation" using the actual outcome-generating probabilities (see Wulff et al. 2018, pp 157-159, for a discussion of such estimations in the literature).



Bag A

80% are worth $2.00
20% are worth $0.00

Bag B

100% are worth $1.00

Sample Each Bag

$2.00

(a) DfD Treatment (b) DfE Treatment

Figure 2: Screenshots from Experiment 1.

2.1 Experiment 1: DfD and DfE Treatments

In Experiment 1, we replicate the GAP between decision from description (DfD) and decision from Experience (DfE). In our experiment, subjects face 18 distinct binary choice problems like those discussed in the previous section: one lottery pays an amount c with probability 1, and the other an amount $x > c$ with probability p and $y < c$ otherwise. These lotteries vary p across 0.1, 0.15, 0.2, 0.8, 0.85 and 0.9 and vary payoffs x, y and c. The outcomes are chosen so that the sure amount c is symmetric around the expected value (EV) of the lottery. This will allow us to get a rich picture of behavior, and is crucial for identification in structural exercises reported in Section 6, below. Details on parameters are provided in Online Appendix B and we will use identical lotteries in each of the experiments reported later in the paper.

Lotteries are described to subjects as “bags,” containing 20 “coins,” each of which is worth a different amount of money. After choosing a lottery, a single coin is drawn from the bag to determine the subject’s payment.

Treatments. Experiment 1 consists of two treatments. In the Decision from Description treatment (DfD), the subject is explicitly told the properties of each lottery (i.e., the contents of each bag); Figure 2a shows a screenshot. A pair of radio buttons below the lottery description allows the subject to make and submit a choice between the two lotteries (i.e., a choice of which lottery will be realized to determine the subject’s payment).

In the Decision form Experience (DfE) treatment, the subject is instead shown two buttons, one for each of the two lotteries/bags. Figure 2b shows a screenshot. When the subject clicks on the button, she is shown a single realization of the lottery (i.e., a single draw from the bag, *with* replacement). The subject is told nothing about either lottery ahead of time and thus must learn all of their properties (e.g., the number of possible outcomes, the relevant probabilities, the payoffs in each outcome etc.) by sampling. The subject in the Figure 2b example has just clicked the “Sample Bag A” button and drawn $2. Each sample is shown for 0.5 seconds. Subjects are allowed to sample as many (or as few) times as they like from the two bags, with no time constraints. Below the sampling buttons are the same two radio buttons shown beneath the lottery descriptions in DfD, and the subject can choose one of the lotteries to determine her payment whenever she is ready.

Repetition. Unbeknownst to subjects, four of the 18 lotteries are randomly selected (at the subject level) to be repeated. This randomization is included to allow us to measure the noisiness of subjects' decisions (useful for the structural estimation in Section 6). Thus subjects face, in total, 22 lotteries in an order that is randomized at the subject level.

Stages. Each session in the experiment proceeds in two stages. In Stage 2, subjects experience their main treatment: 22 randomly ordered lottery choices under DfD or DfE, depending on treatment. In Stage 1, subjects face the same 22 binary choice tasks under DfD (in a different random order). We included Stage 1 for several reasons. First, doing this allows us to examine the description-experience GAP both within-subject (by comparing Stage 1 and Stage 2 in the DfE treatment) and between-subjects (by comparing Stage 2 in the DfE vs. DfD treatment). Second, including Stage 1 is useful for fixing prior beliefs about lotteries and linking DfD and DfE behavior, both of which will be useful for structural estimation in Section 6.

Implementation. We ran 99 subjects through the DfD treatment and 99 subjects through the DfE treatment on Prolific in September of 2023. We paid all subjects \$6 and selected ten percent of them to be paid based on a lottery outcome from a randomly selected task. The median subject spent 18 minutes in the experiment and the average subject earned \$18.67 per hour. Instructions to subjects, including four comprehension questions, are included in Online Appendix H.

2.2 Results

We focus on results from the main part of the experiment (Stage 2) and on the between-subjects contrast between treatments; in Online Appendix G we show virtually identical results in within-subjects comparisons between Stage 1 and 2 in the DfE treatment. Panel A of Figure 3 plots the fraction of times subjects chose the risky lottery in DfD versus DfE, aggregated by probability of winning, p, and pooling across values of c.[8] This figure repli-

[8] Figure 9, included farther below, provides an overview of choice proportions at the task level resolution. We will also make use of variation across values of c in our structural analysis in Section 6. See Online Appendix G for graphs that break the analysis down by c.

cates the main stylized patterns from the DfE literature—risk taking is more pronounced in DfD than DfE for small probabilities, and this difference reverses for large probabilities, with DfE eliciting higher levels of risk taking. The size of the gap in choice proportions between treatments is relatively small at low probabilities, but becomes very large at large probabilities.[9] As we show farther below, the GAP is also strongest in tasks in which the expected value of the lottery exceeds the sure amount – the main type of task studied in the early DfE literature. The fact that the size and even direction of the GAP are sensitive to task characteristics like these is consistent with prior work (Glöckner et al. 2016).

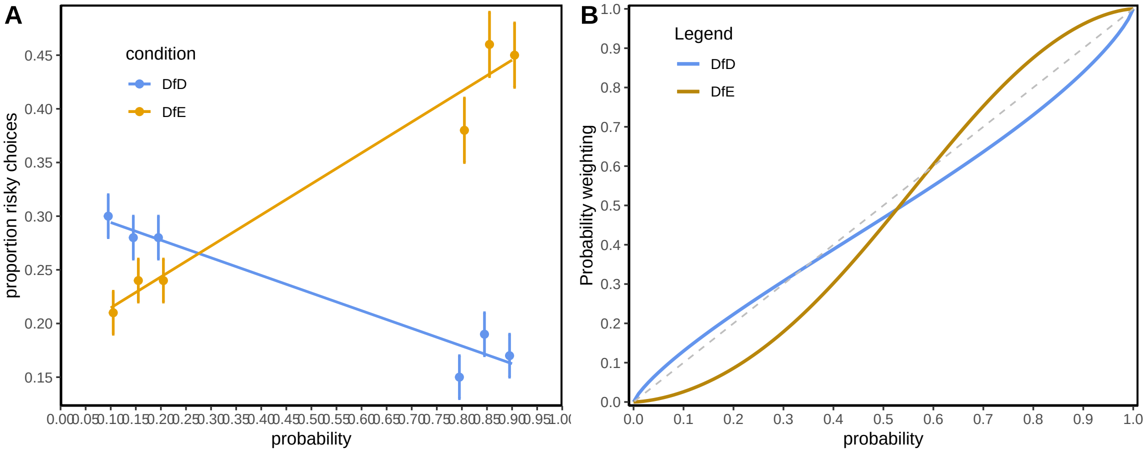


Figure 3: The GAP: Decisions from Description vs Decisions from Experience

The figure shows choice proportion for the lottery in DfD and DfE, ordered by probability of winning. Panel A shows the between subject comparison using the DfD and DfE treatments, fitted with a linear regression line. The error bars indicate 1 standard error. Panel B shows structural estimates of the LLO probability weighting function for each treatment, using median parameter estimates.

Quantifying the GAP. To get a better idea of the size of the decision-experience GAP in our data, and to relate it to typical findings in the literature, we can aggregate the evidence across tasks using the tools of meta-analysis.[10] Let $\pi_d = {}^{R_d}/_{N_d}$ be the proportion of risky choices in DfD, where R_d is the number of risky choices, and N_d the number of observations. Let $\pi_e = {}^{R_e}/_{N_e}$ be the proportion in DfE. We define the difference in choice proportions as g, where we encode the difference in the direction of the standard gap, so that $g = \pi_d - \pi_e$ for $p < 0.5$ and $g = \pi_e - \pi_d$ for $p > 0.5$. This difference will be approximately normally

[9] This is consistent with the fact that the risk seeking for small probabilities documented using certainty equivalents in DfD is typically small or nonexistent in binary choice settings like ours (Bouchouicha et al. 2023).

[10] The meta-analytic tools we use are identical to a "measurement error model". That is, the assumption is that each single choice proportion is observed with some error. Meta-analysis then allows us to aggregate across the choice proportions while eliminating measurement error and thus correcting our analysis for multiple testing across many moderate (and not statistically independent) samples.

distributed, with variance $\pi(1-\pi)\left({}^{1}/_{N_d}+{}^{1}/_{N_e}\right)$, where $\pi=\frac{\pi_d+\pi_e}{N_d+N_e}$. We can now use g and its associated standard error, se, for meta-analytic aggregation across tasks, indexed by i:

$$g_i \sim \mathcal{N}(\,\widehat{g}_i\,,\,se_i^2\,)$$
$$\widehat{g}_i \sim \mathcal{N}(\,\omega\,,\,\tau^2\,),$$

where g and se are data, $\widehat{g}$ is the unknown true effect, and ω and τ are parameters capturing the meta-analytic mean and standard deviation across tasks, respectively (further details and the code used can be found in Online Appendix D). We quantify the GAP by meta-analytically aggregating the differences in choice proportions across tasks in a direction that is consistent with the standard GAP. Between subjects we find an average GAP of 15.7 percentage points (*pp*), with a 95% credible interval of [9.7 , 21.8] pp. The GAP is clearly significant and large: in their meta-analysis of the GAP, Wulff et al. (2018) report a meta-analytic average of 9.7 pp, meaning we find a somewhat larger description-experience gap than is typical in the literature.

Reversals in Likelihood Dependence. We can also use meta-analysis to test whether choice proportions exhibit likelihood-dependence, and whether the nature of this dependence is different in DfE and DfD. To do this, we analyze the choice proportions π_i directly (instead of examining differences in choice proportions g_i) so that we estimate $\pi_i \sim \mathcal{N}(\,\widehat{\pi}_i\,,\,se_i\,)$. We then use meta-regression to assess the dependence of the choice proportions on the probability of winning, by letting $\widehat{\pi}_i \sim \mathcal{N}(\,\lambda_0+\lambda\times p_i\,,\,\tau^2\,)$, where $\widehat{\pi}_i$ is the unknown true choice proportion.

Beginning with DfD, we obtain a coefficient of $\lambda=-0.172$, with a 95% credible interval of $[-0.242\,,\,-0.098]$ (within-subject analysis yields very similar results). Risk taking thus clearly decreases in the probability of winning in DfD, matching the typical finding in the prospect theory literature. This contrasts sharply with findings using the DfE data, where meta-analytic regression of choice proportions on true outcome-generating probabilities produces a $\lambda=0.284$, with a credible interval of $[0.226\,,\,0.347]$. Probability-dependence of risk taking is thus strong in both DfD and DfE, but runs in exactly opposite directions. This is the typical finding in the DfE literature.

In Panel B of Figure 3, we illustrate this reversal by once again showing structural estimates of an LLO parameterization of the probability weighting function. We structurally estimate this model on choice data for each subject in the dataset using Bayesian hierarchical techniques, and plot the resulting probability weighting function for the median parameters in DfD and DfE (Online Appendix F discusses the details of the estimation, discusses identification, and provides the code). Because of the reversal of the direction of likelihood dependence in DfE, the standard pattern of probability weighting documented in the prior literature for DfD is reversed in DfE. Once again, this is perhaps the most salient finding from the literature on decisions from experience.

3 Removing Bias from DfE

What is responsible for the GAP and the reversals of likelihood dependence that produce it? Given the relative novelty of DfE (and the widespread tendency to interpret probability distortions in DfD as outgrowths of true preferences) it is natural to begin by trying to understand potential biases in DfE. In this section, we highlight two basic biases we should expect in DfE for even a rational decision-maker. The first (long-emphasized in the literature, e.g. Fox & Hadar 2006, Hertwig & Pleskac 2010) is *sampling bias*: unless the DM collects a large sample, she runs the risk of drawing misleading samples that systematically distort beliefs particularly at extreme probabilities. The second (which has not been emphasized in the literature so far) we will call *inference bias*: because the DM's sample is finite, she cannot be entirely confident in the sample she draws. This will make it optimal to combine such samples with her prior beliefs in a Bayesian fashion, distorting her posterior beliefs. As we will show, these biases should be related to one another and can jointly produce the GAP.

In this section we build a model of DfE designed to show where these two biases come from, examine their relationship and motivate our next empirical steps. To fully specify a model of DfE, we must describe not only how people form beliefs about probabilities and payoffs (the beliefs that matter for understanding these biases), but also how these beliefs co-evolve with higher order beliefs about the structure of the lotteries (e.g., the number of outcomes in each lottery's support). To close the model, it is therefore necessary to make a number of detailed modeling choices about the evolution of these structural beliefs that are unrelated to the qualitative properties of inference and sampling biases and that therefore do not directly impact the way we interpret and design our experiments. In the Online Appendix we propose such a fully specified model.[11] But in this section, for expositional ease, we

[11] In the full version of the model in Online Appendix A, we close the model by assuming that (i) subjects mainly use samples to build beliefs about the comparative properties of the two choice options (which seems likely given the choice subjects face), (ii) that subjects know that they are making a risky choice and that the choice is therefore not between two degenerate lotteries (which seems likely given the lotteries subjects

abstract from these issues of higher order belief formation altogether by (i) assuming that subjects already know the structure of the lotteries (i.e., assuming beliefs evolve according to a Beta instead of the Dirichlet used in the full model in the Appendix)[12], (ii) assuming that subjects quickly identify which lottery is risky during sampling and (iii) focusing attention in the model on the way subjects evaluate the risky arm. In the fully specified, general model in Online Appendix A we discuss the implications of these assumptions, but argue that they are qualitatively irrelevant to the key matters at hand.

To model the way beliefs change as the DM samples the simple binary lotteries from Experiment 1, let α be the number of draws in which the DM observed payment x and β the number of draws in which she observed payment y.[13] We model this sampling process using a Beta distribution with parameters α and β, producing a representation of the probability p of earning x (e.g. in lottery R described above) equal to $\mathbb{E}[\,\widehat{p}\,|\,p\,] = \frac{\alpha}{\alpha+\beta}$ (i.e.the sampled mean probability $\widehat{p}$, given the true probability p).[14] We will assume that the DM's beliefs are represented in a log-odds form. This is not necessary for any of our qualitative conclusions in what follows, but (i) it is increasingly supported in neuroscience both empirically and theoretically[15] and (ii) it will allow us to neatly connect our characterization to the

exclusively see in Part 1 of the experiment) and (iii) make a few other technical assumptions required to fully specify the joint inference problem. The main implication of (ii) is that inferences in which the outcomes observed in both choice options are attributed probability close to 1 will carry very high noise, in a sense to be made precise below. Within the formalism of the model, this assumption mainly serves to explain why subjects take more than 1 sample from each option.

[12] In our experiment, this is in fact a fairly realistic assumption, given that subjects entering DfE have all just made a number of lottery choices, all with the same structure.

[13] For simplicity, we focus attention here on the evolution of beliefs about the non-degenerate lottery, which are the beliefs relevant for understanding these biases. Although beliefs about the degenerate lottery clearly matter quantitatively, the DM's decision is comparative between the two options which means this has little bearing on qualitative properties of the model. Of course, in Online Appendix A.1 we remove these simplifications in the general version of the model.

[14] It is important to emphasize that our model *does not require us to assume* that the DM knows the structure of the decision problem. We use a Beta distribution here purely for expositional simplicity, and because binary lotteries is all a DM will ever experience in our experiments. Our model generalizes to any number of outcomes by using a Dirichtlet distribution—the multi-dimensional generalization of the Beta—to represent the different states. Indeed, we can use Dirichlet distributions defined over all possible outcomes to explicitly model the inference process of the DM about the underlying state space in DfE—an important element that distinguishes our approach from some of the DfE literature in economics, which has assumed that the DM (often counterfactually) knows the objective state space or which has (in some papers) provided this information ex ante in experiments (Abdellaoui et al. 2011, Aydogan 2021, Cubitt et al. 2022). Online Appendix A provides details of the inference process, and of how the model we use here can be generalized to N states of nature.

[15] It is common in neuroscience to assume that the brain represents the sort of evidence encoded by α and β in terms of log-odds. This is in part because of its computational efficiency for the brain, a straightforward consequence of the fact that new evidence can be simply added to pre-existing evidence, which is a much less computationally expensive operation than, e.g., multiplication. (Indeed, Gold & Shadlen (2002) show how just such a computationally tractable choice rule was used by Alan Turing to decode the Nazi navy's Enigma code. In the absence of modern computing power, being able to additively combine evidence proved

standard LLO functional form of the probability weighting function (used to characterize our findings in Experiment 1 and in our structural estimation below).

Inference Bias. Because α and β are finitely sampled, they produce noisy beliefs about the true probabilities. Note that such beliefs would be noisy in finite samples even if the samples were accurate on average, since the DM can never be 100% sure of whether a *given* sample correctly reflects the underlying outcome-generating probability. We can thus represent the typical inference on the log-odds as follows:

$$ln\left(\frac{\widehat{p}}{1-\widehat{p}}\right) = ln\left(\frac{\alpha}{\beta}\right) \; = ln\left(\frac{p}{1-p}\right) + \varepsilon, \tag{2}$$

where the error term ε captures uncertainty in the representation of the true log-odds due to the finite sample.

The log-odds formulation gives rise to approximately normally distributed errors even with relatively few observations (see e.g. Gelman et al. 2014, section 5.6), which we will (given our Experiment 1 results) assume. Following the characterization of the logit-normal distribution by Atchison & Shen (1980), it is straightforward to obtain the average noise from the sampling draws representing the odds:

$$\varepsilon \sim \mathcal{N}(\,0\,,\,\nu^2\,) \;\;,\;\; \nu^2 = \digamma'(\alpha) + \digamma'(\beta), \tag{3}$$

where $\digamma'$ represents the trigamma function. We will refer to ν^2 as the *inference noise*, and it is clear that this noise will decrease in the number of draws $\alpha + \beta$, at a decreasing rate (an inherent characteristic of the Beta distribution).[16]

Given this inference noise, a Bayesian DM will rationally combine the results of her sampling with her prior beliefs about the log odds. Continuing with our log-odds characterization of beliefs, assume the prior takes logit-normal form

crucial for this process.) It is also in part because of the empirical success of such representations. For instance, Zhang & Maloney (2012) describe log-odds representations as "ubiquitous", discussing a long list of findings which can be fit by log-odds representations. Glanzer et al. (2019) identify a unique empirical signature of log-odds representations, and argue that such representations underlie neural representations in general.

[16] The sum $\alpha + \beta$ is known as the *concentration* of the Beta distribution, which can be interpreted as a measure of confidence in the mean belief. Equivalently, we can thus interpret the *precision* of the log-odds presentation, ν^{-2}, as a measure of confidence in the sampled log-odds. Olschewski & Scheibehenne (2024) present a discussion of different types of noise arising when decision-makers need to infer (and bet on) means of a series of sampled numbers, and present a concept of "Thurstonian uncertainty" that resembles what we here call inference noise.

$$ln\left(\frac{p}{1-p}\right) \sim \mathcal{N}\left(ln\left(\frac{p_0}{1-p_0}\right)\,,\,\sigma^2\right). \tag{4}$$

Even if samples are themselves unbiased, combining information from noisy samples with prior beliefs will produce *inference bias.* As we show in more detail in Appendix A.3, the average posterior expectation of the log-odds being inferred, $\widehat{\ell o}$, conditional on the true log-odds, ℓo, will take the following form:

$$\begin{aligned}\mathbb{E}\left[\widehat{\ell o}\,\middle|\,\ell o\right] &= \frac{\sigma^2}{\sigma^2+\nu^2}\,ln\left(\frac{p}{1-p}\right) + \frac{\nu^2}{\sigma^2+\nu^2}\,ln\left(\frac{p_0}{1-p_0}\right)\\ &= ln\left(\frac{p}{1-p}\right) + (1-\gamma)\left[ln\left(\frac{p_0}{1-p_0}\right) - ln\left(\frac{p}{1-p}\right)\right].\end{aligned} \tag{5}$$

where $\gamma = \frac{\sigma^2}{\sigma^2+\nu^2}$ is the Bayesian evidence weight. The second line of this equation shows this *inference bias*: the true log odds, shown in the first term, are distorted by the bias captured by the second term. The larger the inference noise ν, the smaller the Bayesian evidence weight γ, and the larger the inference bias will be.

Sampling Bias. The preceding discussion assumed that α and β are, on average, sampled in an unbiased way (i.e., $ln\left(\frac{\alpha}{\beta}\right) = ln\left(\frac{p}{1-p}\right)$ on average). However, the binomial distribution will produce samples that are skewed towards 0 in lotteries with small probabilities, and skewed towards 1 in lotteries with large probabilities (Fox & Hadar 2006, Hertwig & Pleskac 2010). Unbiased samples in any given task will only obtain if the DM takes very large samples. As it turns out, this is not the case in our data. As we show in more detail in Online Appendix G (and in figure 4 below), subjects in our DfE treatment chose to sample only 8 draws on average—far too few to produce reliably unbiased samples. As a result, we should expect the ratio of α and β observed by subjects to produce systematically biased impressions of the log odds (even setting inference bias aside). This is particularly true of samples taken from lotteries with extreme probabilities, where sampling bias is most likely and where the gap between description and experience is most severe.[17]

To understand why DMs tend to undersample in DfE (and to set up some structure and notation that will be useful in our structural estimates later in the paper), we model the DM's choice problem in DfE. As we show in Online Appendix A, expected value maximization in the simple choice problems from Experiment 1 entails a choice rule in which the

[17] For instance, close to 50% of subjects taking 7 samples from a lottery providing an objective probability of 0.1 of winning a prize x will never observe that prize. Even after 10 draws, only about 40% will draw a sequence correctly representing the underlying probability.

DM trades off the log-odds against the log cost-benefit ratio, $ln\left(\frac{c-y}{x-c}\right)$. For expositional simplicity, we will assume in this section that the log cost-benefit ratio (unlike the log-odds) is objectively perceived, though clearly it will in fact be learned by sampling just as the log odds are. This assumption will have no impact on our qualitative predictions here but greatly simplifies the exposition.[18] In Appendix A.3, we show that this yields the following *discriminability* equation:

$$\psi = \frac{\gamma \times ln\left(\frac{\widehat{p}}{1-\widehat{p}}\right) - ln\left(\frac{c-y}{x-c}\right) - ln(\theta)}{\nu \times \gamma}, \tag{6}$$

where $\theta \triangleq \left(\frac{1-p_0}{p_0}\right)^{1-\gamma} = \delta^{-1}$ can be interpreted as a measure of "risk aversion" generated by the distorting influence of the prior.[19] Note that we now use $\widehat{p}$ instead of p because in any given task, the inferred probability may be affected by both inference bias *and* sampling bias. As the DM samples from the lottery, ψ adjusts, with positive values supporting the risky option and negative values the safe option. The DM stops sampling once ψ reaches a sufficiently extreme value, passing a discriminability threshold (see Appendix A.3 for details). When exactly this threshold is reached will depend on θ, with larger values of θ making it harder to pass the threshold needed for a choice of the risky option. This is the sense in which θ captures average risk aversion in the model.[20]

Why might this choice process lead to under-sampling, and thereby sampling bias? A major ingredient is inference bias itself. At low probabilities, the skewed nature of the binomial distribution will tend to produce exclusive initial draws in favor of the low outcome y. This will result in high levels of inference noise initially (cfr. Online Appendix A.3), and thus shrinkage towards the ("risk averse") prior θ. Sampling bias and inference bias thus reinforce one another, pushing the DM to take few samples and resulting in an over-selection

[18] It is straightforward to extend the model to include inference bias in cost-benefit perceptions—see Vieider (2024*b*). In our quantitative analysis (our structural model) below, we will take explicit account of the effects of sampling on the DM's beliefs about the cost-benefit ratio.

