Is Prospect Theory Really a Theory of Choice?*

Ranoua Bouchouicha[1], Ryan Oprea[2], Ferdinand M. Vieider[1], and Jilong Wu[1]

[1]*RISLαβ, Department of Economics, Ghent University*
[2]*Department of Economics, University of California Santa Barbara*

December 10, 2024

Abstract

Using new experiments we show that the fourfold pattern of risk attitudes, "the most distinctive implication of prospect theory" (Tversky and Kahneman, 1992, p. 279), does not actually appear in lottery choice. Instead, it is a special phenomenon of explicitly elicited lottery valuations. Conducting a meta-analysis, we show that the same is true of decades worth of previous data, calling into question prospect theory's ability to predict risky choice. We show theoretically (and verify empirically) that both probability weighting and its failure to produce the fourfold pattern in lottery choice are a consequence of cognitive imperfections in evaluations of lotteries.

1 Introduction

In this paper we experimentally show that *prospect theory* – the most influential behavioral theory of decision making under risk (Kahneman and Tversky, 1979, Tversky and Kahneman, 1992, Wakker, 2010) – doesn't actually predict how people *choose between* lotteries. Rather, it predicts a behavior that turns out to be very different: the way people explicitly assign monetary values to (i.e., assign certainty equivalents to) lotteries. What's more, meta-analytically re-evaluating data from decades of prior research, we show that *this has always been true* in the data but has gone unnoticed in most studies for interesting reasons.[1] We offer (and experimentally validate) a theoretical explanation for this empirical rift between lottery choice and lottery valuation – an explanation that sheds significant light on the cognitive foundations of the behaviors described by

*Ranoua Bouchouicha and Ferdinand Vieider gratefully acknowledge funding from the Research Foundation—Flanders (FWO) under the project "Causal Determinants of Preferences" (G008021N). A previous version of this manuscript was circulated under the title "Choice lists and 'standard patterns' of risk-taking". We are indebted to Mohammed Abdellaoui, Ben Enke, Thomas Epper, Paul Feldman, Olivier L'Haridon, Pietro Ortoleva, Larry Samuelson, and Christian Walter for helpful comments and discussions. All errors remain our own.

[1]The only exception we are aware of is Harbaugh, Krause and Vesterlund (2010), discussed in more detail below.

prospect theory and risky choice more generally.

Our investigation is focused on the *fourfold pattern of risk attitudes*, an empirical regularity that Tversky and Kahneman (1992) call "the most distinctive implication of prospect theory" (p. 279). The fourfold pattern summarizes prospect theory's main predictions about attitudes towards unmixed lotteries, suggesting that people tend to be risk averse towards high probability gains and risk seeking towards low probability gains, but tend to have exactly the reverse risk postures in each case towards losses. The fourfold pattern is foundational to prospect theory both because it is the primary expression of several of the theory's key components (including, notably, probability weighting and likelihood insensitivity), but also because it is responsible for some of the theory's most important empirical rationalizations, e.g., its ability to account for the coexistence of gambling and insurance – a foundational issue in decision making under risk (Vickrey, 1945; Friedman and Savage, 1948; Markowitz, 1952).

Experimental evidence. Our first contribution is a new experiment that carefully tests for and compares the qualitative incidence and quantitative intensity of the fourfold pattern in the two canonical environments studied in the literature. First, we explicitly elicit subjects' values (certainty equivalents) for a series of lotteries in the standard way used in the literature, *multiple price lists* (MPLs). We call these our *Valuation* tasks. Second, we decompose the individual binary choices contained in the MPLs from the Valuation tasks into individual *binary choices* between certain payments and lotteries. We call these our (binary) *Choice* tasks. Because these tasks include *identical component lottery choices*, this design maximizes the comparability of the two tasks and maximies the likelihood of finding isomorphic behavior in the two cases. Indeed because of the nature of our design, under the lens of prospect theory, these are *identical tasks.*

Nonetheless, we report dramatic differences in the predictive accuracy of prospect theory (i.e., the appearance of the fourfold pattern) in these two settings. In our Valuation tasks we find highly conventional evidence of the fourfold pattern as is typically found in elicitations of certainty equivalents in the prospect theory literature (e.g., Tversky and Kahneman, 1992; Gonzalez and Wu, 1999; Bruhin, Fehr-Duda and Epper, 2010; L'Haridon and Vieider, 2019). In our Choice tasks by contrast, the fourfold pattern collapses: risk seeking disappears for small probability gains and risk aversion disappears for small probability losses. Thus, in direct lottery Choice, the fourfold pattern is replaced by a consistent twofold pattern: people become simply risk averse in gains and risk-seeking in losses. What's more, subjects are far more sensitive to variation in probabilities in Choice than Valuation, suggesting their Choice behavior is far less distorted by probability weighting (and much closer to expected utility theory) than Valuation behavior is.

A tempting explanation is that (perhaps) this failure occurs because binary Choice is more difficult than Valuation, preventing a true, latent fourfold pattern from expressing itself due to an increase in, e.g., noisy behavior. However, the data is starkly inconsistent with such an explanation – if anything it suggests the opposite. While it is true that we find some evidence of increased "multiple switching" in binary Choice, these inconsistencies are tightly concentrated in a likely region of indifference near expected value (as documented by Cubitt, Navarro-Martinez and Starmer, 2015, and Agranov and Ortoleva, 2017), where we should expect indifference-driven noise. By contrast, we find (i) much *lower* rates of preference consistency *between* lotteries and (ii) significantly *weaker* test-retest-reliability across decisions, in Valuation relative to Choice. Preferences measured in Choice are thus far more internally coherent and consistent on net than Valuation, suggesting that (if anything) Valuation (and by extension the fourfold pattern that uniquely arises in it) may provide a more artificial portrait of 'true risk preferences' than Choice. A second tempting explanation is that our findings are driven by loss aversion set off by changes in the reference point across environments. In Online Appendix A.3 we show that prospect theory cannot rationalize the data in this way because no consistent loss aversion parameter can reproduce these results.

Meta analysis. Why do we find this fundamental failure of prospect theory on its "home turf" of direct lottery Choice (i.e., the domain the theory was ultimately designed to explain)? After all, researchers have been empirically estimating prospect theory functionals on this same kinds of task for *decades*. Is it possible there is something unusual about our experiment that leads to this finding? To investigate, we conduct a large-scale meta-analysis in which we search for all prior papers that (i) use data from standard Valuation or Choice tasks and (ii) estimate prospect theory parameters on those data. Using these estimates, we are able to calculate certainty equivalents from 141 prior prospect theory estimates and compare the incidence and intensity of the fourfold pattern implied by prior Valuation and Choice estimates.

Remarkably, we find almost universal evidence of the exact same pattern in the prior literature that we found in our experiment. Estimates from prior Valuation tasks overwhelmingly produce certainty equivalents that exhibit the full fourfold pattern. By contrast estimates from prior Choice tasks virtually never do. Instead (as in our data) estimated certainty equivalents from prior