[19] Note that we do not *assume* the prior to entail risk aversion. We rather treat it as a free parameter through which any underlying risk aversion of the DM may manifest in the model.

[20] Since the prior can be expected to change only very slowly, we will take it to be fixed for the purposes of interpreting results from the experiment. Indeed, as in Experiment 1, in all of our experiments we employ a two-part within-subject design, in which subjects see the same tasks in the first and in the second part, making such an assumption especially empirically plausible. What is assumed fixed in our modeling, are the parameters p_0 and σ. Both the threshold parameter θ and the likelihood-discriminability parameter γ will be endogenous due to variation in inference noise, ν. Beyond that, we do not require any specific assumptions about the prior. In particular, the prior over the unit interval of probabilities could take any shape, including one with multiple peaks. The mapping back onto the log-odds scale will ensure that the prior is well behaved, and that it will typically abide by the normality assumption. This is indeed a standard assumption for log-odds transformations in statistics—see e.g. Gelman et al. (2014), section 5.4.

of the safe lottery. Conversely, high probability lotteries will often produce exclusive early evidence of the prize x. This will again result in high levels of inference noise, but inference bias and sampling bias now pull in opposite directions. This will delay arrival at the discriminability threshold and lead to much more sampling than at low probabilities. This will yield a reduction in average sampling bias concomitantly to a reduction in inference bias, thus resulting in higher levels of risk taking (especially by DMs who continue to draw "optimistic" samples that overestimate the true probability). As we discuss in more depth in Appendix A, taken together this will tend to produce risk averse choices at low probabilities and risk-tolerant choices at high probabilities: exactly the pattern we documented in the previous section for DfE.

Evidence of Sampling and Inference Bias in DfE. This line of argument suggests that the relatively low risk taking at small probabilities and high risk taking at large probabilities typically observed under DfE is due not only to sampling bias (a bias the literature has previously discussed), but also to inference bias. Importantly, our model suggests a distinctive test for inference bias by predicting that (i) sampling behavior should vary with the probability of the prize, p; and (ii) this dependence should vary according to the subject's pre-existing level of risk aversion as captured by the prior mean. This follows from equation (5), which shows that inference bias is a function not only of inference noise, but also of the difference between the outcome-generating log-odds and the prior mean, i.e. of $ln\left(\frac{p_0}{1-p_0}\right) - ln\left(\frac{p}{1-p}\right)$. Lower levels of p_0 – which, in the model, produce a more risk averse prior – will intensify shrinkage for large probabilities while attenuating it for small probabilities, thus increasing the likelihood-dependence of sampling behavior. Highly risk averse DMs should thus sample more at higher relative to lower probabilities; this pattern should be weakened or even reversed for DMs with risk tolerant priors.

Intuitively, the more risk averse a DM is to start with, the more evidence will be required to convince her to choose the risky option R. Inference bias will thus ultimately determine both sampling and observed risk taking in our DfE data. Uniform samples of y observed for small probabilities agree with a risk averse prior, so that discriminability quickly becomes very negative and triggers a choice of the safe option. Uniform samples of the prize x for large probabilities, however, clash with risk averse priors, thus increasing samples. This will reduce inference bias and shrinkage, and thus produce risk taking especially by DMs drawing relatively "optimistic" samples when compared to the true probability p.

To test this, we categorize DfE subjects (from Experiment 1) according to their risk aversion

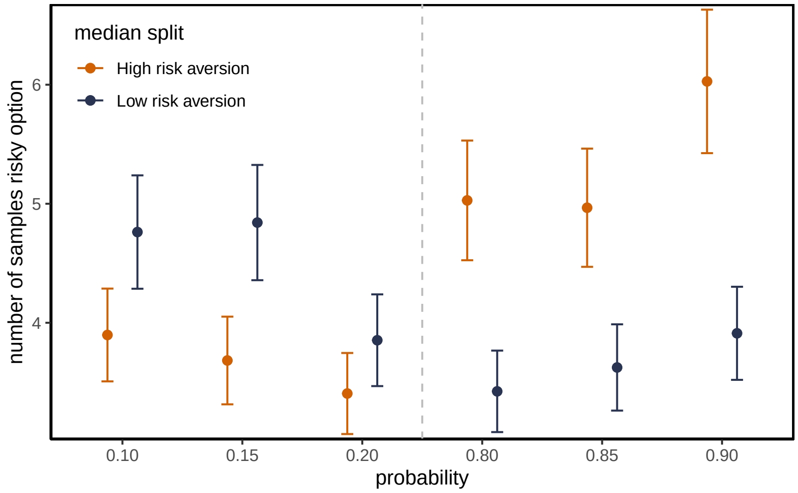


Figure 4: Samples by probability and risk aversion

The figure shows the number of samples taken from the risky option by probability and risk aversion. Risk aversion is assessed as the proportion of safe choices in the first, DfD part of the experiment, after removing repeated tasks. The categorization is obtained using a median split. Error bars show ±1 standard error.

using their propensity to choose risky lotteries in Stage 1 by simply counting the number of Stage 1 risk averse choices subjects make. Splitting the sample at the median, we classify subjects as exhibiting Low or High ex ante risk aversion.[21] In Figure 4, we plot the mean number of samples taken from the risky option in Stage 2 as a function of probability p for the High and Low risk aversion subsamples. We find clear evidence of the predicted pattern: highly risk averse subjects sample substantially more at high than at low probabilities; relatively risk tolerant (Low risk aversion) subjects show (somewhat weaker) evidence of the reverse sampling pattern. Given that most subjects in our sample are risk averse this results in an overall average tendency for subjects to sample more for larger than smaller probabilities (see Online Appendix G for regressions using continuous measures of risk aversion, and which confirm these patterns statistically). The results thus strongly support the idea that inference bias is a major determinant of beliefs and therefore sampling behavior in DfE.

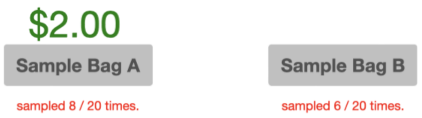
Along with this evidence of inference bias, our DfE data also shows clear evidence of sampling bias—the other bias driving the GAP according to our model. DfE samples are quite small for both relatively risk averse and risk tolerant subjects: as Figure 4 shows, the av-

[21] As we will explain shortly, we expect behavior in DfD to also be affected by inference bias, so that this measure is only a *proxy* for risk aversion as captured by the prior. The structural estimations in section 6 show that our results are robust to using a theoretically cleaner measure of risk aversion.

erage sample for any one lottery is never greater than 6. Clearly this is far too small of a sample to produce an accurate estimate of extreme probabilities in which the GAP is largely concentrated. In Online Appendix G, we show that this under-sampling leads to systematic bias in α and β in our DfE data, producing systematically biased values of p. For small probability lotteries, our subjects gather samples that produce a smaller probability than the true one in 66% of cases overall, and an accurate sample in some 3.4% of cases. For large probability lotteries this is reversed, with 55% of samples over-estimating the true probability, and only 2.2% resulting in a correct estimate. Sampling bias is clearly more severe at low than high probabilities, a predicted consequence of inference bias in our model (see Online Appendix G for statistical evidence). The inverse probability weighting observed in DfE is thus driven by an interaction of sampling bias and inference bias.

De-Biasing DfE. These results suggest that the GAP is driven by systematic sampling and inference bias, both of which derive from the same source: under-sampling which produces both sampling bias and (by generating noisy beliefs) inference bias. This in turn gives us a first clue as to how to close (and thereby explain) the GAP: by forcing subjects to sample more than they naturally would choose to do, we will simultaneously reduce and perhaps even remove both sampling and inference bias. In particular, by forcing subjects to observe a balanced sample (that correctly represents the true probabilities in frequencies, i.e. a sample such that $\frac{\alpha}{\alpha+\beta} = p$) we remove sampling bias, and by making this sample sufficiently large we reduce ν and thereby reduce inference bias. In the next section, we design an experiment that will allow us to test this hypothesis.

3.1 Experiment 2: DfE+forced Treatment



sampled 8 / 20 times.

sampled 6 / 20 times.

Figure 5: Screenshot from the DfE+forced treatment (Experiment 2).

In Experiment 2, we attempt to eliminate the decision-experience GAP by forcing DfE subjects to sample from each lottery (i) using a representative sample (removing scope for sampling bias) and (ii) via a relatively large number of draws (removing scope for inference bias). We do this using the DfE+forced treatment, pictured in Figure 6. This treatment is identical to DfE except that subjects are required to sample all twenty "coins" from each bag (lottery) *without replacement* before making a choice between lotteries. Below each button, the subject is shown how many times she has sampled from each bag and the total number of draws she must make in total (set to 20 in this treatment). The radio buttons for submitting the final lottery choice do not appear on the subject's screen until she has sampled all 20 coins from each bag.

In terms of the model, requiring the subject to exhaustively sample a frequentist representation of each lottery means that subjects observe samples α, β for each lottery such that $\frac{\alpha}{\alpha+\beta} = p$, removing scope for sampling bias. By setting the number of elements in the frequentist representation to 20 (20 "coins" in each bag), we force subjects to sample far more times than they are observed to do in the DfE treatment, reducing ν and therefore reducing scope for inference bias.

In all other respects the experiment is identical to the DfD and DfE treatment. Subjects are assigned the same 18 lotteries, repeat four of these lotteries selected at random and make choices in these lotteries in a DfD treatment in Stage 1 before entering the DfE+forced treatment in Stage 2. The experiment was conducted on Prolific in September of 2023 using 96 subjects.

3.2 Results

As we've just shown, (i) subjects in DfE sample on average fewer than five times from each option producing highly noisy beliefs, subject to inference bias; and (ii) draw unbalanced samples that are overly pessimistic at low and optimistic at high probabilities for the majority of subjects. Theoretically, our DfE+forced treatment in Experiment 2 resolves (i) by quadrupling the size of the sample and resolves (ii) by balancing the composition of the sample subjects observe. Our key prediction is that this removal of the inference and sampling biases will reduce or eliminate the positive likelihood dependence (i.e., the reverse probability weighting) characteristic of DfE behavior, thus narrowing the GAP.

In Panel A of Figure 6 we plot choice behavior from DfE+forced, and reproduce behavior from DfE for comparison. As predicted, forced sampling produces a dramatic effect on behavior, particularly in reducing the high levels of risk taking observed for large probabilities. Importantly, as predicted, DfE+forced does this largely by virtually eliminating likelihood dependence, suggesting that it mostly removes both sampling and inference bias from DfE. Regressing choice proportions observed after forced sampling on the probability of winning

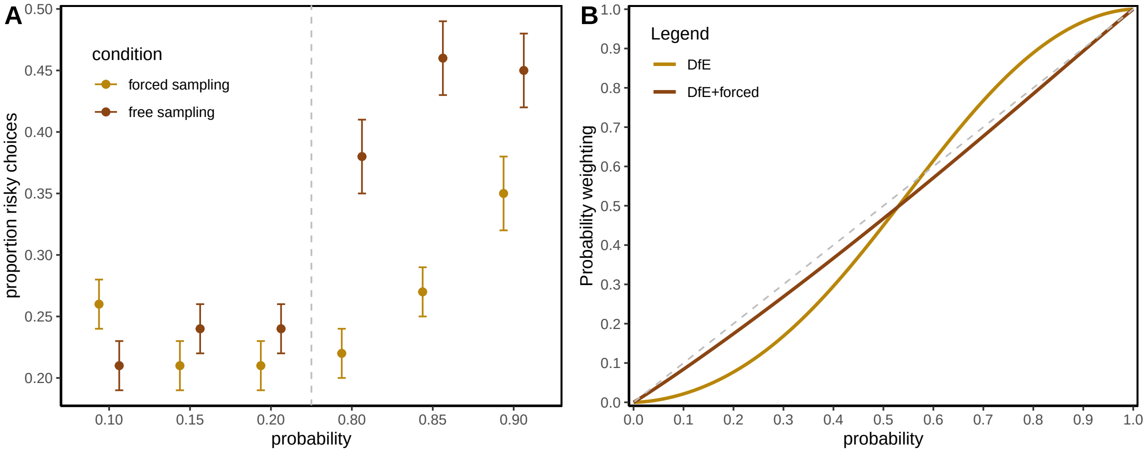


Figure 6: Effects of forced sampling in DfE on choice proportions

The figure shows the effects of forced complete sampling in DfE from both options. Panel A shows the effect of forced sampling on choice proportions of the risky option in DfE, and compares it to DfE with free sampling. Panel B shows the LLO function in DfE and compares it to the equivalent function in DfE+forced. The functions shown are based on the median estimates from a Bayesian hierarchical model, which are $\gamma = 1.64$ and $\delta = 0.81$ in DfE, and $\gamma = 1.02$ and $\delta = 0.87$ in DfE+forced. The error bars indicate ± 1 standard error.

the prize, we find that the strong positive effect of the winning probability on choice proportions, documented in DfE, disappears. With forced sampling, choice proportions no longer significantly depend on probability, with $\lambda = 0.080$ and a CrI of $[-0.020\ ,\ 0.182]$, which is significantly smaller than in DfE.

Panel B of Figure 6 plots median estimates of the LLO probability weighting function estimated based on the DfE and DfE+forced data. Clearly the inverse probability weighting from DfE disappears altogether, replaced with a mildly convex function indicating consistent risk-taking across probabilities. This suggests that forced sampling eliminates (or at least severely reduces) sampling and inference bias in DfE. But comparison with estimates from DfD (Panel A in figure 1) shows that this does not close the GAP. Indeed, a meta-analysis of the gap between DfE+forced and DfD reveals that the GAP remains substantial as well as statistically significant, with a meta-analytic mean of 8.4 pp and a CrI of [0.053 , 0.115] pp. A similar failure of forced sampling to close the GAP has been documented in prior work (Ungemach et al. 2009, Cubitt et al. 2022).[22] This leaves an important question: why doesn't eliminating sampling and inference bias via forced sampling lead behavior in DfE to converge to the standard probability weighting observed in DfD experiments?

[22] Wulff et al. (2018) use an alternative strategy whereby they only use observations in which the samples in DfE correctly reflect the underlying probability to analyze the GAP. They conclude that the GAP remains even using just those samples, consistent with what we find here. For a thorough discussion of attempts to eliminate sampling bias in DfE by various means, see their discussion on pp. 151–152.

4 Closing the Gap: Removing Bias from DfD

Forced sampling leads to dramatic changes to DfE behavior, but it does not close the GAP. Instead of pushing choices to standard probability weighting, forced sampling causes behavior to converge to behavior that resembles standard expected utility behavior with moderate levels of risk aversion. The shift suggests that sampling and inference bias indeed distort DfE behavior, but it also suggests that these biases aren't fully responsible for the GAP.

The key hypothesis in our paper is that the reason forced sampling in DfE doesn't remove the GAP is because very similar biases also afflict DfD. Probability weighting itself, we hypothesize, is a consequence of inference bias (one of the two biases affecting DfE) distorting decision making in DfD. Because of this, in order to fully close the GAP, we have to remove biases in DfD in a manner symmetric to the way we removed biases in DfE.

A long line of research in neuroscience suggests that perception and mental representation itself depends on a kind of *neuronal sampling* akin to the explicit sampling in DfE. This is a consequence of the fact that precisely representing quantities like probabilities requires the use of a large number of neurons and therefore comes at great cost to the brain. Gold & Shadlen (2001) and Gold & Shadlen (2002) for instance forcefully argue that it is efficient for the brain to summarize evidence for or against an uncertain hypothesis using a neuron (or population of neurons) in favor of the hypothesis and an "anti-neuron" (or population of anti-neurons) summarizing the evidence against. Because efficient means of representation are unavoidably finite, the brain tends to represent even precisely described data imprecisely.[23] The idea that such imprecise representations of probabilities, payoffs and other quantities produces inference bias has been formalized in recent years as *noisy coding models* (Natenzon 2019, Khaw et al. 2021, 2023, Vieider 2024*b*).

To model this for the simple binary lotteries in our experiments in a way that emphasizes finite neuronal sampling and therefore facilitates comparisons to our DfE model, assume that the intensity of beliefs is coded as a count of spikes or neuronal firings ("action potentials") in the neuron(s) $\alpha_0 > 0$ in favor of the prize (e.g., the better paying outcome), and a count of spikes $\beta_0 > 0$ in the anti-neuron(s) in favor of the opposite (where the subscript

[23] In particular, any perception can potentially be affected by some noise due to the finite mental resources at our brain's disposal, and the need to encode a multiplicity of complex stimuli using just a series of "spikes"—the electrical firing rates or 'action potentials' emitted by neurons. It is thus important to understand how neurons can efficiently encode probabilistic information to represent beliefs.

0 simply serves to distinguish the parameters from those used above to designate literal samples in DfE).[24] We can think of the spike counts described by α_0 and β_0 as direct analogues to samples in DfE – a virtual sampling process ("virtual draws") that produces evidence for the winning versus losing outcomes in our binary lottery. Because neuronal resources allocated to the encoding of beliefs (the number of neurons and their spike counts) are necessarily finite, α_0 and β_0 will be finite, producing immediate imprecision in beliefs (see e.g. Heng et al. 2020 for a stylized model of neural activation along these lines) that mirror the imprecision in finite samples in DfE. Indeed, with finite neuronal resources, this sampling process will again be a Beta distribution with parameters α_0 and β_0, producing a representation of the probability p of earning x (e.g. in lottery R described above) equal to $\mathbb{E}[\,\widehat{p}\,|\,p\,] = \frac{\alpha_0}{\alpha_0+\beta_0}$ (i.e.the neurally coded mean probability $\widehat{p}$, given the true probability p).

Because of this, all of our main conclusions regarding inference bias hold in DfD just as in DfE: finite neuronal sampling leads to sampling noise parameterized by ν^2 (which here we will call "coding noise"), producing the same inference bias described in (5). The key difference in DfD relative to DfE is that there is no longer scope for sampling bias. Because the DM is directly told the true probability, it is natural to assume $\mathbb{E}[\,\widehat{p}\,|\,p\,] = \frac{\alpha_0}{\alpha_0+\beta_0} = p$. We can thus directly use equation (6) and simply substitute p for $\widehat{p}$, because the perceived probability will be correct *on average* (see online appendix A for a step-by-step derivation and discussion).

Even though the inferred probability will be correct *on average*, the finite neuronal machinery means that the Beta distribution will entail some uncertainty around this unbiased mean.[25] Given this uncertainty, a Bayesian agent will continue to suffer from inference bias in DfD, and her posterior beliefs will continue to be distorted by her prior as in (5). Re-examining the first line of that expression, we see that it is simply equal to the popular linear in log odds (LLO) probability weighting function (equation (1) discussed and estimated in Section 2), with $\delta = \left(\frac{p_0}{1-p_0}\right)^{1-\gamma}$. Thus, noisy neuronal sampling itself serves as an explanation for the probability weighting widely observed in DfD.[26]

[24] The efficiency of this neuronal architecture lies in the fact that the balance between neurons and anti-neurons is robust to fluctuations in general neural activation over time (Gold & Shadlen 2001, 2002).

[25] Given encoding by parameters α_0 and β_0, infinite precision can only be achieved in the limit as α_0 and β_0 jointly approach infinity. In that limit, the Beta will indeed approach a Dirac-Delta distribution having all it probability density in a single point. Given that firing rates or spike rates of neurons are limited by physiological factors and that a DM's availability of neurons for a given decision task is also necessarily limited, coding noise is an inevitable feature of neuronal representations and calculations.

[26] In DfD, the DM's decisions are governed by the same discriminability expression described for DfE (expression (6)), however it influences behavior in a different way since the DM does not make explicit sampling decisions. In DfD, equation (6) instead simply quantifies the *choice probability* between the two

Closing the Gap. If it is true that probability weighting in DfD is a result of the same inference bias afflicting DfE, why doesn't the removal of sampling bias in our DfE+forced treatment lead to a closure of the GAP? Our model provides a crisp explanation: by forcing an increase in the samples in DfE (i.e., increasing α and β) and ensuring that those samples are balanced, we *simultaneously* remove both sampling and inference bias. Because inference bias is removed simultaneously with sampling bias, DfE behavior never converges to probability weighting (which, we hypothesize, is a consequence of sampling noise via inference bias). This explanation for the failure to close the GAP is therefore firmly rooted in the hypothesis that probability weighting is a consequence of inference bias: it is the fact that probability weighting is driven by inference bias that prevents it from appearing in DfE as samples grow large.

This explanation suggests both a method for closing the GAP, and a distinctive test for the hypothesis that inference bias is responsible for probability weighting in classic DfD experiments. By forcing subjects to sample *fully redundant information* in the DfD treatment in a manner symmetric to forced sampling in DfE, we predict that (i) probability weighting will disappear (or reduce in severity) in DfD and (ii) the GAP between description and experience will shrink or even close.

In our model, in which probabilistic information is coded in the brain as virtual samples, observed samples will simply be added to the 'virtual draws' formed mentally on the basis of the described distribution. Defining α_0, β_0 as the virtual draws formed initially on the basis of the described probability, we obtain the following updating equations after T samples:

$$\begin{aligned} \alpha_T &= \alpha_0 + \sum_{t=1}^{T} s_t \\ \beta_T &= \beta_0 + \sum_{t=1}^{T} (1 - s_t), \end{aligned} \tag{7}$$

where $s_t = \{0, 1\}$ designates the sample obtained in draw t, which takes the value 1 whenever a prize x is drawn, and the value 0 whenever the lower outcome y is drawn.[27] Sampled

choice options: it predicts that the DM (stochastically) chooses the risky option as a function of the positive value of ψ and the safe option in function of its negative value. The expression ψ follows a standard normal distribution, so that its standard normal cumulative distribution function is a Probit choice probability. Specifically, the choice probability in (6) constitutes a Probit link function, which is entered into a Bernoulli distribution to map it into Binary choice outcomes taking the values of 1 (choice of the risky option) or 0 (choice of the safe option).

[27] In practice, s may well take values that are different from (precisely) 1 and (precisely) 0, depending on the precise nature of neuronal activation in response to sampling.

successes and failures will thus simply be added to the virtual draws from the initial representation.

Given that in DfD the initial representations based on the description are correct on average, adding balanced draws (draws representative of the true distribution) will again result in a correct representation, but will result in a reduction in coding noise and therefore inference bias, reducing or eliminating probability weighting. This is a distinctive implication of noisy coding explanations for probability weighting that seems inconsistent with other classical explanations.[28] Since, as we saw in Experiment 2, there is little residual evidence of either inference or sampling bias in DfE after forced sampling, we predict that this will also close the description-experience GAP.

4.1 Experiment 3: the DfD+Forced Sampling Treatment

In Experiment 3, we attempt to eliminate the decision-experience GAP by forcing DfD subjects to redundantly sample large, representative samples from each lottery. In DfD+forced we show subjects the same information about lotteries as we do in the DfD treatment (pictured in Figure 4), but we also provide subjects the sampling tools pictured in Figure 6 below the explicit description, and force subjects to draw 20 times from each just as in DfE+forced. Indeed, the DfD+forced treatment is identical to the DfE+forced treatment, except that lotteries are fully described to the subject prior to, during and after sampling.

4.2 Results

Panel A of Figure 7 shows the effect of forced sampling in DfD, by plotting average choice proportions for DfD+forced and (for comparison) DfD. As predicted, we find that forced sampling has *exactly the reverse effect* on DfD as on DfE. At small probabilities, we find a sizeable *decrease* in risk taking at most probabilities.[29] For large probabilities, on the other hand, risk taking *increases* with forced sampling in DfD. Thus, just as predicted by noisy coding models like ours, providing completely redundant information to subjects has

[28] We also emphasize that although the specifics of our model make this implication easier to describe and facilitates comparison between DfD and DfE, the prediction that redundant information like this should reduce probability weighting likely applies to any explanation for probability weighting rooted in residual noise in fully described probabilities (i.e., any "noisy coding" style explanation).

[29] The exception is $p = 0.15$. This is, however, in part caused by the aggregation across different values of sure payments, c. For this particular probability, the changes across different sure amounts go in opposite directions canceling each other out – see Online Appendix G for the plot broken down by values of c.

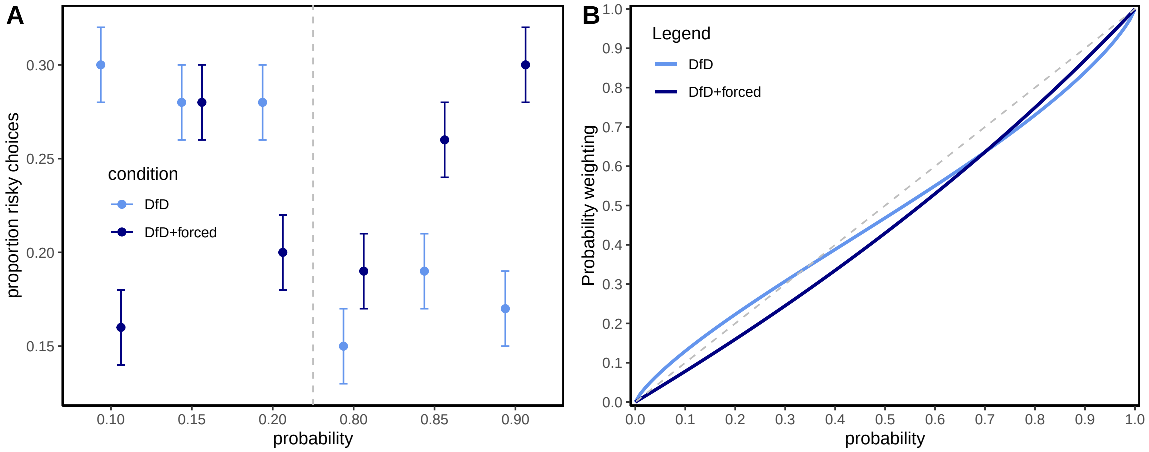


Figure 7: Effects of forced sampling on choice proportions

The figure shows the effects of forced complete sampling in description-based choice. Panel A shows the effect of forced sampling in DfD on choice proportions for different probabilities. The error bars indicate ± 1 standard error. Panel B shows the effect of forced sampling in DfD using estimates of the LLO probability weighting function. The figure is based on the median parameter values from a Bayesian hierarchical estimation, which are $\gamma = 0.80$ and $\delta = 0.88$ in DfD, and $\gamma = 0.99$ and $\delta = 0.75$ in DfD+forced.

a sizable effect on choices in DfD.

The asymmetric effect on low and high probabilities is exactly the effect we would expect if likelihood sensitivity in DfD (and probability weighting) was driven by inference bias rather than conventional preferences. As such, the treatment has the effect of eliminating standard probability weighting in DfD. As we showed in Section 2, choice proportions are negatively related to p in DfD in regressions. Estimating the same regressions on DfD+forced, this negative dependence of choice proportions on the probability of winning in the lottery vanishes in the DfD+forced treatment, with $\lambda = 0.058$, with a credible interval of $[-0.039\,,\, 0.157]$ indicating that likelihood-dependence is not significantly different from 0. Panel B of Figure 7 plots the LLO fit of the probability weighting function based on the median parameter values in DfD+forced and (for reference) for the median parameter values in standard DfD. The plot shows that standard DfD probability weighting *entirely* disappears after showing subjects redundant samples, strongly supporting our hypothesis that probability weighting in DfD is a consequence of inference bias due to coding noise.

Closing the GAP. What does the elimination of sampling and inference bias via forced sampling (in both DfE and DfD) do to the description-experience GAP? Figure 8 combines data from Figures 6 and 7 to answer this question. Panel A shows choice probabilities for DfE+forced and DfD+forced, revealing that the GAP has largely disappeared (in the two small-probability tasks in which a gap still seems to occur, it goes in opposite directions, so that the effects cancel each other out). Likewise, Panel B plots LLO functions based on median parameters in each treatment and shows that estimated behavior converges. Probability weighting and reverse probability weighting are replaced with modest, uniform risk aversion in both DfD and DfE.

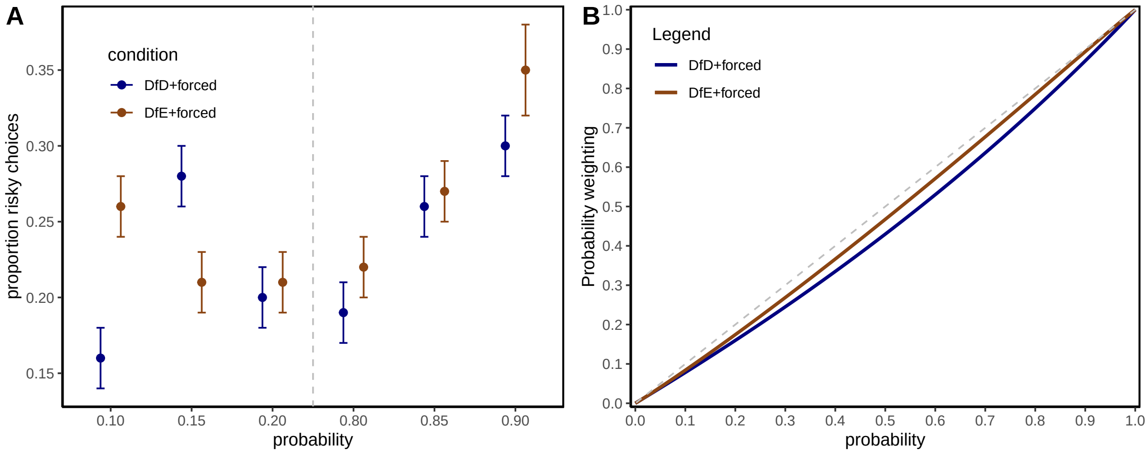


Figure 8: Comparison of behavior after forced sampling in DfE and DfD

The figure shows the effects of forced complete sampling. Panel A shows the effect of forced sampling in DfD and DfE on choice proportions for different probabilities. The error bars indicate ± 1 standard error. Panel B shows the effect of forced sampling in DfD and DfE using estimates of the LLO probability weighting function. The functions are based on the median parameter estimates from a Bayesian hierarchical model. The parameter values are $\gamma = 1.02$ and $\delta = 0.87$ in DfE+forced, and $\gamma = 0.99$ and $\delta = 0.75$ in DfD+forced.

To show this more systematically (and at the choice task level), we can meta-analytically aggregate the choice proportions across tasks. Panel A in figure 9 shows the original GAP between description-based choice and experience-based choice with free sampling. We again use a measure g capturing the difference in choice proportions, defined so that positive values correspond to behavior typically documented in the literature for the standard GAP—more risk taking in DfD than DfE for small probabilities, more risk taking in DfE than DfD for large probabilities.

Panel A shows that in the *absence* of forced sampling (DfE vs. DfD from Experiment 1), the GAP is significant in 12 out of 18 tasks when looking at the raw choice proportions, and in 13 out of 18 tasks in the meta-analytic posterior. The exceptions in which the GAP is not statistically significant at conventional levels are small probability tasks with $c \geq px$. Panel B compares description-based and experience-based choice proportions after forced sampling (DfE+forced vs. DfD+forced), and shows that the GAP disappears in these treatments. Indeed, we find no significant gap for *any* of the 18 choice proportions in the meta-analytic posterior. In the one case in which we see a significant gap in the raw

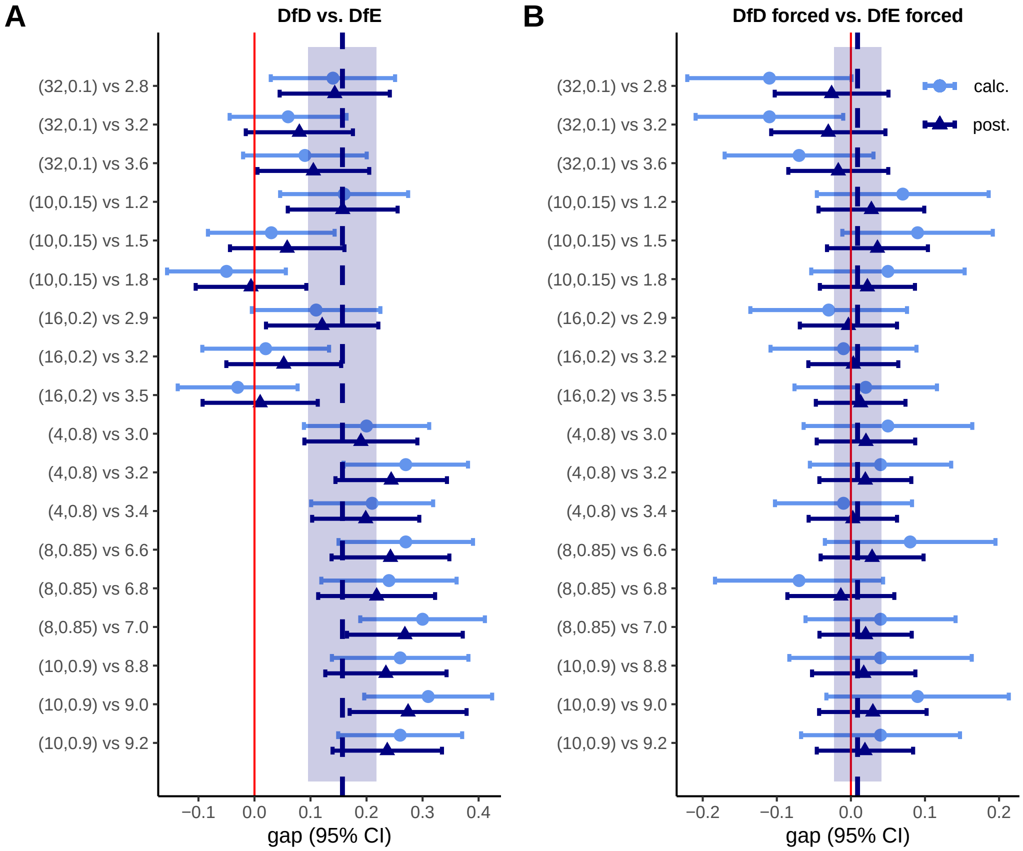


Figure 9: Meta-analysis of the GAP

Panel A shows a forest plot of the gap for our standard implementation of DfD versus DfE. Panel B shows a forest plot of the GAP after forced sampling both from description and from experience. The light blue circles, labeled 'calc.', indicate the raw differences in choice proportions in the data, g. The dark blue triangles, labeled 'post.', indicate the inferred posterior parameters, $\widehat{g}$. The thick, dashed vertical line indicates the meta-analytic posterior mean, ω, and the shaded rectangle indicates the 95% credible interval around that estimate.

choice proportions, this gap goes in the *opposite* direction of the standard GAP. At 0.9 pp (95% credible interval of $[-2.3\,,\,4.1]$ pp), the meta-analytic posterior mean is arbitrarily close to 0. The GAP has closed.

We can further examine the effect of forced sampling by restricting our attention to the 6 'standard tasks' in which the GAP is strongest to start with (and which represent the typical type of task used in the early DfE literature, with $px > c$). Figure 10 shows the meta-analytic evidence. Panel A shows the usual GAP between DfD and DfE. It is significant in all but one task in the raw data, and in all of them in the meta-analytic posterior. The meta-analytic mean indicates an average gap of 18.6 pp in these tasks. Panel B shows the same tasks for DfD with forced sampling versus DfE with forced sampling. None of the raw choice proportions is significantly different from 0, nor is any of the meta-analytic

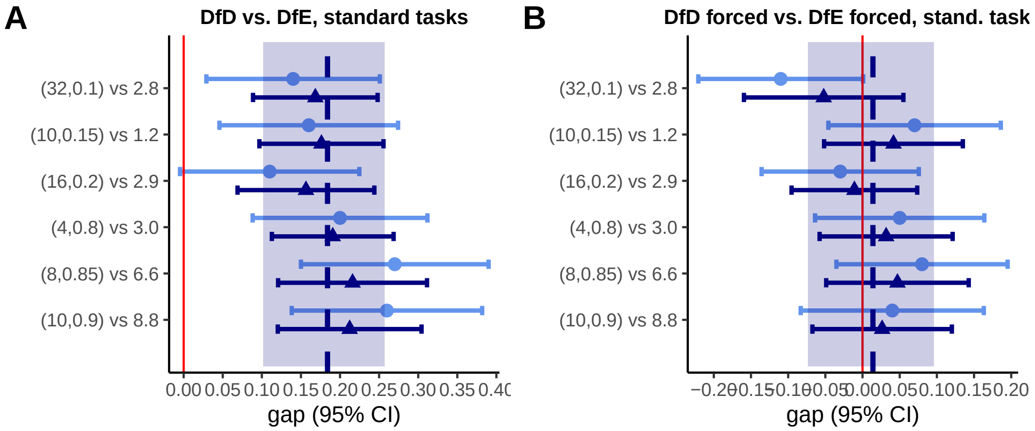


Figure 10: Meta-analysis of the GAP, 'standard tasks'

Panel A shows the gap for a standard implementation of DfD versus DfE. Panel B shows a forest plot of the GAP after forced sampling both from description and from experience. The light blue circles, labeled 'calc.', indicate the raw differences in choice proportions in the data, g. The dark blue triangles, labeled 'post.', indicate the inferred posterior parameters, $\widehat{g}$. The thick dashed line indicates the meta-analytic posterior mean, μ, and the shaded rectangle indicates the 95% credible interval around that estimate.

posterior means, $\widehat{g}$. What's more, the GAP falls from 18.6 to 1.4 percentage points, meaning the meta-analytic mean falls once more arbitrarily close to (and statistically insignificantly different from) 0. Once more, the GAP has closed.

Our results thus provide strong evidence that the decision-experience GAP is a consequence of the two biases we expect forced sampling to remove. First, inference bias creates classic probability weighting in DfD. Second, sampling bias (intensified by inference bias) creates underweighting of rare events, creating the inverse of probability weighting in DfE. Forced sampling removes each of these biases, generating inverse responses in DfE and DfD that drive behaviors in the two settings together, removing the GAP.

5 Structural Estimation: The Influence of Coding Noise

Finally, we use structural estimation to more deeply assess the hypothesis that both probability weighting and the description-experience gap are a consequence of cognitive noise – and that our treatments eliminate these patterns by eliminating this noise. We structurally estimate our model from choice data based on our discriminability expression (6). The key parameter driving both probability weighting and the GAP in our model (and, therefore, our focus in this section) is γ, the weight the DM puts on her perception of the log-odds in the decision process. We will refer to this as "likelihood-discriminability," mirroring the name given the equivalent parameter in the LLO function, "likelihood-sensitivity." In the model, γ is an inverse function of coding noise: the smaller coding noise ν becomes, the closer γ will come to 1, producing perfect discriminability of log-odds and an elimination of inference bias (and hence of probability weighting). Importantly, this parameter is estimated, in part, using inconsistencies in subjects' choices across repeated instances of the same task (recall, four random tasks were repeated for each subject) which give us direct, subject-level measures of *behavioral noise*. This analysis therefore relies on new data, not reported in the previous analysis.

We estimate the model using Bayesian hierarchical techniques, which optimally combine individual-level information with group-level evidence (Gelman & Hill 2006, Gelman et al. 2014). This allows us to study distributions of individual-level parameters based on relatively few decision tasks (details and code are provided in Online Appendix G). We normalize the variance of the prior to $\sigma = 1$ throughout, so that coding noise is measured relative to the variance of the prior, ν/σ. This is done without loss of generality and to improve comparability across studies, simply leading to a rescaling of the equation (see Natenzon 2019 for an equivalent simplification).[30] We execute tests on distributional differences and correlations in individual-level parameters based on the means of the individual-level posteriors throughout. All comparisons are within-subject, leveraging our two stage design, unless specified otherwise. We report four main findings:

First, we find that, conditional on the information subjects have about probabilities, estimates of γ indicate strong (and similar) levels of noisy coding and inference bias in DfD and DfE, with γ estimates well below the unbiased benchmark of 1. To estimate γ in a way that makes DfE and DfD estimates comparable, we estimate the model in DfE on the actually experienced probabilities (i.e., probabilities implied by the sample subjects have drawn), rather than the lottery's true probabilities.[31] Because of this, we must make an assumption on how subjects perceive the log cost-benefit ratio in cases in which the subject fails to sample both lottery outcomes before making a choice. Panel A in Figure 11 shows

[30] We estimate the model on choice data while leveraging our within-subject design. That is, we estimate the model using the data from both treatments, and assuming that the parameters governing the prior remain the same across the two treatments, while leaving the other model parameters free to vary. This allows us to maximize the informative content of our sparse choice stimuli. See Online Appendix G for details.

[31] We assume throughout that the initial Beta parameters, before any samples are observed, are $\alpha = \beta = 0.1$. This assumption derives from our general inference framework, based on a diffuse Dirichlet space – see Online Appendix A.1 for details. While values smaller than 1 are plausible (they imply that subjects expect relatively few outcomes in our general inference framework), our results are not sensitive to variations of this value within that range.

the cumulative distribution function of individual-level γ estimates under the assumption that DMs are "naive" in the sense that they judge costs and benefits to be equal in such cases. In panel B, we instead assume DMs are sophisticated in the sense that they realize that larger log-odds imply larger log cost-benefits; the correlation measuring the degree of sophistication thus must be estimated as an endogenous parameter (see Online Appendix G for details and additional results).

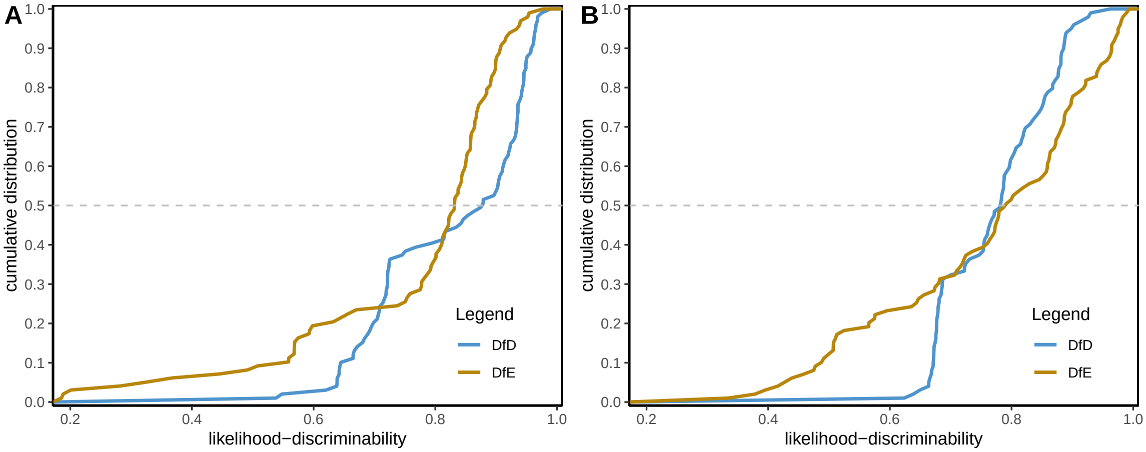


Figure 11: Structural estimates, DfD versus DfE

The figure shows structural estimates of the model parameters. Panel A compares likelihood-discriminability γ in DfD and DfE for a naive decision maker, who assumes costs and benefits to be equal when one of the outcomes has not been observed. Panel B compares likelihood-discriminability, γ, for a sophisticated DM, who (correctly) infers that log-odds and log costs-benefits are correlated in the choice problems. The correlation coefficient is thereby estimated endogenously from the data (see Online Appendix G for details).

Regardless of the approach taken, two findings stand out from Figure 11. First, in both DfD and DfE, γ falls well below the unbiased benchmark of 1, suggesting a strong role for inference bias in both settings as predicted by our model. Second, the distributions of γ estimates are similar in both DfD and DfE.[32] This is important because our model explains the GAP between these settings not via differences in γ but rather via the very different effects the model predicts γ has in DfD vs. DfE environments. The results therefore assure us that the model parsimoniously explains differences in lottery choices across treatments, conditional on the information available to subjects.

Second, we show that forced sampling in DfD and DfE results in a sharp increase in γ towards 1 (the unbiased benchmark), suggesting that the intervention influences behavior (as predicted by the model) by severely reducing coding noise and with it scope for inference

[32] For the naive estimates pictured in panel A, likelihood-discriminability γ is somewhat smaller in DfE than in DfD ($p = 0.006$). For the sophisticated estimates in panel B, the two distributions produce roughly equal deviations above and below 0.5, and are not significantly different ($p = 0.979$).

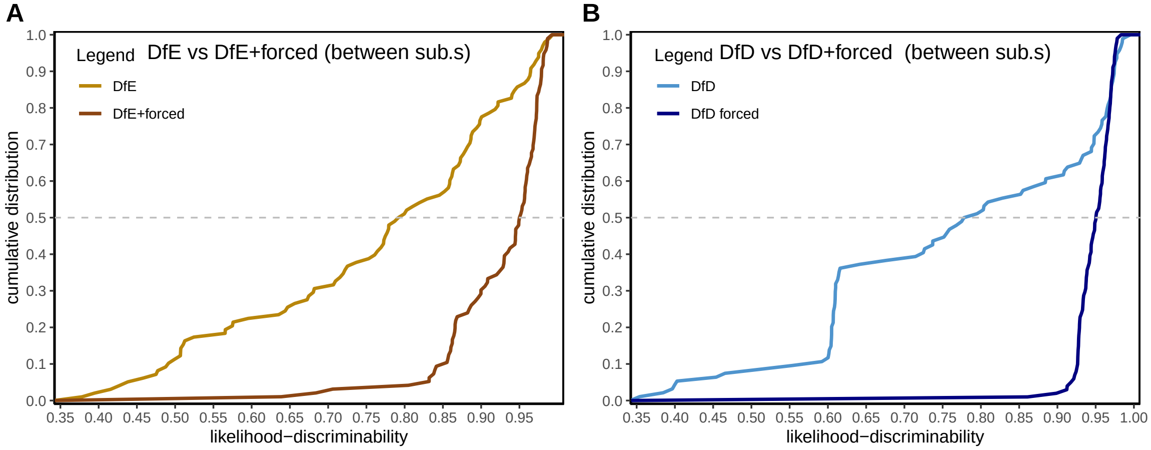


Figure 12: Structural estimates, DfD vs DfD+forced and DfE vs DfE+forced

The figure shows structural estimates of likelihood-discriminability γ. Panel A compares likelihood-discriminability in DfE and DfE+forced. Panel B compares likelihood-discriminability in DfD and DfD+forced.

bias. In panel A and B of Figure 12 respectively we plot CDFs of estimated individual-level mean γ estimates in DfE[33] and DfD with and without forced sampling.[34] In both cases, forced sampling causes a sharp rightward shift in the γ parameter, with medians in both cases of about 0.95 suggesting a near elimination of coding noise and inference bias.[35]

Third, we show that forced sampling in DfD and DfE – which, recall, caused a convergence in behavior between the two treatments – also causes a convergence in γ. This suggests (as our model predicts) a causal linkage between the two findings: joint convergence of γ in the two treatments towards 1 (signalling the disappearance of inference bias) causes lottery choice patterns to converge, suggesting (as predicted by the model) that coding noise was responsible for their initial divergence. Panel A of Figure 13 directly compares γ in DfD+forced and DfE+forced. Over most of the distribution, the panel shows that discriminability converges across the two treatments, suggesting that subjects are similarly free of inference bias in the two settings – a finding that matches the similar revealed risk aversion in choices in the two settings. Indeed, non-parameteric tests detect no significant difference between the two distributions ($p = 0.376$).[36]

[33] In DfE we plot estimates that assume subjects make sophisticated inferences about the cost-benefit ratio, as discussed above.

[34] For this analysis, we use a between-subject comparison in both cases since DfE vs DfE+forced can only be compared between subjects; in DfD, replacing this with within-subject comparisons yields very similar results (cfr. Online Appendix G).

[35] Estimates also reveal a sharp reduction in cross-subject variance. This too is a prediction of the model, since the treatment is predicted to have similar impacts on both initially high and low noise subjects.

[36] Nonetheless, as is clear from the graph, discriminability is somewhat lower in the left hand tail of the DfE distribution. We hypothesize that this is due to limitations on subjects' memory, highlighting the value

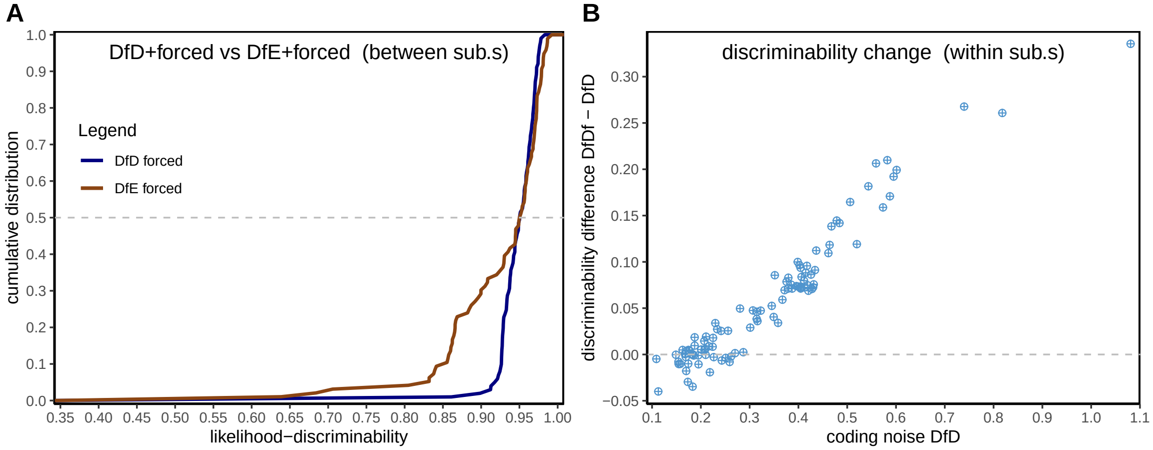


Figure 13: Effects of forced sampling, Structural estimates

The figure shows structural estimates of the model parameters. Panel A directly compares likelihood-discriminability γ in DfD+forced and DfE+forced. Panel B compares likelihood-discriminability, γ, in DfD without and with forced sampling. Panel B plots coding noise in first stage DfD against the change in likelihood-disciminability when forced sampling is introduced.

Finally, Panel B of Figure 13 illustrates the reason for this effect by plotting coding noise ν (measured in the DfD choices in Stage 1 of the experiment) against the difference between γ in DfD and DfD+forced (defined as $\gamma_2 - \gamma_1$, with subscripts indicating the stage of the experiment), exploiting our within-subject design. The figure shows clearly that the effect of sampling is most pronounced for those subjects who had the largest coding noise to begin with. These results strongly support an additional prediction of the model: that sampling should have the strongest effect on subjects who have relatively high coding noise to start with (i.e., relatively small 'spike counts' α_0 and β_0). This is a consequence of the fact that the reduction in coding noise decreases at a decreasing rate with further samples (that is, the smaller α_0 and β_0, the larger the effect of adding actual samples based on equation 7). The figure thus shows in a particularly sharp way how strong the effect of forced sampling is on likelihood-discriminability in the DfD treatment.

6 Robustness: Voluntary Sampling in DfD

Finally, we report a fourth experiment designed to test two additional implications of our model, providing a robustness check for our explanation for probability weighting and the GAP. First, noisy coding explanations for probability weighting predict that subjects re-

to subjects of having an explicit description of the outcomes and probabilities on the screen (in DfD+forced) to guard against inattention and working memory limitations.

main, in an important sense, uncertain about the properties of fully described lotteries. Because of this, subjects should not only be *responsive* to additional information – they should also *demand* such information to make their beliefs more precise. Second, our model suggests a second way to close the GAP beyond the one pursued in Experiments 2 and 3. Instead of removing the GAP by removing inference bias (in DfE and DfD) and sampling bias (in DfE) via forced sampling, we can instead *introduce* the same kind of sampling bias afflicting DfE to DfD by allowing subjects in DfD to (if they so choose) under-sample their options.

We test these implications using our DfD+free treatment. This treatment simply combines the sources of information from our DfD and DfE treatments: subjects in DfD+free are explicitly told the properties of their lottery options (as in DfD) but can (if they choose) sample from each of these lotteries randomly as much (or as little) as they like (as in DfE). We ran this experiment on Prolific with 101 subjects in September 2023 using the exactly same parameters and structure used in our other treatments. We provide details of the results in Online Appendix C and here highlight two main effects of this treatment.

First, we find strong evidence that subjects do in fact voluntarily seek out redundant information on the properties of fully-described lotteries. This seems to be particularly strong evidence that subjects (as noisy coding models like ours predict) are residually uncertain about the properties of their choice options. The average subject takes just under two samples, on average, but takes more than 4 in early rounds and some subjects take as many as 9 on average. More than 90% of subjects sample at least once in this treatment. However, as in DfE, the sampling subjects conduct is far too modest to generate anything other than highly biased information.

Second, and as a result of this, we find that this sampling causes DfD+free behavior to converge sharply towards DfE behavior, closing the GAP in a second way. Free sampling in DfD+free produces a severe change in behavior relative to DfD, causing a complete reversal in the direction of likelihood dependence. In DfD+free subjects take *even more* risk at higher probabilities than in DfD+forced, thus converging towards behavior in DfE. Probability weighting thus reverses in DfD+free, suggesting that the inference bias suffered by subjects in DfD becomes overwhelmed by sampling bias in DfD+free. There is some apparent difference between DfD+free and DfE at high probabilities (and likelihood-dependence in DfD+free, albeit positive, is significantly weaker than in DfE), suggesting that subjects' beliefs continue to be influenced by the explicit provision of descriptive information. However, we cannot meta-analytically reject the hypotheses that simply giving subjects the option to sample in DfD is sufficient to eliminate the description-experience GAP (details in Online Appendix C).

This treatment tests especially distinctive implications of our explanation for probability weighting and the GAP. After all, it seems highly unlikely under any explanation other than noisy coding that subjects would voluntarily seek out additional information in the DfD treatment. The fact that they do seems to directly suggest that subjects are in some important sense uncertain of the information explicitly given to them. Likewise, it is hard to explain why subjects provided unbiased initial information should be susceptible to bias from under-sampling, unless those subjects have a noisy understanding of what that initial information represents. Because of this, we view these results as strong additional support for our hypothesis that probability weighting and the GAP are rooted in cognitive imprecision.

7 Discussion

In this paper we show that probability weighting and the description-experience gap – two key phenomena in the lottery choice literature – are a consequence of the incomplete and imprecise ways decision makers perceive and represent information. Reducing the imprecision of subjects' beliefs by forcing them to observe redundant information causes probability weighting to disappear and closes the description-experience gap. In addition to shedding significant light on a key mystery in the literature, we believe there are two broader implications of our findings.

First, our results provide some of the most direct (and therefore strongest) evidence to date of "noisy coding" explanations for non-standard behaviors like probability weighting (Zhang et al. 2020, Khaw et al. 2023, Frydman & Jin 2023, Vieider 2024*b*). Noisy coding models hypothesize that descriptive failures of benchmark models like expected utility theory (von Neumann & Morgenstern 1944, Savage 1954) and exponential discounting (Samuelson 1937) are a consequence, not of non-standard preferences, but rather of what we have called inference bias, driven by limitations in the way the brain encodes information. In particular these models hypothesize that decision making is subject to the same kinds of Bayesian distortions that have been shown for decades by psychologists to shape perception: noisily processed valuations are systematically distorted by decision makers' prior beliefs, a form of noise-driven bias that can account for a number of behavioral anomalies. Our empirical approach is particularly direct because it relies on a direct manipulation of the representational noise that lies at the root of noisy coding models. By providing subjects with completely redundant information, we are able to severely reduce this noise and thereby cause probability weighting to disappear – a treatment effect that is difficult to account for via alternative explanations that are not similarly rooted in cognitive imprecision.

Second, because our findings show that valuations are fundamentally shaped by cognitive frictions, they call into question the common interpretation of behavioral anomalies like probability weighting (or its reversal in decision from experience) as expressions of subjects' welfare-relevant *preferences* for risk. When we eliminate noise in subjects' beliefs (in DfD) and remove scope for sampling bias (in DfE), we find that probability weighting and its reversal both disappear. What they are replaced with is strikingly neoclassical: subjects in both DfD and DfE show no evidence of likelihood dependence and instead show evidence of modest risk aversion that is broadly consistent with expected utility theory. Our results therefore suggest that we should be cautious in interpreting anomalous behavior in domains like lottery choice as rejections of standard models of preferences like expected utility theory. Our results suggest that even the simplest choice problems are powerfully shaped by limitations in human cognition like those expressed by noisy coding models. This has obvious implications for our normative interpretations of anomolous behaviors like probability weighting, and for the policies we design in response to them.

References

Abdellaoui, M. (2000), 'Parameter-Free Elicitation of Utility and Probability Weighting Functions', *Management Science* **46**(11), 1497–1512.

Abdellaoui, M., L'Haridon, O. & Paraschiv, C. (2011), 'Experienced versus Described Uncertainty: Do we Need Two Prospect Theory Specifications?', *Management Science* **57**(10), 1879–1895.

Atchison, J. & Shen, S. M. (1980), 'Logistic-normal distributions: Some properties and uses', *Biometrika* **67**(2), 261–272.

Aydogan, I. (2021), 'Prior beliefs and ambiguity attitudes in decision from experience', *Management Science* **67**(11), 6934–6945.

Aydogan, I. & Gao, Y. (2020), 'Experience and rationality under risk: re-examining the impact of sampling experience', *Experimental economics* **23**(4), 1100–1128.

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.

Bohren, J. A., Hascher, J., Imas, A., Ungeheuer, M. & Weber, M. (2024), A cognitive foundation for perceiving uncertainty, Technical report, National Bureau of Economic Research.

Bouchouicha, R., Oprea, R., Vieider, F. M. & Wu, J. (2023), Choice lists and 'standard patterns' of risk taking, Technical report, Ghent University Discussion Papers.

Bruhin, A., Fehr-Duda, H. & Epper, T. (2010), 'Risk and Rationality: Uncovering Heterogeneity in Probability Distortion', *Econometrica* **78**(4), 1375–1412.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P. & Riddell, A. (2017), 'Stan: A probabilistic programming language', *Journal of Statistical Software* **76**(1), 1–32.

Cubitt, R., Kopsacheilis, O. & Starmer, C. (2022), 'An inquiry into the nature and causes of the description-experience gap', *Journal of Risk and Uncertainty* pp. 1–33.

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

Enke, B. & Graeber, T. (2023), 'Cognitive uncertainty', *Quarterly Journal of Economics* .

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.

Frydman, C. & Jin, L. J. (2022), 'Efficient coding and risky choice', *Quarterly Journal of Economics* **136**, 161–213.

Frydman, C. & Jin, L. J. (2023), On the source and instability of probability weighting, Technical report, National Bureau of Economic Research.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2014), *Bayesian data analysis*, Vol. 2, CRC press Boca Raton, FL.

Gelman, A. & Hill, J. (2006), *Data analysis using regression and multilevel/hierarchical models*, Cambridge university press.

Glanzer, M., Hilford, A., Kim, K. & Maloney, L. T. (2019), 'Generality of likelihood ratio decisions', *Cognition* **191**, 103931.

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.

Gold, J. I. & Shadlen, M. N. (2001), 'Neural computations that underlie decisions about sensory stimuli', *Trends in cognitive sciences* **5**(1), 10–16.

Gold, J. I. & Shadlen, M. N. (2002), 'Banburismus and the brain: decoding the relationship between sensory stimuli, decisions, and reward', *Neuron* **36**(2), 299–308.

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

Green, D. M., Swets, J. A. et al. (1966), *Signal detection theory and psychophysics*, Vol. 1, Wiley New York.

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.

Heng, J. A., Woodford, M. & Polania, R. (2020), 'Efficient sampling and noisy decisions', *Elife* **9**, e54962.

Herold, F. & Netzer, N. (2023), 'Second-best probability weighting', *Games and Economic Behavior* **138**, 112–125.

Hershey, J. C., Kunreuther, H. C. & Schoemaker, P. J. H. (1982), 'Sources of Bias in Assessment Procedures for Utility Functions', *Management Science* **28**(8), 936–954.

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.

Hertwig, R. & Erev, I. (2009), 'The description–experience gap in risky choice', *Trends in cognitive sciences* **13**(12), 517–523.

Hertwig, R. & Pleskac, T. J. (2010), 'Decisions from experience: Why small samples?', *Cognition* **115**(2), 225–237.

Kahneman, D. & Tversky, A. (1979), 'Prospect Theory: An Analysis of Decision under Risk', *Econometrica* **47**(2), 263 – 291.

Khaw, M. W., Li, Z. & Woodford, M. (2021), 'Cognitive imprecision and small-stakes risk aversion', *The Review of Economic Studies* **88**(4), 1979–2013.

Khaw, M. W., Li, Z. & Woodford, M. (2023), Cognitive imprecision and stake-dependent risk attitudes, Technical report.

L'Haridon, O. & Vieider, F. M. (2019), 'All over the map: A worldwide comparison of risk preferences', *Quantitative Economics* **10**, 185–215.

Ma, W. J., Kording, K. P. & Goldreich, D. (2023), *Bayesian Models of Perception and Action: An Introduction*, MIT press.

Natenzon, P. (2019), 'Random choice and learning', *Journal of Political Economy* **127**(1), 419–457.

Netzer, N. (2009), 'Evolution of time preferences and attitudes toward risk', *American Economic Review* **99**(3), 937–55.

Netzer, N., Robson, A., Steiner, J. & Kocourek, P. (2021), Endogenous risk attitudes, Working paper.

Olschewski, S. & Scheibehenne, B. (2024), 'What's in a sample? epistemic uncertainty and metacognitive awareness in risk taking', *Cognitive Psychology* **149**, 101642.

Oprea, R. (2022), 'Simplicity equivalents', *Working Paper* .

Prelec, D. (1998), 'The Probability Weighting Function', *Econometrica* **66**, 497–527.

Preston, M. G. & Baratta, P. (1948), 'An Experimental Study of the Auction-Value of an Uncertain Outcome', *The American Journal of Psychology* **61**(2), 183.

Robson, A. J. (2001*a*), 'The biological basis of economic behavior', *Journal of Economic Literature* **39**(1), 11–33.

Robson, A. J. (2001*b*), 'Why would nature give individuals utility functions?', *Journal of Political Economy* **109**(4), 900–914.

Robson, A. J. & Samuelson, L. (2011), The evolutionary foundations of preferences, *in* 'Handbook of social economics', Vol. 1, Elsevier, pp. 221–310.

Samuelson, P. A. (1937), 'A Note on Measurement of Utility', *The Review of Economic Studies* **4**(2), 155–161.

Savage, L. J. (1954), *The Foundations of Statistics*, Wiley, New York.

Simon, H. A. (1959), 'Theories of decision-making in economics and behavioral science', *The American Economic Review* **49**(3), 253–283.

Steiner, J. & Stewart, C. (2016), 'Perceiving prospects properly', *American Economic Review* **106**(7), 1601–31.

Tversky, A. & Kahneman, D. (1992), 'Advances in Prospect Theory: Cumulative Representation of Uncertainty', *Journal of Risk and Uncertainty* **5**, 297–323.

Tversky, A. & Wakker, P. P. (1995), 'Risk Attitudes and Decision Weights', *Econometrica* **63**(6), 1255–1280.

Ungemach, C., Chater, N. & Stewart, N. (2009), 'Are probabilities overweighted or underweighted when rare outcomes are experienced (rarely)?', *Psychological Science* **20**(4), 473–479.

Vieider, F. M. (2024*a*), Bayesian estimation of decision models, Technical report.
URL: *https://fvieider.quarto.pub/bstats/*

Vieider, F. M. (2024*b*), 'Decisions under uncertainty as bayesian inference on choice options', *Management Science, forthcoming* .

von Neumann, J. & Morgenstern, O. (1944), *Theory of Games and Economic Behavior*, Princeton University Press, New Heaven.

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

Wu, G. & Gonzalez, R. (1996), 'Curvature of the Probability Weighting Function', *Management Science* **42**(12), 1676–1690.

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

Zhang, H. & Maloney, L. T. (2012), 'Ubiquitous log odds: a common representation of probability and frequency distortion in perception, action, and cognition', *Frontiers in neuroscience* **6**, 1.

Zhang, H., Ren, X. & Maloney, L. T. (2020), 'The bounded rationality of probability distortion', *Proceedings of the National Academy of Sciences* **117**(36), 22024–22034.

Online Appendices

A Model derivation

A.1 A general inference model

In Decisions from Experience (DfE), subjects need not only learn the outcomes and underlying probabilities, but also the whole structure of the decision problem (i.e., the number of outcomes in the lottery's support). In the body of the paper we assume away this component of the inference problem for simplicity and to focus our discussion on the influence of sampling and inference bias. Here, for completeness, we propose a stylized model of how such higher order learning could take place based on the sort of sampling from the two options that occurs in DfE. We argue that expanding the model in this way has little qualitative impact on our findings.

We start by discussing the inference process. Assume a DM believes that outcomes will range from 0 to some upper limit u, outcomes beyond which are not considered plausible.[37] Take two objective probability distributions over all outcomes underlying the two choice options, $\{p_0, p_1, ..., p_u\}$ and $\{q_0, q_1, ..., q_u\}$, where subscripts indicate monetary outcomes. In DfE, DMs will infer the probability distributions from the draws they observe. Let the initial likelihood at time $t = 0$, before any draws are taken, be encoded in two $u + 1$-dimensional Dirichlet distributions, $\mathcal{D}_A(\pi_j) \propto \prod_{j=0}^{u} \widetilde{p}_j^{\;\pi_j-1}$ and $\mathcal{D}_B(\omega_j) \propto \prod_{j=0}^{u} \widetilde{q}_j^{\;\omega_j-1}$, where $\widetilde{p}_i \triangleq \frac{\pi_i}{\sum_j \pi_j}$ and $\widetilde{q}_i \triangleq \frac{\omega_i}{\sum_j \omega_j}$ represent the subjective expectations of the probabilities attributed to an outcome i in the two choice options A and B. Given the ex ante exchangeability of the two choice options, the two Dirichlets will have the same parameters at time $t = 0$. We assume that DMs consider any given outcome as equally likely in the two choice options, so that $\pi_i = \omega_i \; \forall \; i$ at $t = 0$. This assumption directly follows from the exchangeability of the two options before any draws have been observed, and is implemented in our experiment by randomizing the risky and safe options in positions A and B.

We assume that what matters for decisions is the direct comparison between the two choice options. To capture this in our model, we map the inferences based on the Dirichlets encoding draws from the two choice options into a *comparative Dirichlet* which entails a statewise comparison between to two options. That is, what matters for choices are events in which one option pays a given outcome, while the other option pays a different outcome. In our experiment, these will be the events under which the risky option pays x while the

[37] In principle, u can take any value, as long as it is finite. In our experiment, we tell subjects beforehand that all outcomes will range between \$0 and \$ 35 inclusive, thus setting their expectations about this range.

safe option pays $c < x$, and the event under which the risky option pays y while the safe option pays $c > y$ (see below for a generalization). The probabilities of the comparative events e_1 (obtain $x > c$ rather than c) and e_2 (obtain c rather than $y < c$) can now be obtained from the single-state Dirichlets $\mathcal{D}_A(\pi_j)$ and $\mathcal{D}_B(\omega_j)$ defined for the two options, since $P[e_1] = P[x \cap c] = \widetilde{p}_x \times \widetilde{p}_c$ and $P[e_2] = P[y \cap c] = \widetilde{p}_y \times \widetilde{p}_c$. Given that for finite samples $\widetilde{p}_c < 1$ and $\widetilde{p}_x + \widetilde{p}_y < 1$, the inferred probabilities will generally be subadditive, that is, $P[e_1] + P[e_2] \leq 1$ (with 1 being the limiting case as samples tend to infinity). This implies that we can express the subjective beliefs in the comparative states of the world once again by a Dirichlet, $\mathcal{D}(\delta_i) = \prod_{i=1}^{u} P[e_i]^{\lambda\widehat{\delta}_i - 1}$, where $\lambda \triangleq \sum_{i=1}^{u} \delta_i$ is the concentration of the new Dirichlet, and $\widehat{\delta}_i \triangleq {}^{\delta_i}/_{\lambda}$ captures the mean belief about a given state i. While some probability mass will thus remain attributed to 'non-observed outcomes', this part will drop out of the main choice equation below.

This justifies the assumption of the Beta distribution in the main text: while the latter imposes additivity in $\widehat{p}_x$ and $\widehat{p}_y$, that assumption serves to simplify our discussion, but has no substantive implications for our conclusions (given that the non-observed states receiving the remaining probability mass drop out of the discriminability equation). If, say, a third outcome from the risky option were to be observed at some point, this would add a new comparative state to the comparison (see below). In the text we further discussed inference bias in terms of the samples taken from the risky option only. More generally, however, the samples from the safe option will also count. While a precise closed-form solution does not exist for that case, we can approximate the samples by the total samples for each state, where the samples from the safe option are simply added to the samples indicating each comparative sample in the sum of the trigamma function. This means that our discussion in the main text may *quantitatively* underestimate the samples, but that this more general case will not qualitatively affect any of the conclusions drawn.

In the main text, we implicitly assume that subjects know which of the two options is the risky one and which the safe. In reality, subjects need to infer this from the samples they take. We make three assumptions in this regard. The first, and most substantively relevant, is that subjects make inferences on the choice environment (including potentially the intentions of the experimenter). This entails that choices between two non-degenerate options are deemed extremely unlikely. Practically, this entails that noise will remain high until a plausible set of outcomes has been observed (Figure 14 below illustrates this for our experimental stimuli).[38] The second assumption is that we assume the initial parameters

[38] This assumption seems particularly defensible in our DfE experiments, since all subjects assigned to this

of the two choice option Dirichlets to be sparse, i.e. $\pi_i, \omega_i \ll 1 \; \forall \; i$. This assumption implies that subjects do not expect a very diffuse probability distribution with many different outcomes. Practically, this helps explain why samples are relatively small, since it keeps the probability mass assigned to unobserved outcomes low in the comparative Dirichlet.

An additional assumption in the main text is that subjects can infer which of the two options is the risky one. This obtains trivially once a subject has observed all three outcomes used in our experiment (the two in the risky option, and the one in the safe option, which constitute a 'plausible minimal outcome set' inasmuch as they indicate a non-degenerate choice, or equivalently, they map into two comparative states with a meaningful tradeoff between log-odds and log-cost benefits). This indeed follows directly from the two assumptions above: that subjects expect non-degenerate choices, and that the initial parameters are sparse (meaning that they do not necessarily expect more outcomes once they have observed a plausible outcome set). The inference is somewhat less trivial as long as only one outcome has been observed from each choice option.

We illustrate this based on the choice options we provide in the experiment. For small probabilities, subjects are overwhelmingly likely to observe the lower outcome y. Given that in our experiment y is always equal to 0, and that we tell subjects that they will only ever face non-negative amounts, this immediately identifies this choice option as the risky one. For large probabilities, where subjects may observe two strictly positive amounts from the two options, this is less obvious. We thus furthermore assume that the parameters of the option-specific Dirichlets before any samples are taken will be characterized by sparsity increasing in outcomes. That is, for any $j > i$, where the two indices are non-negative outcomes, $\omega_j = \pi_j \leq \pi_i = \omega_j$ at time $t = 0$, before any samples have been taken. In practice, this entails that subjects consider smaller outcomes more likely than larger outcomes. Notice that this is the equivalent of a pessimistic prior for the inference process, and that it is thus fully coherent with both our model and our empirical results.[39]

treatment have all finished making dozens of binary DfD choices for lotteries with one degenerate and one non-degenerate lottery.

[39] In principle, this inference process could be modelled as a probabilistic process resulting in stochastic assessments of the riskiness of the two choice options after each sample. Such a model would follow a very similar structure as our discriminability model, and we do thus not formalize it here. Such a model would be most relevant for large probability lotteries in cases where only one outcome has been observed from each option. The notion that subjects infer the structure of such choice problems from sampling draws is indeed supported by the observation that samples from the *safe* option increase in the objective probability of winning for both risk averse and risk seeking subjects in our data.

A.2 Noisy log-odds representation

In our actual experiment, subjects will experience exactly 1 outcome from the sure option, and no more than 2 from the risky option. We can thus use the 2-dimensional special case of the comparative Dirichlet distribution discussed above – the Beta distribution (see above for an explicit discussion of this simplifying assumption). In particular, the parameter α will encode the 'good state', in which the lottery pays a prize $x > c$, whereas β will encode the 'bad state', under which the lottery pays an outcome $y < c$. The perceived or sampled probability of the good state favoring the lottery will thus be $\mathbb{E}[\widehat{p}] = \frac{\alpha}{\alpha+\beta}$.

We start from an optimal choice rule entailing expected value maximization. The DM will thus choose the lottery over the sure amount whenever $\widehat{p}x + (1-\widehat{p})y > c$, or equivalently whenever

$$ln\left(\frac{\widehat{p}}{1-\widehat{p}}\right) > ln\left(\frac{c-y}{x-c}\right).$$

The transformation into log-odd space is convenient for computational reasons, but otherwise inconsequential (see Vieider 2024*b*, for an alternative derivation). The choice rule entails that the log-odds in favor of the lottery will be traded off against the log of the ratio of costs ($c-y$, potentially get the lower outcome y when c could have been had) and benefits ($x-c$; obtain the prize x instead of the lower sure amount c). Here, we will assume without loss of generality that the log cost-benefits are perceived objectively. This is a simplifying assumption that allows us to focus on the likelihood dimension, where most of the action takes place. It is straightforward to generalize the derivation to include the noisy coding of costs and benefits as well (cfr. Vieider 2024*b*).

The mean of the sampled log-odds can simply be derived from the two parameters containing the counts of successes and failures:

$$\mathbb{E}\left[ln\left(\frac{\widehat{p}}{1-\widehat{p}}\right)\right] = ln\left(\frac{\alpha}{\beta}\right)$$

Given limited samples, however, even samples that are accurate on average will contain some error on single draws, driven by natural sampling variation around the true mean. Averaging across all probabilities, we will thus observe

$$ln\left(\frac{\alpha}{\beta}\right) = ln\left(\frac{p}{1-p}\right) + \varepsilon,$$

which, following Atchison & Shen (1980), could equivalently be written as the difference of the digamma functions of the two parameters, $\digamma(\alpha) - \digamma(\beta)$.

Log-odds tend to follow approximately normal distributions, giving rise to a *logit-normal* (Atchison & Shen 1980). This suggests that $\varepsilon \sim \mathcal{N}(0, \nu^2)$. The error variance ν^2, in turn, again derives from the properties of the logit-normal distribution, and is given by the sum of trigamma functions of the two parameters, i.e. $\nu^2 = \digamma'(\alpha) + \digamma'(\beta)$.

A.3 Sampling bias and inference bias

We can illustrate the properties of these equations based on some examples. In DfE, the DM starts from a position of complete ignorance, i.e. zero discriminability. As they start sampling from the two options, they update the parameters α and β of the Beta distribution. This has two effects. One, the DM draws inferences about the underlying log-odds generating the observations. Two, as draws accumulate, uncertainty about the inferred log-odds is increasingly reduced.

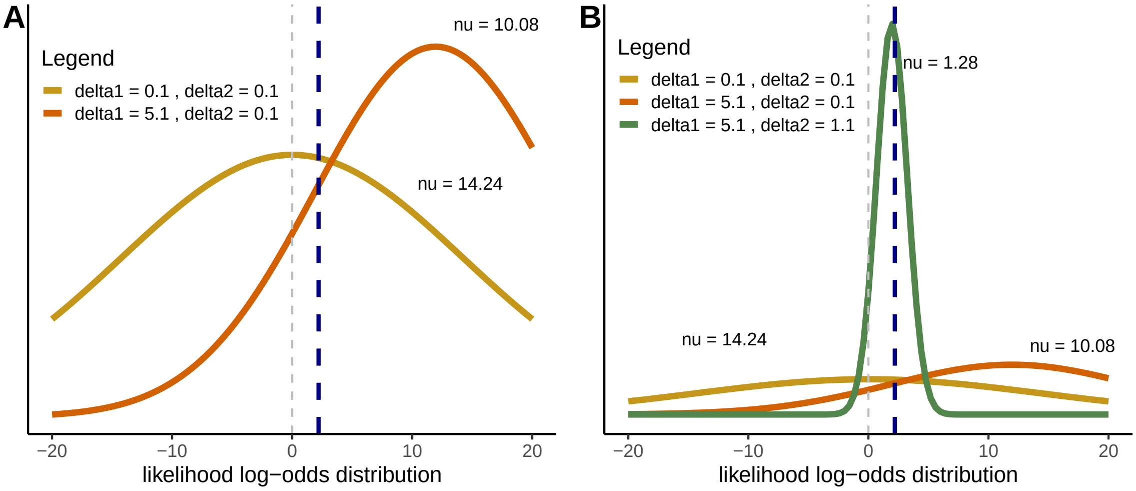


Figure 14: Inferences on log-odds: examples

The figure shows inferences on log-odds after a small number of draws. The dashed vertiical line indicates the true log-odds.

Figure 14 illustrates the process with two examples. Panel A shows the case of 5 samples being drawn from the risky option, all of which yield x. Starting from initial comparative Dirichlet parameters of 0.1, the true log-odds are now substantially over-estimated. Coding noise, however, remains large because the DM rightly infers that they have not observed all possible outcomes yet (i.e., the experimenter is unlikely to confront them with a degenerate choice – something that is implicit in the logit-normal formulation we use). This illustrates the importance of our assumption that there must be a minimum set of plasible outcomes, i.e. at least two states of nature that entail a meaningful tradeoff.

Panel B illustrates what happens when an outcome y is drawn from the risky option. The true log-odds are now somewhat *under*-estimated. At the same time, confidence in the inferred log-odds has increased dramatically. This illustrated the inter-play between sampling bias and inference bias in our setting. Inference bias thereby is a function of whether both outcomes from the risky option have been observed, and also of how many samples they are based on. As samples increase, both inference bias and sampling bias will thus decrease.

Optimal Combination with Bayesian prior

Given the noise in inferences, it will be optimal to combine the observations with a Bayesian prior. The optimality of this operation derives from the fact that – even though it will introduce systematic bias into the estimates under the form of regression to the mean – it will minimize the mean squared error across many estimates (see Ma et al. 2023, chapter 4, for an illustration). The reason for this is that the reduction in variance of the estimator will more than make up for the introduction of bias.

The objective for the mind now becomes to infer the log-odds from the underlying samples (whether they be true samples or virtual/neural samples – we drop the subscripts here and derive the equation just once). The inference problem for any given choice task will thus be as follows:

$$\mathbb{E}\left[ln\left(\frac{p}{1-p}\right) \mid \alpha, \beta\right] = \frac{\sigma^2}{\sigma^2+\nu^2}\, ln\left(\frac{\alpha}{\beta}\right) + \frac{\nu^2}{\sigma^2+\nu^2}\,\mu,$$

where we redefine $\mu = ln\left(\frac{p_0}{1-p_0}\right)$ in the main text, and where the Baysian evidence weight or "likelihood-discriminability" parameter is given by $\gamma \triangleq \frac{\sigma^2}{\sigma^2+\nu^2} = \frac{1}{1+\nu^2/\sigma^2}$. A step-by-step derivation of this equation can be found in Vieider (2024a), chapter 2.

In DfD, the "virtual draws" encoded in α and β (referred to as α_0 and β_0 in the main text) are unobservable. We can, however, estimate the equation by aggregating across multiple similar probabilities. This will yield the expectation over repeated stimuli of the posterior expectation above, which takes the following form:

$$\mathbb{E}\left[\mathbb{E}\left[ln\left(\frac{p}{1-p}\right) \mid \alpha, \beta\right] \mid p\right] = \frac{\sigma^2}{\sigma^2+\nu^2}\, ln\left(\frac{p}{1-p}\right) + \frac{\nu^2}{\sigma^2+\nu^2}\,\mu,$$

which now allows us to substitute the true log-odds for the sampled log-odds. Choice to choice fluctuations in the samples will be reflected in the variance of the distribution, which takes the form $\gamma^2\nu^2 = \frac{\sigma^4\nu^2}{(\sigma^2+\nu^2)^2}$.

Proof. The proof exploits the well-known property of the normal distribution whereby $z \sim \mathcal{N}(\widehat{z}, \tau^2)$ implies $bz + a \sim \mathcal{N}(b\widehat{z} + a, b^2\tau^2)$. To obtain the response distribution above, let $ln\left(\frac{\alpha}{\beta}\right) = z$, $\frac{\sigma^2}{\sigma^2+\nu^2} = b$, $\frac{\nu^2}{\sigma^2+\nu^2}\,\mu = b$, $ln\left(\frac{p}{1-p}\right) = \widehat{z}$, and $\nu = \tau$. □

Note that the problem does not change in any substantive way if we abandon the assumption of draws correctly reflecting the underlying distribution on average when real samples are taken in DfE. We then simply change the objective probability p to the sampled probability $\widehat{p}$ in the equations above. Sampling bias in $\widehat{p}$ will then occur on top of the inference bias, which still results in regression to the mean of the prior, just like represented above.

Stochastic choice rule

We can now trade off the inferred log-odds, as derived above, against the log-cost benefits, as suggested by our optimal choice rule. Letting $\mu \triangleq ln\left(\frac{p_0}{1-p_0}\right)$, we obtain $\delta = ln\left(\frac{p_0}{1-p_0}\right)^{1-\gamma}$, and by extension, $\theta = \delta^{-1} = ln(\frac{1-p_0}{p_0})^{1-\gamma}$. Putting everything on the scale of the standard deviation of the response distribution derived in the previous section yields the z-score describing the choice probability of the lottery:

$$pr[(x,p;y) \succ c] = \Phi\left[\frac{\gamma\, ln\left(\frac{p}{1-p}\right) - ln\left(\frac{c-y}{x-c}\right) - ln(\theta)}{\gamma\,\nu}\right],$$

where Φ is the standard normal cumulative distribution function. In DfD (as well as DfD+forced and DfE+forced), the probability will correspond to the correct one, and the model can thus be simply estimated on choice data by plugging the probit link function above into a Bernuoulli distribution (see below).

In DfE, we need to slightly amend the function above. In particular, we will now substitute sampled probabilities $\widehat{p}$ for the true probabilities above (adding a constant to both numerator and denominator to make sure it is defined—see discussion of the inference process above). An additional assumption concerns the log cost-benefit ratio when either x , y, or c have not yet been observed. The simplest assumption is that of a "naive" decision maker, who assumes the ratio to be 1 in that case (and hence its logarithm to be 0). However, this is just a special case of what a more sophisticated decision maker would do. Multiplying the log cost-benefit ratio by an additional parameter ρ, conditional on one of the outcomes not yet having been observed, allows for a more flexible specification whereby the DMs can (correctly) infer a positive correlation between log-odds and log cost-benefits. The "naive" DM discussed above is then just a special case for whom $\rho = 0$.

N-dimensional generalization

The general inference framework discussed at the beginning of this section is fully general. While we have described it for the particular case of comparisons used in our experiment, it can just as easily be applied to comparison between multi-outcome lotteries.

The inference framework introduced above remains directly applicable, with the two option-specific Dirichlet simply counting instances of different outcomes. We do, however, need to make an additional assumption when it comes to the construction of the comparative Dirichlet: our assumption here is simply that subjects order the outcomes in each Dirichlet by size in order to come up with the comparative distribution. The comparative Dirichlet is then constructed over k comparative states constructed based on the ranked outcomes.

Take two lotteries offering outcomes $\boldsymbol{x} = \{x_1, ..., x_k\}$ and $\boldsymbol{y} = \{y_1, ..., y_k\}$ under the comparative events $e_1, ..., e_k$, where each comparative event is characterized by a probability $\widehat{p_i}$, which could be different from the true underlying probability p_i. We assume that the outcome are ordered such that $x_1 \geq x_2 \geq ... \geq x_k$ and $y_1 \geq y_2 \geq ... \geq y_k$. We further assume that $\boldsymbol{x}$ is riskier than $\boldsymbol{y}$ in the sense of having wider spread or variance (entailing that $x_k < y_k$).[40] The optimal choice rule , which once again entais expected value maximization, takes the following form:

$$\sum_{i=1}^{k-1} \frac{\widehat{p_i}}{\widehat{p_k}} \frac{(x_i - y_i)}{(y_i - x_k)} > 1, \tag{8}$$

which sums the relative costs and benefits of the two lotteries. This equation has been used for instance in signal detection theory (Green et al. 1966). The reference outcome, here taken to be the worst outcome x_k, is arbitrary, since the choice rule enshrines within it all pairwise comparisons (similar to what happens in multinomial logits). To see this, let $V_{ik} \triangleq \frac{P[e_i]}{P[e_k]} \frac{(x_i - y_i)}{(y_i - x_k)}$. It is then straightforward to derive any binary comparison as $V_{ij} = {}^{V_{ik}}/_{V_{jk}}$. This expression thus maps the k-dimensional simplex into a (k-1)-dimensional log-odds

[40] The last assumption is not essential, but it ensures that the equation such as written here below is well-defined without the introduction of additional indexing.

representation.

The log-odds of each single state, now given by $ln\left(\frac{\widehat{p}_i}{\widehat{p}_k}\right)$, can be treated exactly as described above. Following Vieider (2024*b*) and assuming that the different states will be processed in parallel in a neural network, the stochastic choice equation then takes the following form:

$$P[\boldsymbol{x} \succ \boldsymbol{y}] = \sum_{i=1}^{k-1} \Phi \left[\frac{\gamma \times ln\left(\frac{\widehat{p}_i}{\widehat{p}_k]}\right) + \mathbb{1} \times ln\left(\frac{\mathbb{1}(x_i - y_i)}{y_i - x_k}\right) - ln(\theta)}{(k-1)\,\nu \times \gamma} \right].$$

where $\mathbb{1} = 1$ if $x_i - y_i > 0$ and else $\mathbb{1} = -1$, thus assuring that the logarithm is defined (we implicitly assume that the lowest outcome in the safer option is larger than the lowest outcome in the risky option, so that $y_i > x_k \; \forall \, i$). The multiplication of the log relative-outcome ratio by $\mathbb{1}$ further makes sure that this ratio enters with the appropriate sign, since it could now favor either choice option in any given state $i < k$.

B Experiments

Choice stimuli

We selected our choice stimuli from those in the early DfE literature (Hertwig et al. 2004), but generalized them so as to allow us to structurally estimate our model, and to obtain a more balanced picture of the behavior. We assured identification of the structural estimations using simulations (R code available upon request), which allowed us to find the optimal compromise between number and type of task and the length of the experiment. The limiting factor derived in particular from the forced sampling experiments, where subjects had to take 40 samples by tasks, as well as expressing their final choice.

We thus chose 6 different lotteries—3 with a small probability, and 3 with a large probability of winning. We then obtained three choice tasks by lottery by setting the sure amount c equal to the expected value, and by adding or subtracting a fixed amount. This provides some valuable variation for the structural estimations, and results in the following 18 unique tasks (4 randomly selected ones of which were repeated in the experiment):

Table 1: Choice tasks

small p	large p
(31,0.10) vs. 2.8	(4,0.80) vs. 3.0
(31,0.10) vs. 3.2	(4,0.80) vs. 3.2
(31,0.10) vs. 3.6	(4,0.80) vs. 3.4
(10,0.15) vs. 1.2	(8,0.85) vs. 6.6
(10,0.15) vs. 1.5	(8,0.85) vs. 6.8
(10,0.15) vs. 1.8	(8,0.85) vs. 7.0
(16,0.20) vs. 2.9	(10,0.90) vs. 8.8
(16,0.20) vs. 3.2	(10,0.90) vs. 9.0
(16,0.20) vs. 3.5	(10,0.90) vs. 9.2

Choice tasks are describes as usual, with (x, p) designating a lottery providing a prize x with probability p or else 0, and c designating the sure amount.

C Experiment 4: the DfD+free Treatment

Forcing subjects to sample from described options shows dramatic effects on behavior. This raises the question of whether subjects will sample voluntarily when given a description of the options, even when they are not forced to do so. We designed experiment 4 to answer this question by introducing a DfD+free treatment. We show subjects the same information about lotteries as we do in the DfD treatment, but we also provide subjects the sampling tools just like in experiment 3. Other than in experiment 3, however, the radio buttons to indicate a choice appear from the very start. Subjects are told explicitly that they can sample if they want to but that they do not have to, and that they can also indicate their decision directly without sampling. We ran this treatment on Prolific with 101 subjects in September 2023.

C.1 Results

Subjects do indeed sample when given the possibility to do so, even when choice options are fully described. Across subjects and tasks, they take an average of 1.74 samples, almost all of which are taken from the risky option. This average, however, hides significant heterogeneity within it. Some subjects take as many as 9 samples per task on average, whereas others sample very little. That being said, only 8% of subjects never sample at all. Samples also change significantly over time, starting at more than 4 on average in the first round, and then declining to about 2 by round 5, to settle on an average of 1.4 thereafter. The fact that DMs do sample fully redundant information seems remarkable in our context, given the high opportunity costs of subjects on Prolific.

We next examine what happens to choice behavior once free sampling is introduced. Our updating equations raise an intriguing question: may free sampling in DfD introduce (limited) sampling bias into DfD? The question arises simply because, although subjects are given an objective description of the odds, our updating equation (7) suggests that actual samples drawn are simply added to the neural samples representing the evidence in favour and against the lottery. Small samples, however, will suffer from the same issues we have seen in DfE: they will tend to be biased against observing the rare event, so that our model predicts that they will yield biased updates of the true log-odds.

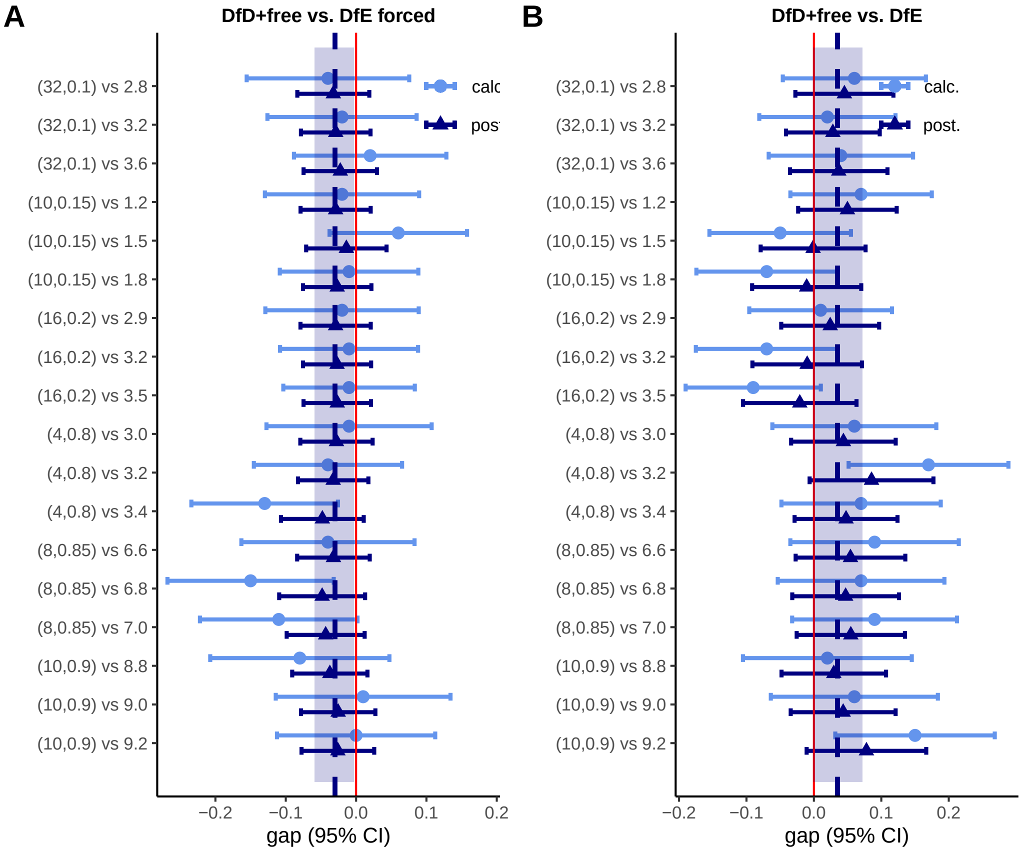


Figure 15: Meta-analysis of the GAP after free sampling from description

Panel A shows a forest plot of the gap between DfD+free versus DfE+forced. Panel B shows a forest plot of the GAP between DfD+free and DfE. The light blue circles, labeled ‘calc.’, indicate the raw differences in choice proportions in the data, g. The dark blue triangles, labeled ‘post.’, indicate the inferred posterior parameters, $\widehat{g}$. The thick, dashed vertical line indicates the meta-analytic posterior mean, μ, and the shaded rectangle indicates the 95% credible interval around that estimate.

Figure 15 shows behavior resulting from free sampling from fully described options, and compares it to various benchmarks using our meta-analytic technique. The difference in choice proportions are coded in the usual way, consistent with the standard description-experience gap. Panel A examines the GAP between DfD+free and DfE+forced. While only 2 of the raw differences in choice proportions are significant at conventional levels (and none of the posterior differences), the meta-analytic mean indicates a small *negative* GAP of 3pp, with its 95% credible interval of $[-0.059\,, -0.002]$ pp. indicating a statistically significant effect.[41] Panel B examines the GAP between DfD+free and DfE (also with free sampling). At 3.5 pp., the point estimate of the GAP is now again positive, consistent with the direction of the standard gap. However, it is (just) not significantly different from 0 at conventional levels, with a 95% credible interval of $[-0.001\,, 0.072]$. In other words, free sampling from described options closes the GAP with DfE. By levelling the playing field and allowing for free sampling in both cases, we thus again close the GAP, but now by making DfD more similar to DfE.

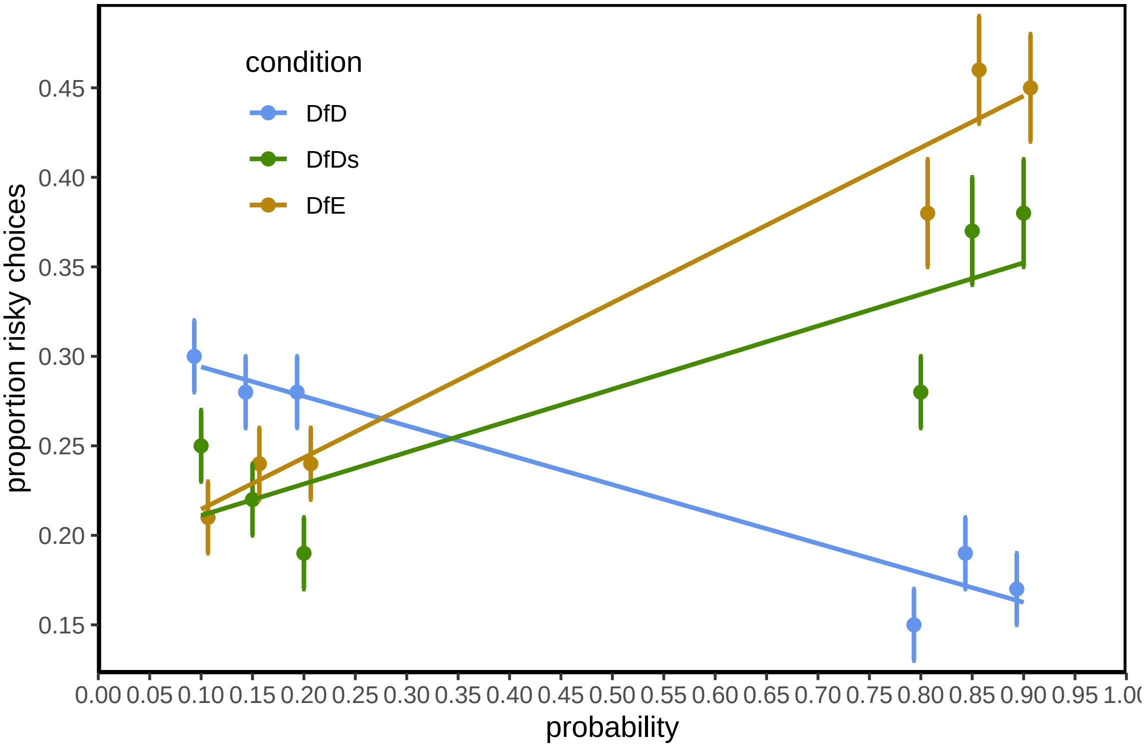


Figure 16: Structural estimates, DfD versus DfE

The figure shows structural estimates of the model parameters. Panel A compares likelihood-discriminability γ in DfE and DfE+forced. Panel B compares likelihood-discriminability, γ,in DfD and DfD+forced.

[41] The significant result notwithstanding the relatively small effect seems in particular due to the fact that almost all differences in choice proportions go in the same direction. Even though few of them are significant individually, this produces a highly coherent effect at the aggregate level, reducing the variance and hence the confidence bounds.

Figure 16 shows the raw choice proportions in DfD+free, and directly juxtaposes them with the choice proportions in DfD and in DfE. The difference from DfD is very large, with somewhat less risk taking for small probabilities, and much more risk taking for large probabilities. This results in a positive dependence of choice proportions on the probability of winning. Indeed, a meta-regression of the choice proportions on the probability of winning indicates a significantly positive slope, with $\lambda = 0.174$, and a 95% credible interval of $[0.141\,,\,0.241]$. This further suggests that sampling bias now affects DfD, just like predicted by our model. Nonetheless, the influence of the description is strong enough to keep the likelihood-dependence significantly smaller than observed under DfE (where we have observed a $\lambda = 0.284$, with a credible interval of $[0.226\,,\,0.347]$). This shows that samples from description—while affecting probabilities in the way predicted by our model—are still balanced against the description provided on the screen, with choices indicating an aggregation of the two types of information.

The feat of closing the GAP with DfE by acting on DfD is remarkable inasmuch it achieves something that acting on DfE alone has never achieved—it closes the GAP by manipulating one of the two sides only. This further speaks to the mechanism predicted by our model: studies intervening on DfE have generally removed sampling bias by increasing the number of samples taken.[42] This, in turn, has decreased sampling bias while at the same time strongly reducing inference bias, with the upshot that the GAP was not eliminated. In the DfD+free treatment, we achieve the opposite: guided by our model, we introduced a limited degree of sampling bias into DfD, all the while keeping inference bias relatively large due to the few samples added. This closes the GAP when people can sample freely.

D Meta-analytic estimation

Our meta-analysis follows the "standard" equations presented in the main text. We estimate the model in Stan (see Vieider 2024*a* for a tutorial on the use of Stan for decision models; chapter 4 contains a part specifically dedicated to meta-analysis). Here is the Stan code used to estimate the model:

[42] While some studies have tried to use the sampled probabilities in such a way as to create matched probabilities from description, thus potentially circumventing this issue, such approaches have run into the limitation of producing many degenerate choices between different sure amounts of money. This is due to the fact that many subjects make a choice without ever having observed on of the outcomes—something that is indeed consistent with our model, and which further showcases the effects of inference bias. See Wulff et al. (2018), p151 and onwards, for a discussion of these issues and an overview of attempts to close the GAP in the literature.

```
//footnotesize
data{
    int<lower=1> N; \\number of observation
    vector[N] gap; \\difference in choice proportions
    vector<lower=0>[N] se; \\standard error of the difference
}
parameters{
    vector[N] gamma; //true, estimated gap (called g_hat in paper)
    real mu; //meta-analytic mean (omega in paper)
    real<lower=0> sigma; //variance
}
model{
//regularizing priors
    sigma ~ normal( 0 , 1 );
    mu ~ normal( 0 , 1 );

    // measurement error model:
    gap ~ normal( gamma , se );

    // likelihood:
    gamma ~ normal( mu , sigma );
}
```

The meta-regression is introduced into the same code simply by modifying the mean *mu*, making it dependent on the probability of winning:

```
//footnotesize
data{
    int<lower=1> N;
    int<lower=1> K; //dimension of design matrix
    vector[N] gap;
    vector<lower=0>[N] se;
    matrix[N,K] X; //design matrix of explanatory variables
}
parameters{
    vector[N] gamma;
    real mu;
    real<lower=0> sigma;
     vector[K] beta;
}
model{
    sigma ~ normal( 0 , 1 );
    mu ~ normal( 0 , 1 );

    // measurement error model:
    gap ~ normal( gamma , se );
    // likelihood:
    gamma ~ normal( mu + X * beta , sigma );
}

}
```

E Structural estimation

We implement our structural equations based on the discriminability equation in the main text, using the objective probability of winning, p, in DfD, DfD+forced, and DfE+forced. We use the sampled probability $\widehat{p}$ in DfE, and complement this with an assumption about the log-cost benefits in the case that one of the outcomes has not yet been observed when the decision is taken, as described above.

We keep the model as simple as possible in order to maximize our comparative power and to keep the model parsimonious. This means, first of all, that we normalize the coding noise variance by division with the variance of the prior, so that $\gamma = \frac{1}{1+\frac{\nu^2}{\sigma^2}}$. This helps both identifiability and comparability across treatments but happens without loss of generality, since it is the ratio between coding noise variance and prior variance that determines behavior (see also Natenzon 2019). Another assumption that we maintain throughout the paper is that the mean of the prior, μ, remains unaffected over the course of the experiment. We exploit this in the estimation by letting μ be the same across the 2 parts of the experiment, whereas ν and as a consequence γ and θ are all allowed to vary freely.

We estimate the model using a Bayesian hierarchical setting in Stan (Carpenter et al. 2017). The hierarchical setting allows us to pool information from the aggregate estimation, which provides the priors, and from individual-level parameter estimates, which contribute to the aggregate in proportion to their precision. The aggregation equation follows exactly the equation we describe for our Bayesian inference process. Vieider (2024*a*) provides a step-by-step tutorial on the estimation fo decision models in Stan.

Below, we include an commented version of the code we use in DfD, DfD+forced, and DfE+forced (the code used in DfE is very similar, and only has an additional parameter ρ, as well as including the truely observed log-odds as data; it is available upon request). We define the varianbles at the level of the individual *choices*. This allows us to implement a literal specification of our model, where task-specific quantities are encoded by parameters α and β. These parameters are nested in individual-level parameters, which we use to fit the choice data, and which ensures that the choice-level parameters are identified and well-behaved (since the individual-level parameters act as informative priors). Finally, individual-level parameters are nested within an overall distribution.

We check convergence by making sure that all R-hats are below 1.05. We also carefully check that any divergent iterations do not indicate problems with the posterior (and discard all estimates with more than 1% divergent iterations). The hyperpriors on the aggregate parameter means are given very wide priors, which makes them *mildly regularizing*—they help the convergence of the simulation algorithm by being centered around the region where we expect the parameter values to fall, but they attribute significant probabilitry mass to 1 order of magnitude above the region into which we would expect the parameters to reasonably fall. Our estimates are indeed not sensitive to the choice of the exact parameter values. This follows best practices in Bayesian estimation.

```
data{ \\declare data
    int<lower=1> N; \\number of observations
    int<lower=1> N_id; \\number of subjects
    array[N] int id; \\unique identifier
    array[N] real high; \\outcome x
    array[N] real low; \\outcome y
    array[N] real sure; \\outcome c
    array[N] real p; \\probability
    array[N] int choice_risky;\\choice: 1 if risky
    array[N] int part2; \\dummy to indicate part 2
}
transformed data{
 array[N] real lcb; \\log cost benefit ratio
 array[N] real llr; \\log-odds
    for (i in 1:N){
      lcb[i] = log( (sure[i] - low[i]) / (high[i] - sure[i]) );
      llr[i] = log( p[i]/(1 - p[i]) );
      }
}
parameters{
    vector[3] means; \\aggregate mean parameters on log scale
    vector<lower=0>[3] tau_id; \\aggregate parameter variances
    cholesky_factor_corr[3] L_omega_id; \\decomposed covar matrix
    array[N_id] vector[3] Zid; \\stan dardized individual-level parameters
}
transformed parameters{
// covar and temp parameters
  matrix[3,3] Rho_id = L_omega_id * L_omega_id'; \\obtain covariance matrix
  array[N] vector[3] pars; \\parameter matrix on log scale
// generative parameters:
  vector[N] mu; \\prior mean
  vector<lower=0>[N] kappa; \\concentration part1
  vector<lower=0>[N] kappaf; \\concentration part2
// derived parameters from here
  vector[N] alpha; \\derived parameters—see definitions in text, and below
  vector[N] beta;
  vector[N] nu;
  vector[N] gamma;
  vector[N] theta;
  vector[N] omega;
  vector[N] alphaf;
  vector[N] betaf;
  vector[N] nuf;
  vector[N] gammaf;
  vector[N] thetaf;
  vector[N] omegaf;
  for (i in 1:N){
    pars[i] = means + diag_pre_multiply(tau_id,L_omega_id) * Zid[id[i]];
    mu[i] =  pars[i,1];
    kappa[i] = exp(pars[i,2]);
    kappaf[i] = exp(pars[i,3]);
// define derived parameters
    alpha[i] = kappa[i] * p[i];
    beta[i] = kappa[i] * (1 - p[i]);
    nu[i] = sqrt( trigamma( alpha[i] ) + trigamma( beta[i] ) );
    gamma[i] = 1/( 1 + nu[i]^2 );
    theta[i] = exp( ( gamma[i] - 1) * mu[i]) ;
    omega[i] = nu[i] * gamma[i];
    alphaf[i] = kappaf[i] * p[i];
    betaf[i] = kappaf[i] * (1 - p[i]);
    nuf[i] = sqrt( trigamma( alphaf[i] ) + trigamma( betaf[i] ) );
    gammaf[i] = 1/( 1 + nuf[i]^2 );
    thetaf[i] = exp( ( gammaf[i] - 1) * mu[i] ) ;
    omegaf[i] = nuf[i] * gammaf[i];
  }
}
model{
    vector[N] udiff; \\local vector
\\priors for aggregate (hierarchical) parameters
    tau_id ~ exponential(5);
    L_omega_id ~ lkj_corr_cholesky(4);
    means[1] ~ normal(0, 5);
    means[2] ~ normal(0, 5);
    means[3] ~ normal(0, 5);

\\priors for individual level parameters, standardized:
  for (n in 1:N_id)
      Zid[n] ~ std_normal();

\\the mode:
  for ( i in 1:N ) {
          udiff[i] = ( ( gamma[i] * llr[i] - lcb[i] - log(theta[i]) )/ omega[i] ) * (1 - part2[i
                     ( ( gammaf[i] * llr[i] - lcb[i] - log(thetaf[i]) )/ omegaf[i] ) * part2[i];
          choice_risky[i] ~ bernoulli( Phi( udiff[i] ) );
    }
}
\\code below recovers individual-level parameters
generated quantities{
  vector[N] log_lik;
  vector[N] udiff;

  vector[N_id] mun;
  vector[N_id] kappan;
  vector[N_id] alphan;
  vector[N_id] betan;
  vector[N_id] nun;
  vector[N_id] gamman;
  vector[N_id] thetan;
```

```
   vector[N_id] kappafn;
  vector[N_id] alphafn;
  vector[N_id] betafn;
  vector[N_id] nufn;
  vector[N_id] gammafn;
  vector[N_id] thetafn;

  vector[3] temp;
    for(n in 1:N_id){
      temp = means + diag_pre_multiply(tau_id,L_omega_id) * Zid[n];
      mun[n] = temp[1];
      kappan[n] = exp(temp[2]);
      kappafn[n] = exp(temp[3]);
      alphan[n] = kappan[n]/2;
      betan[n] = kappan[n]/2;
      nun[n] = sqrt( trigamma( alphan[n] ) + trigamma( betan[n] ) );
      gamman[n] = 1/(1 + nun[n]^2 );
      thetan[n] = exp( ( gamman[n] - 1 ) * mun[n] );
      alphafn[n] = kappafn[n]/2;
      betafn[n] = kappafn[n]/2;
      nufn[n] = sqrt( trigamma( alphafn[n] ) + trigamma( betafn[n] ) );
      gammafn[n] = 1/(1 + nufn[n]^2 );
      thetafn[n] = exp( ( gammafn[n] - 1 ) * mun[n] );
      }

    for ( i in 1:N ) {
        udiff[i] = ( ( gamma[i] * llr[i] - lcb[i] - log(theta[i]) )/ omega[i] ) * (1 - comp[i])
                     ( ( gammaf[i] * llr[i] - lcb[i] - log(thetaf[i]) )/ omegaf[i] ) * comp[i];
        log_lik[i] = bernoulli_lpmf( choice_risky[i] | Phi_approx( udiff[i] ) );
    }
}
```

F Prospect Theory Identification and Estimation

Let v_o be the utile of an outcome o, and let π_p be the decision weight a probability p. The condition under which the gain lottery $(x, p; y)$ will be chosen over the intermediate sure outcome c can then be represented under prospect theory as $\pi_p v_x + (1 - \pi_p) v_y \geq v_c$. To be fit to data, this equation will need to be augmented by a stochastic choice model. The most popular choice is the random utility model, which adds an error ε_ℓ to the lottery, and an error ε_c to the sure amount. Assuming that both errors are normally distributed mean 0 ("white noise" errors), and letting $\varepsilon = \varepsilon_\ell - \varepsilon_c \sim \mathcal{N}(0, \tau^2)$, the choice probability can be expressed as follows:

$$Pr[(x,p;y) \succ c] = \Phi\left[\frac{\pi_p v_x + (1-\pi_p)v_y - v_c}{\tau(x-y)}\right],$$

where Φ is the standard normal cumulative distribution function, and where we have made the error variance heteroscedastic across the range of the outcomes in the lottery (see e.g. Bruhin et al. 2010 and L'Haridon & Vieider 2019).

For $y = 0$, the PT model described above is only unique up to a power (Gonzalez & Wu 1999). This issue, however, can be overcome by assuming specific functional forms for the utility and probability weighting functions, which map outcomes and probabilities into utils and decision weights, and by identifying the parameters of such functions *jointly with the error term.* Identification jointly with the error is thereby crucial, since taking the whole stochastic equation to a power would reduce the fit to the data, given that rescaling the (deterministic) PT choice equation would inevitably come at the expense of increasing the error and hence decreasing the fit in our setting. Extensive simulations indeed show that – based on such additional assumptions—we can recover simulated PT parameters from binary choices between lotteries $(x, p; 0)$ and sure amounts c with high degrees of accuracy, as long as there is some orthogonality in the variation of p and x (and consequently, c).

In our estimations, we use power utility throughout, so that:

$$u(x) = x^\eta.$$

This is the most commonly used utility specification by far in the literature. The specification of utility is furthermore relatively unimportant when it comes to describing the shape of the probability weighting function, which is our main quantity of interest. For the latter, we assume the popular linear in log-odds function (Gonzalez & Wu 1999), which creates a direct bridge to our noisy coding model:

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

where δ captures optimism, and γ likelihood-sensitivity, with $\gamma < 1$ describing likelihood-insensitivity as usually observed in DfD, and $\gamma > 1$ likelihood-oversensitivity as usually observed in DfE when estimating the functions o the true underlying probabilities.

We again estimate the model in a Bayesian hierarchical setup in Stan. All the procedures are the same as described in the previous section for the structural estimations of our model. The Stan code takes the following form (see Vieider 2024*a*, for a tutorial and step-by-step explanation of this code):

```
data{
    int<lower=1> N;
    int<lower=1> N_id;
    array[N] int id;
    array[N] real high;
    array[N] real low;
    array[N] real sure;
    array[N] real p;
    array[N] int choice_risky;
}
parameters{
    vector[4] mus;
    vector<lower=0>[4] tau;
    cholesky_factor_corr[4] L_omega;
    array[N_id] vector[4] Z;
}
transformed parameters{
  matrix[4,4] Rho = L_omega*L_omega';
  array[N_id] vector[4] theta;
  vector<lower=0>[N_id] rho;
  vector<lower=0>[N_id] gamma;
  vector<lower=0>[N_id] delta;
  vector<lower=0>[N_id] sigma;
  for (n_id in 1:N_id){
    theta[n_id] = mus + diag_pre_multiply(tau, L_omega) * Z[n_id];
    rho[n_id] = exp(theta[n_id,1]);
    gamma[n_id] = exp(theta[n_id,2]);
    delta[n_id] = exp(theta[n_id,3]);
    sigma[n_id] = exp(theta[n_id,4]);
  }
}
model{
    vector[N] wp;
    vector[N] pv;
    vector[N] udiff;
    vector[N] pu;
    tau ~ exponential(5);
    L_omega ~ lkj_corr_cholesky(4);
    mus[1] ~ normal(-0.5, 0.25);
    mus[2] ~ normal(0, 2);
    mus[3] ~ normal(0, 2);
    mus[4] ~ normal(0, 2);
  for (n_id in 1:N_id)
      Z[n_id] ~ std_normal();
  for ( i in 1:N ) {
        wp[i] = (delta[id[i]]*p[i]^gamma[id[i]])/(delta[id[i]]*p[i]^gamma[id[i]] + (1 - p[i])^g
        pu[i] = ( wp[i] * high[i]^rho[id[i]] + (1 - wp[i]) * low[i]^rho[id[i]] );
        udiff[i] = (pu[i] - sure[i]^rho[id[i]])/(sigma[id[i]]  * (high[i] - low[i]) );
        choice_risky[i] ~ bernoulli_logit( udiff[i] );
    }
}
```

G Additional results

Additional results on free sampling in DfE

Subjects take relatively few samples in our experiment, something that may be explained by the high opportunity costs faced by subjects on Prolific, who—contrary to students in lab or classroom settings—can leave as soon as they are done with the experiment and move on to other earning opportunities. The average number of samples taken is 8, which puts our study at roughly the first tercile of the distribution summarized in the meta-analysis of Wulff et al. (2018). Samples taken, however, generally tend to be lower in tasks comparing lotteries with sure outcomes, as we use here. The average subject on the average task takes 3.3 samples from the safe option, but 4.3 samples from the risky option. However, samples vary greatly between individuals, ranging from 2 on average (1 per option) to about 40.

Panel A in Figure 4 examines the average samples by probability from the risky option at the task resolution. The samples are presented following a median split on risk aversion in the first, description-based, part of the treatment, implemented as the proportion of choices of the sure amount. This aims to test our model prediction according to which samples should vary with the underlying probability depending on the initial risk aversion of the DM. These predictions are strongly supported by the evidence presented in the figure. Risk averse DMs take few samples from small-probability lotteries, but sample significantly more from large-probability lotteries. For the least risk averse half of the sample, we observe a (somewhat weaker) trend in the opposite direction. This aligns with our prediction, according to which

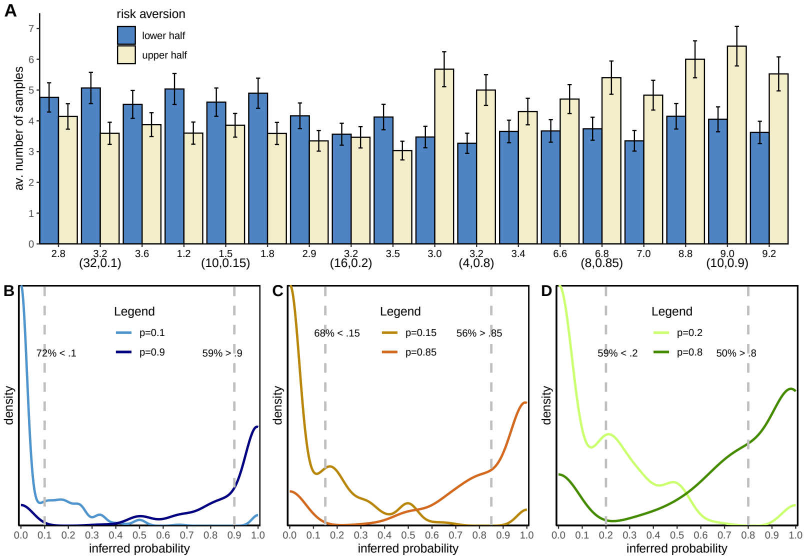


Figure 17: Samples by probability and risk aversion

The figure shows the number of samples taken from the risky option by probability and risk aversion at the task level resolution in Panel A. Risk aversion is assessed as the proportion of safe choice in the first, DfD part of the experiment, after removing repeated tasks. The categorization is obtained using a median split. Error bars show ±1 standard error. Panels B through D show the distribution of sampled probabilities by different actual probabilities.

risk averse DMs should have less of a conflict between noise and sampling bias in small probability lotteries, thus reaching a decision more quickly.

Table 2 shows a more nuanced analysis using regressions on continuous risk aversion measures. Regression I shows that samples taken increase in the probability of winning across the whole sample. Regression II qualifies this effect, showing that the larger overall samples are mainly driven by risk aversion, and that probability-dependence of samples taken interacts strongly with risk aversion. The results thus support the tension between risk aversion and sampling bias for large probabilities predicted by our model. Regression III further shows that the prize x and the absolute deviation from the EV do not matter, but that samples decline slightly over the rounds of the experiment.

The small number of samples taken is reflected in the probabilities people experience. This is illustrated figure 17, panels B through D, which plot distributions of probabilities inferred from the actual samples a DM observed. For small probability lotteries, subjects experience a smaller probability than the true one in 66% of cases overall, while getting a correct picture in some 3.4% of cases. For large probability lotteries this picture is reversed, with 55% of samples over-estimating the true probability, and only 2.2% resulting in a correct estimate. The asymmetry we see between small and large probabilities suggests that the

dep. var:	**number of samples**			**sampling bias**		
	reg. I	reg. II	reg. III	reg. IV	reg. V	reg. VI
probability	0.288	0.331	0.299	-0.029	-0.027	-0.066
	(0.082)	(0.085)	(0.120)	(0.009)	(0.008)	(0.011)
risk aversion		0.569	0.615		-0.016	-0.008
		(0.291)	(0.282)		(0.012)	(0.017)
prob × risk av.		0.618	0.638		-0.025	-0.023
		(0.087)	(0.085)		(0.009)	(0.008)
prize			-0.004			-0.002
			(0.005)			(0.001)
abs. EV dev.			-0.180			-0.012
			(0.230)			(0.020)
round			-0.031			0.001
			(0.005)			(0.001)
constant	3.635	3.705	4.739	0.036	0.046	0.006
	(0.283)	(0.289)	(0.329)	(0.010)	(0.011)	(0.023)
observations	2178	2178	2178	2178	2178	2178
subjects (clusters)	99	99	99	99	99	99

Table 2: Regression analysis of samples

Regressions in the table are based on a Bayesian outlier-robust regression model. Robust regression is implemented by means of a student-t distribution with 2 degrees of freedom, and regressions are programmed with random intercepts to cluster errors at the subject level. Regressions I, II, and III use the total number of samples from the risky option as dependent variable. Regressions IV, V, and VI use the sampling bias, defined as the true probability minus the inferred probability for small probability lotteries, and as the inferred probability minus the true probability for large probability lotteries, as dependent variable. Numbers in parentheses indicate standard errors. Probability and risk aversion are normalized by taking z-scores.

larger samples taken for large probabilities result in a more balanced picture.

This is further tested in regressions IV to VI in table 2. Regression I shows that sampling bias declines in the probability of winning for the whole sample. Regression II refines this picture by controlling for risk aversion and its interaction with the probability of winning, and shows sampling bias is reduced most strongly for risk averse DMs in large-probability lotteries. This is indeed what we would expect, given the sampling behavior discussed above. Interestingly, the only effect not paralleling those for the number of samples is the effect of decision round. In particular, we do not find an increase in bias over the rounds. This suggests that the effect in regression III whereby samples decline over the rounds may be due to learning, in the sense that subjects become more focused in their samples, rather than to fatigue.

Nonparametric within-subject results

Here, we replicate the nonparametric between-subject analysis in the paper by presenting within-subject comparisons wherever this is possible. The descriptions of the figures are self-contained.

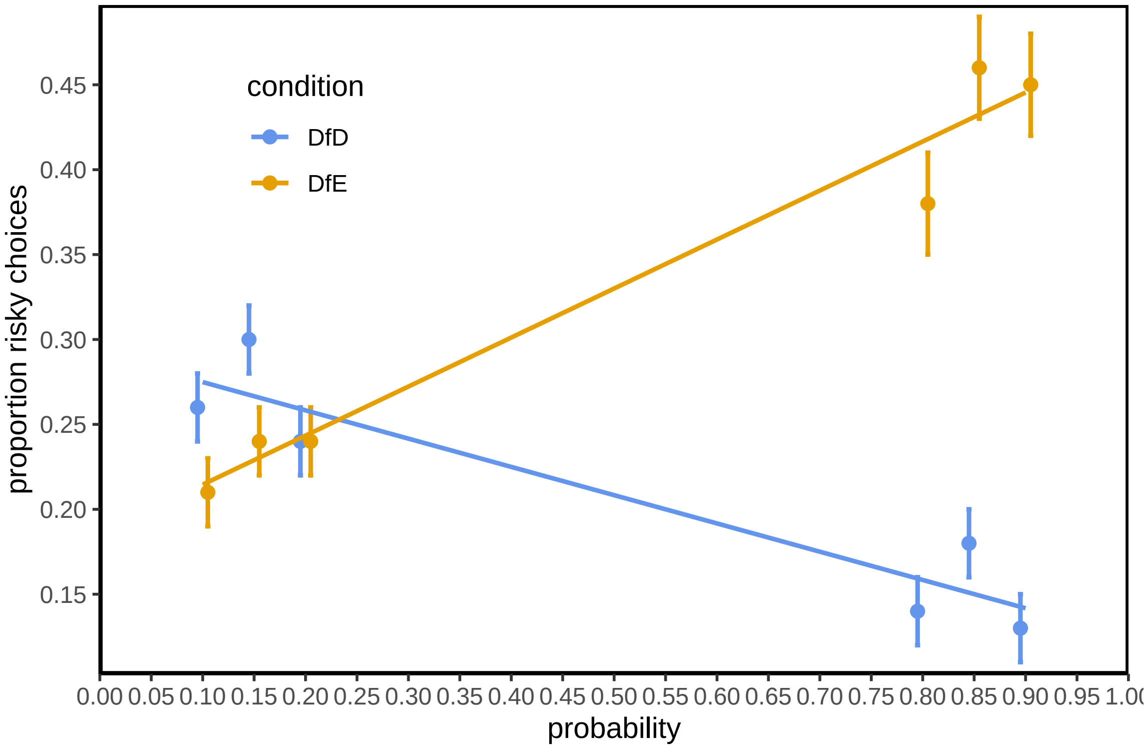


Figure 18: The GAP: within-subject

Choice proportions by probability for the decision-experience gap: DfD versus DfE. Error bars indicate 1 standard error.

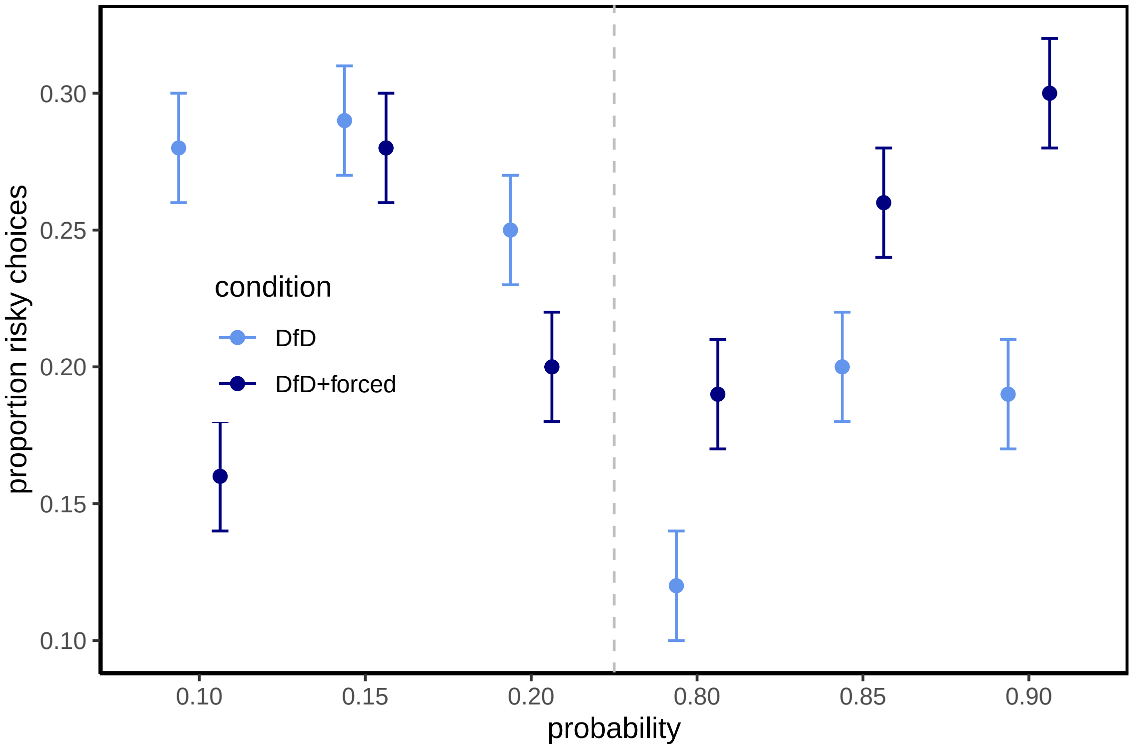


Figure 19: DfD+forced vs DfD within subject

Choice proportions by probability, within-subject comparison between DfD+forced and DfD. Error bars indicate 1 standard error.

Figures at task level

Here, we show all figures for which we averaged across c at the probability level at a task-level resolution. The figure descriptioins are self-contained.

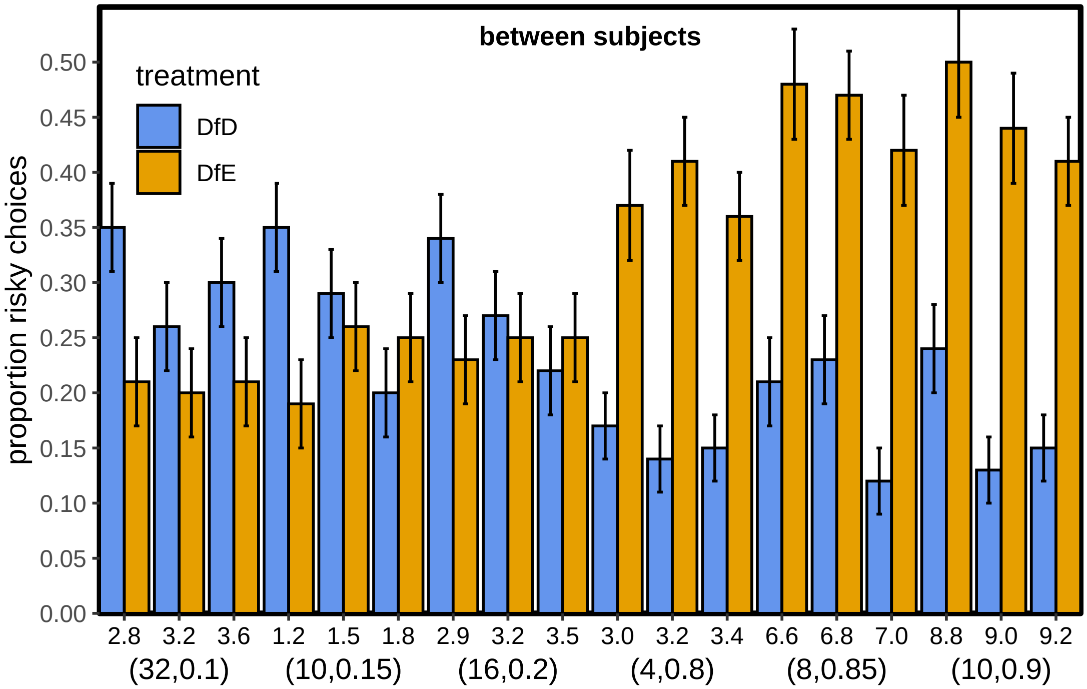


Figure 20: The GAP at the task level (between-subjects)

Choice proportions by task for the decision-experience gap: DfD versus DfE. Error bars indicate 1 standard error.

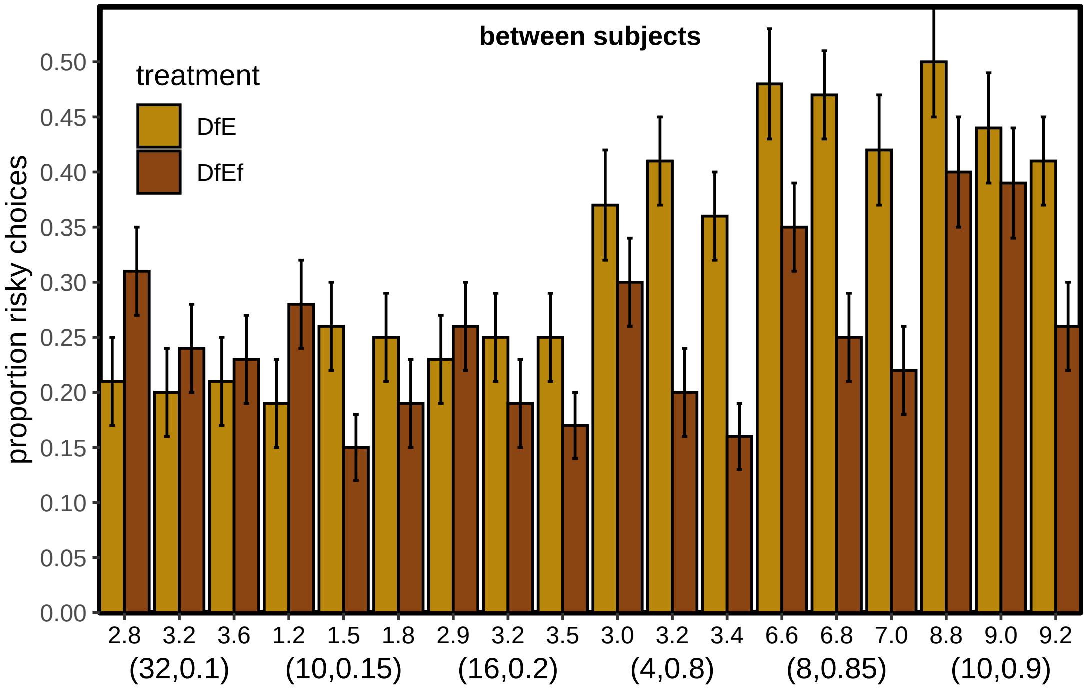


Figure 21: DfE+forced versus DfE at the task level (between-subjects)

Choice proportions by task for DfE+forced compared to DfE. This comparison is only possible between-subjects. Error bars indicate 1 standard error.

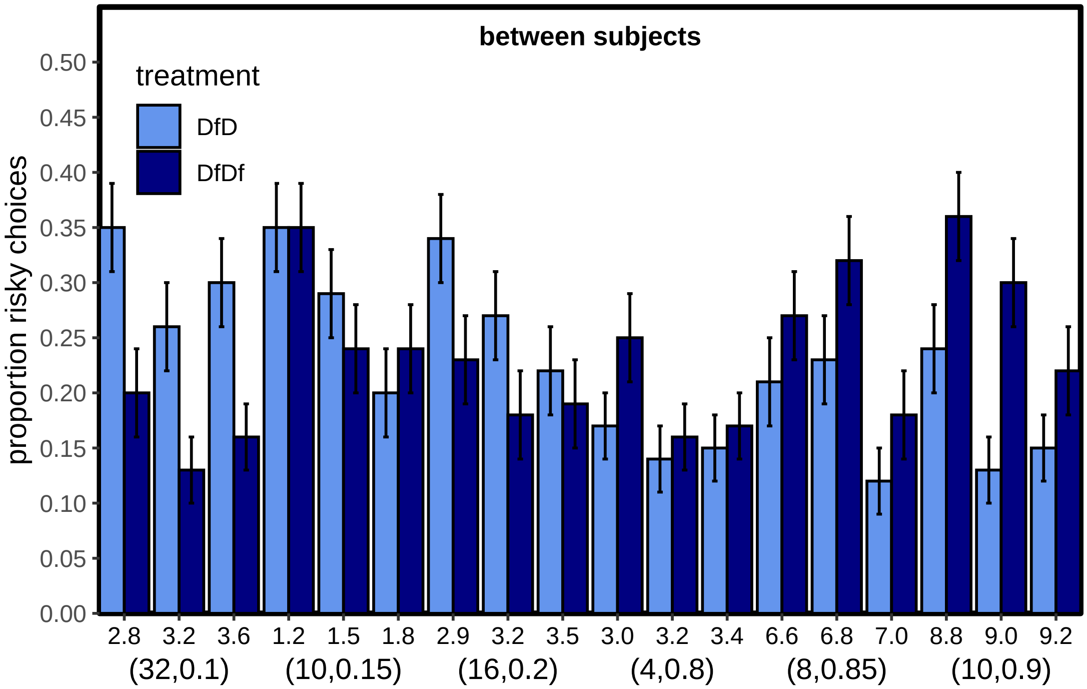


Figure 22: DfD+forced versus DfD at the task level (between-subjects)

Choice proportions by task for DfD+forced compared to DfD. Error bars indicate 1 standard error.

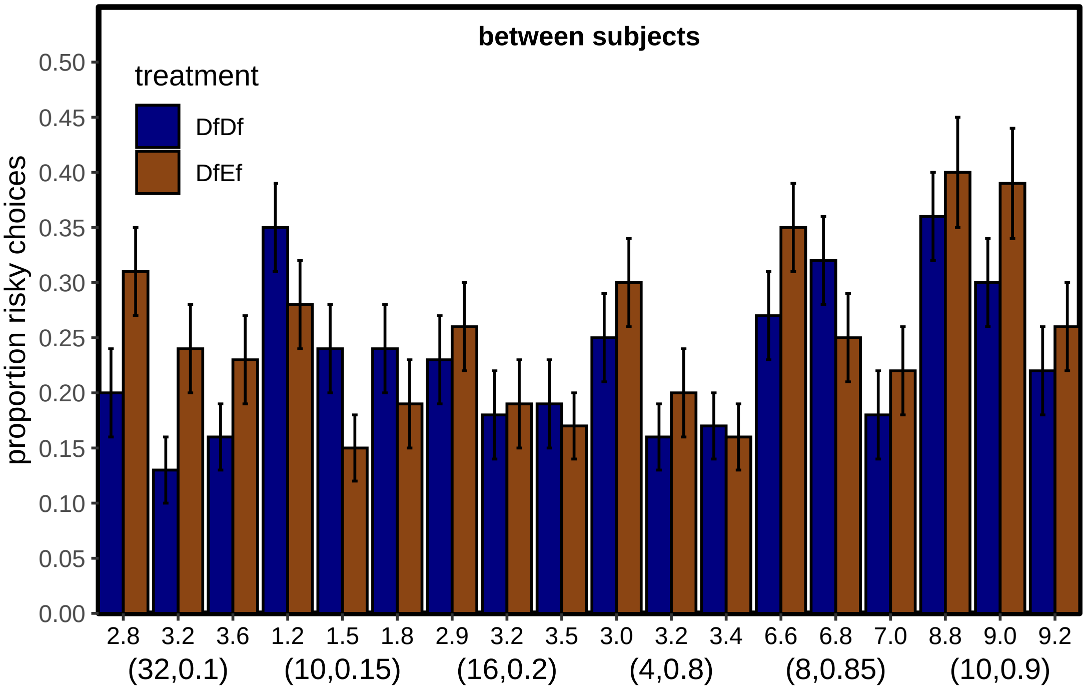


Figure 23: DfD+forced versus DfD at the task level (between-subjects)

Choice proportions by task for DfD+forced compared to DfD. Error bars indicate 1 standard error.

Within-subject structural results

This section contains within-subject structural comparisons for those cases where we used between-subject comparisons in the main text, but within-subject comparisons are possible. The descriptions of the graphs are self-contained.

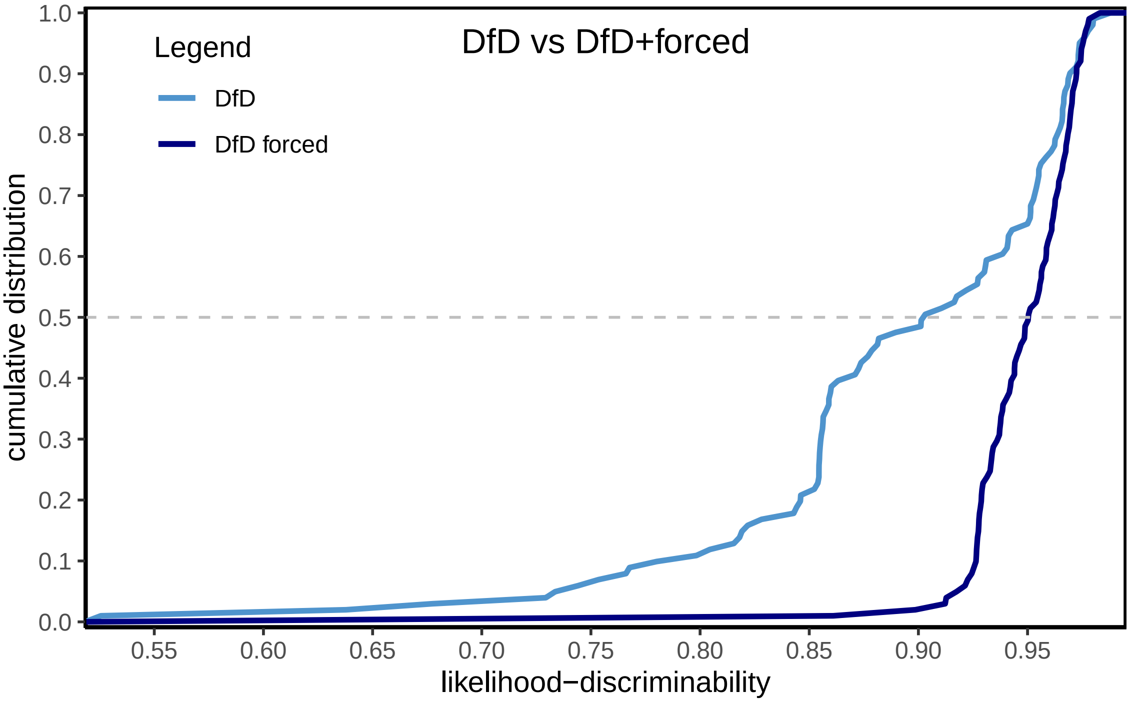


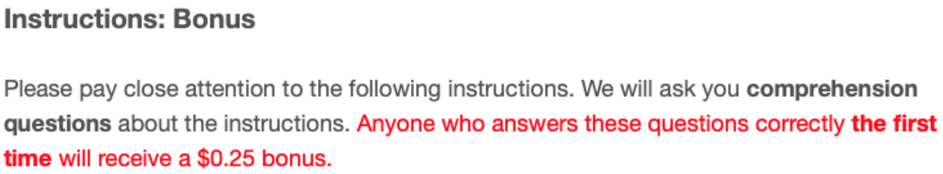
Figure 24: Likelihood-discriminability in DfD vs DfD+forced, within subject

Likelihood-discriminability, γ, empirical cumulative distribution function of individual-level posterior means. Within-subject comparison between DfD and DfD+forced.

H Instructions to Subjects

H.1 Stage 1 Instructions

Subjects in all treatments, were given the following instructions prior to Stage 1.



Instructions: Bonus

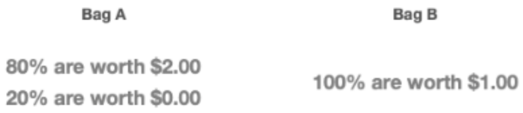
Please pay close attention to the following instructions. We will ask you **comprehension questions** about the instructions. Anyone who answers these questions correctly **the first time** will receive a $0.25 bonus.

Part 1 Instructions: Digital Bags

1. There will be two Parts to this experiment.

2. Part 1 will consist of **several Tasks**. In each Task you will choose between two **digital bags** -- Bag A and Bag B

3. Each bag contains **20 coins** and each coin is worth some amount of money to you as a **bonus**.



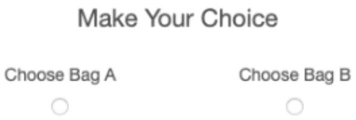
Bag A	Bag B
80% are worth $2.00 20% are worth $0.00	100% are worth $1.00

Example: *In the example above, 80% of the coins in Bag A (i.e. 16 coins) are worth $2, while 20% of the coins (4 coins) are worth $0. On the other hand, 100% of the coins in Bag B are worth $1.*

4. No coin in any bag is worth more than $35.

5. We will **randomly digitally draw one coin** from one of the two bags (Bag A or Bag B), and use that coin to determine how much money to add to your bonus. Each coin in the bag is **equally likely** to be drawn.

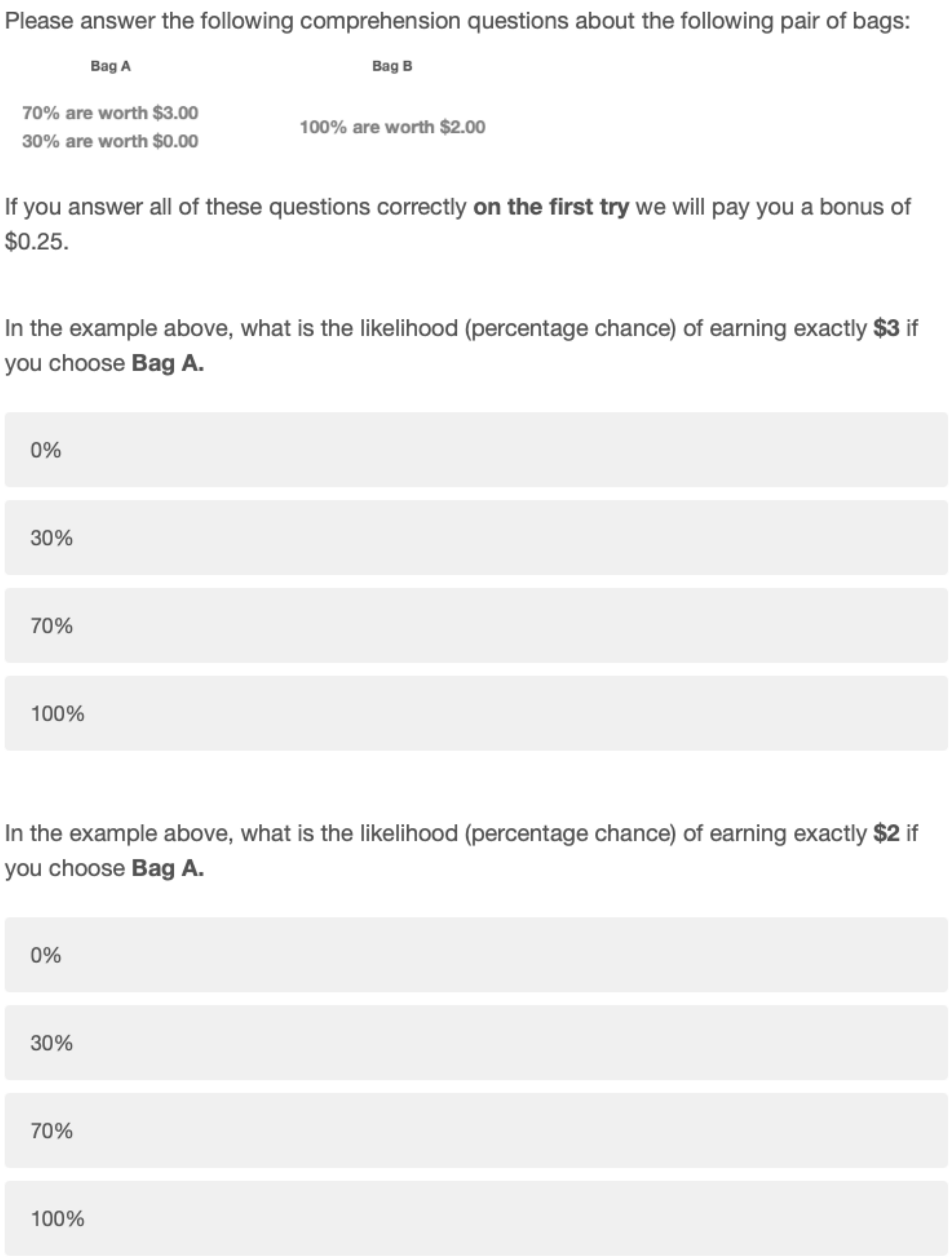
6. Your job is to decide **which bag** you would like us to randomly draw a coin from for your payment, by clicking one of the two buttons as in the example below.



Make Your Choice

Choose Bag A	Choose Bag B
○	○

Example: In the earlier example, if you choose Bag A there is an 80% chance you earn $2 and a 20% chance you earn $0. However, if you choose Bag B there is a 100% chance you earn $1.



Please answer the following comprehension questions about the following pair of bags:

Bag A	Bag B
70% are worth $3.00 30% are worth $0.00	100% are worth $2.00

If you answer all of these questions correctly **on the first try** we will pay you a bonus of $0.25.

In the example above, what is the likelihood (percentage chance) of earning exactly **$3** if you choose **Bag A.**

0%

30%

70%

100%

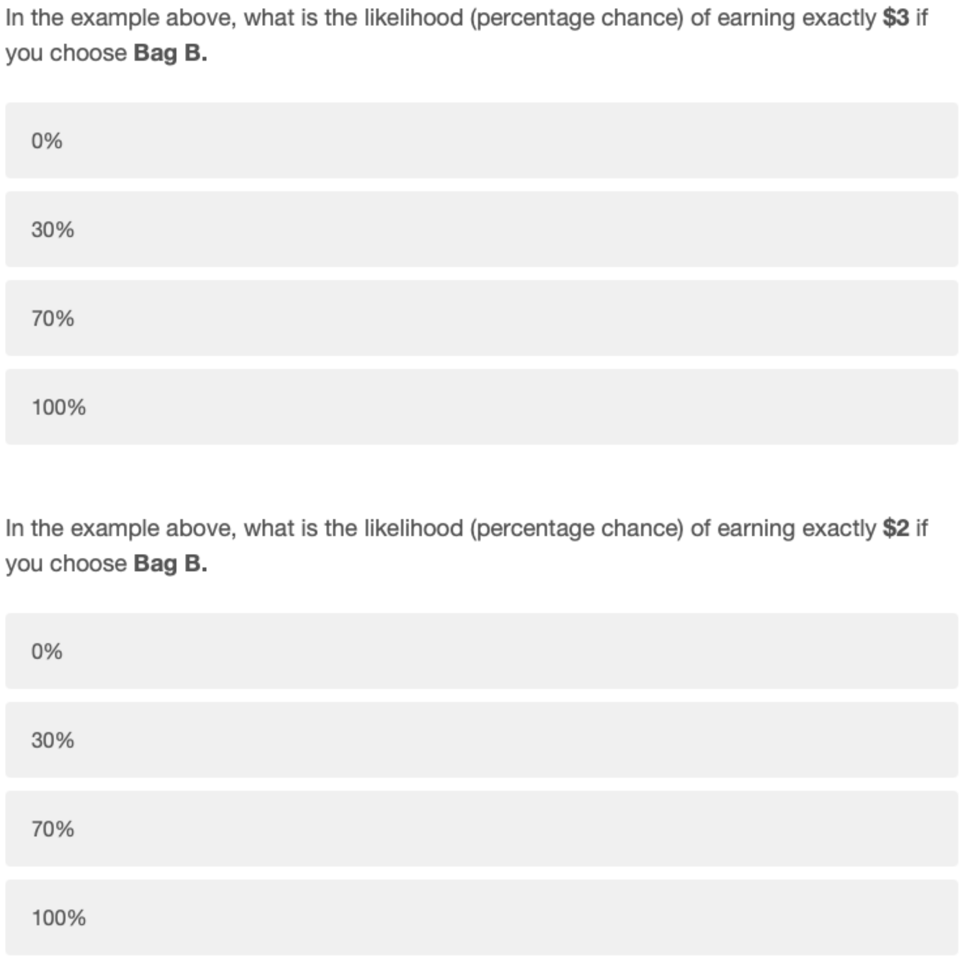
In the example above, what is the likelihood (percentage chance) of earning exactly **$2** if you choose **Bag A.**

0%

30%

70%

100%



In the example above, what is the likelihood (percentage chance) of earning exactly **$3** if you choose **Bag B.**

0%

30%

70%

100%

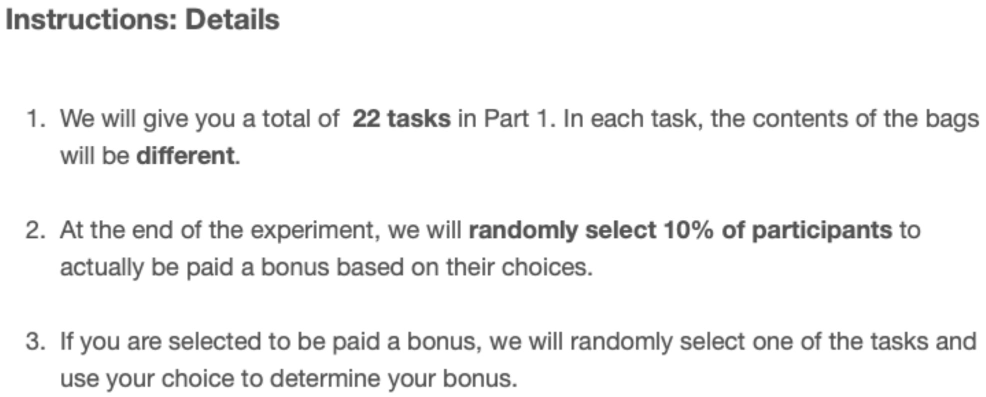
In the example above, what is the likelihood (percentage chance) of earning exactly **$2** if you choose **Bag B.**

0%

30%

70%

100%

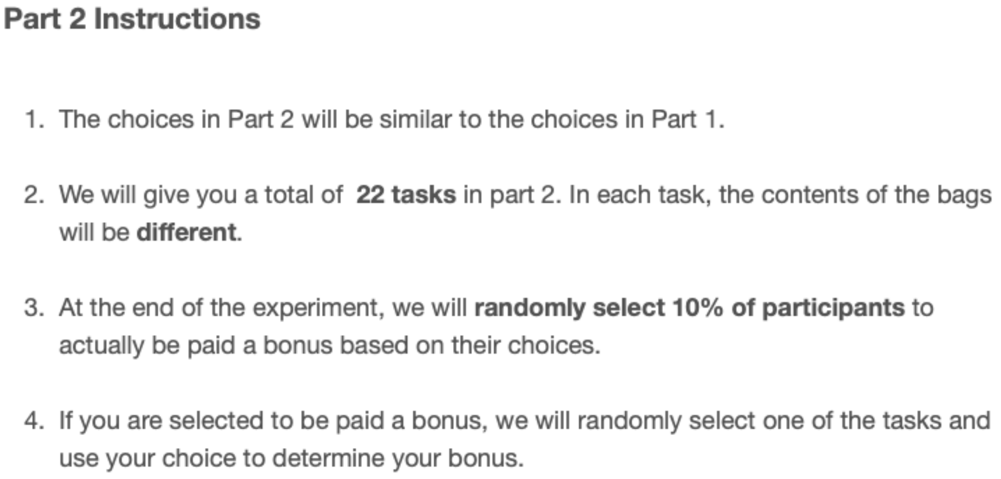


Instructions: Details

1. We will give you a total of **22 tasks** in Part 1. In each task, the contents of the bags will be **different**.
2. At the end of the experiment, we will **randomly select 10% of participants** to actually be paid a bonus based on their choices.
3. If you are selected to be paid a bonus, we will randomly select one of the tasks and use your choice to determine your bonus.

H.2 Stage 2 Instructions

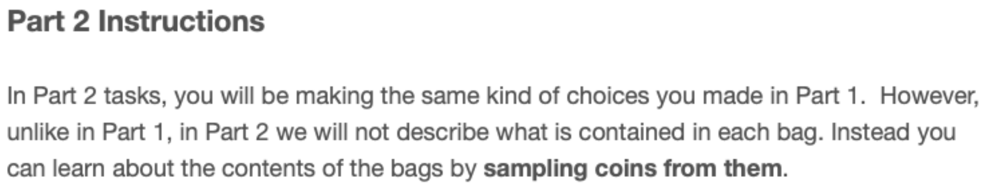
In Stage 2, subjects assigned to the DfD treatment were given the following instructions:



Part 2 Instructions

1. The choices in Part 2 will be similar to the choices in Part 1.
2. We will give you a total of **22 tasks** in part 2. In each task, the contents of the bags will be **different**.
3. At the end of the experiment, we will **randomly select 10% of participants** to actually be paid a bonus based on their choices.
4. If you are selected to be paid a bonus, we will randomly select one of the tasks and use your choice to determine your bonus.

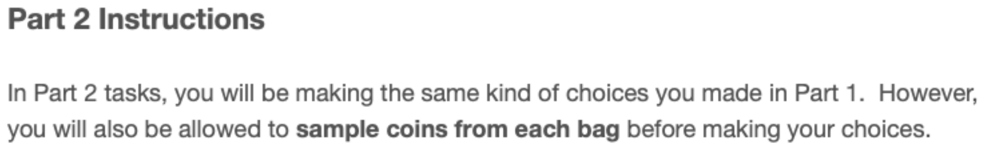
Subjects assigned to DfE or DfE+forced were initially given the following instructions:



Part 2 Instructions

In Part 2 tasks, you will be making the same kind of choices you made in Part 1. However, unlike in Part 1, in Part 2 we will not describe what is contained in each bag. Instead you can learn about the contents of the bags by **sampling coins from them**.

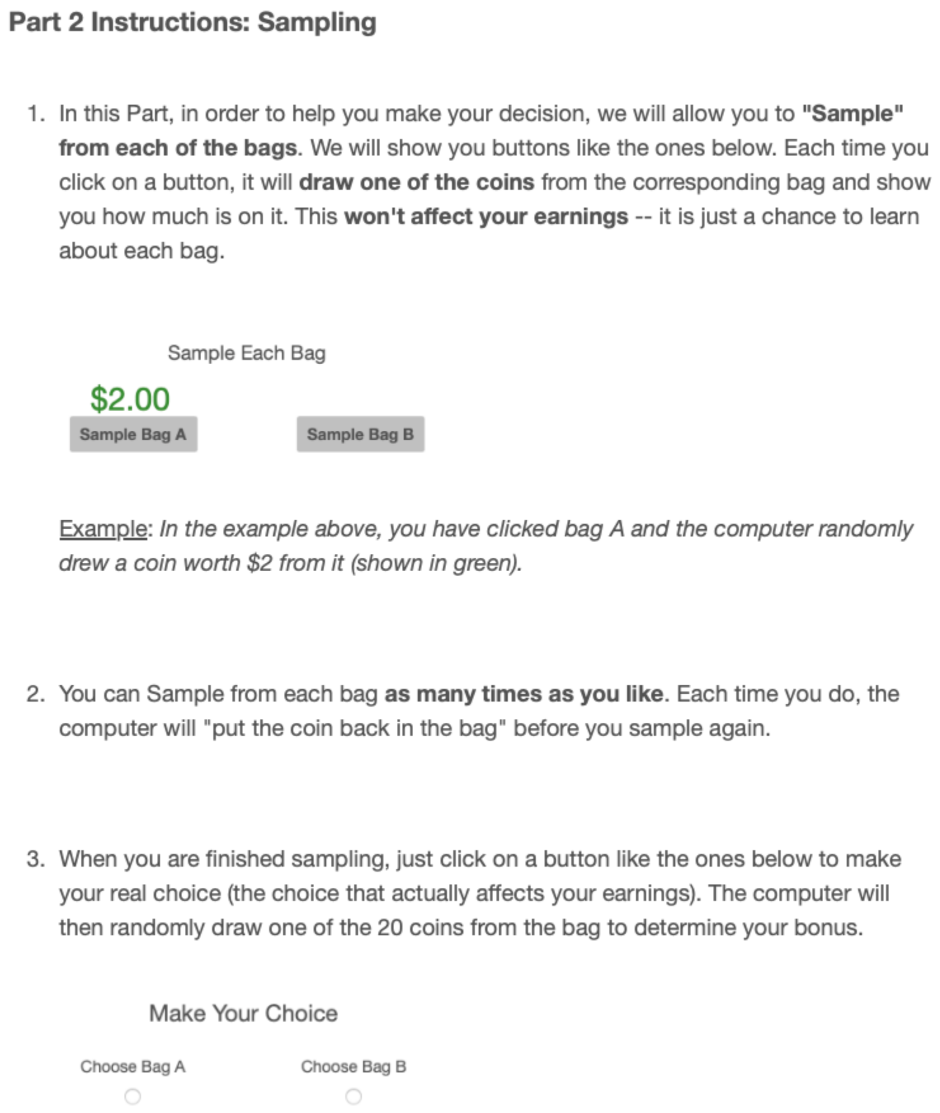
Subjects assigned to DfD+forced or DfD+forced were initially given the following instructions:



Part 2 Instructions

In Part 2 tasks, you will be making the same kind of choices you made in Part 1. However, you will also be allowed to **sample coins from each bag** before making your choices.

After this, subjects in DfE or DfD+free were given the following instructions:



Part 2 Instructions: Sampling

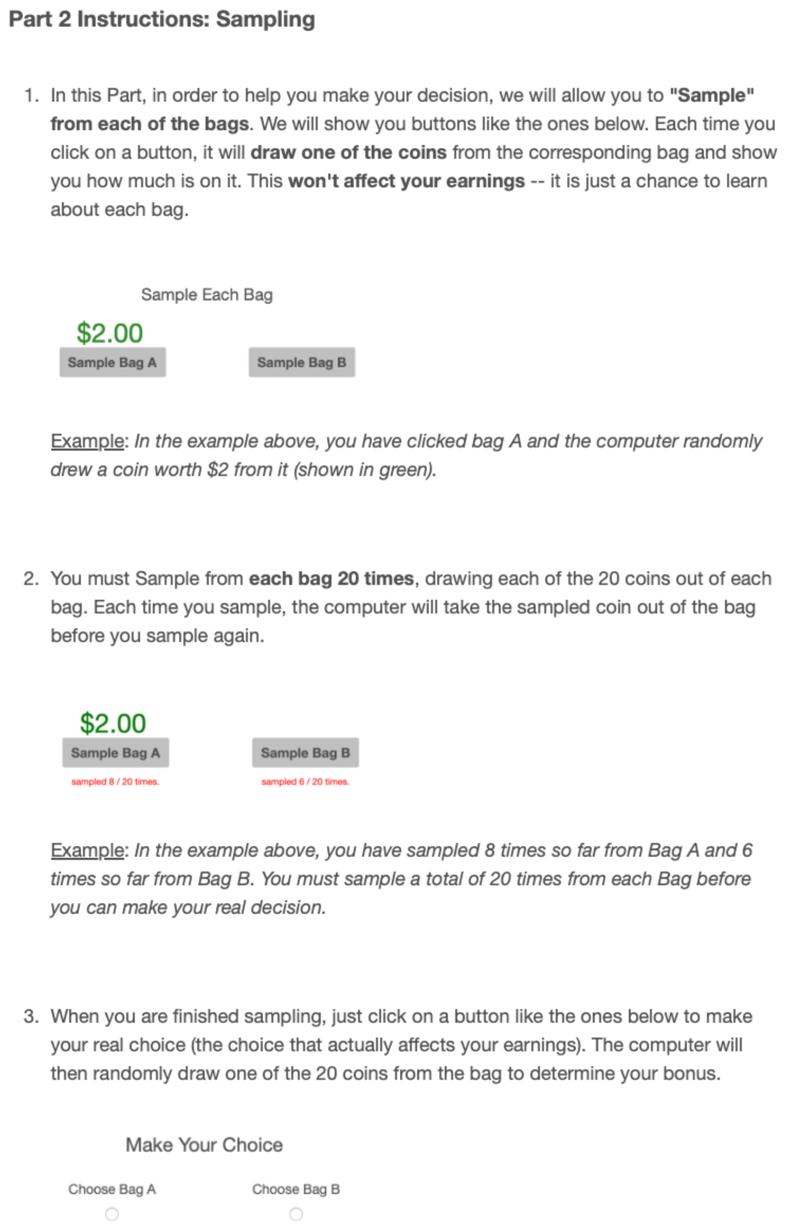
1. In this Part, in order to help you make your decision, we will allow you to **"Sample" from each of the bags**. We will show you buttons like the ones below. Each time you click on a button, it will **draw one of the coins** from the corresponding bag and show you how much is on it. This **won't affect your earnings** -- it is just a chance to learn about each bag.

Example: *In the example above, you have clicked bag A and the computer randomly drew a coin worth $2 from it (shown in green).*

2. You can Sample from each bag **as many times as you like**. Each time you do, the computer will "put the coin back in the bag" before you sample again.

3. When you are finished sampling, just click on a button like the ones below to make your real choice (the choice that actually affects your earnings). The computer will then randomly draw one of the 20 coins from the bag to determine your bonus.

while subjects in DfE+forced or DfD+forced were instead given the following instructions:



Part 2 Instructions: Sampling

1. In this Part, in order to help you make your decision, we will allow you to **"Sample" from each of the bags**. We will show you buttons like the ones below. Each time you click on a button, it will **draw one of the coins** from the corresponding bag and show you how much is on it. This **won't affect your earnings** -- it is just a chance to learn about each bag.

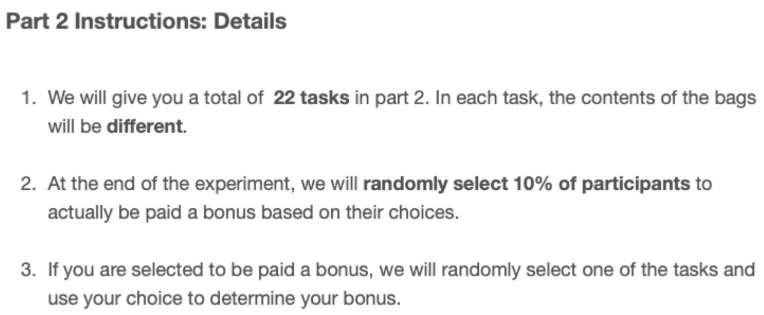
Example: *In the example above, you have clicked bag A and the computer randomly drew a coin worth $2 from it (shown in green).*

2. You must Sample from **each bag 20 times**, drawing each of the 20 coins out of each bag. Each time you sample, the computer will take the sampled coin out of the bag before you sample again.

Example: *In the example above, you have sampled 8 times so far from Bag A and 6 times so far from Bag B. You must sample a total of 20 times from each Bag before you can make your real decision.*

3. When you are finished sampling, just click on a button like the ones below to make your real choice (the choice that actually affects your earnings). The computer will then randomly draw one of the 20 coins from the bag to determine your bonus.

Finally, all subjects were given these instructions prior to the beginning of Stage 2:



Part 2 Instructions: Details

1. We will give you a total of **22 tasks** in part 2. In each task, the contents of the bags will be **different**.

2. At the end of the experiment, we will **randomly select 10% of participants** to actually be paid a bonus based on their choices.

3. If you are selected to be paid a bonus, we will randomly select one of the tasks and use your choice to determine your bonus.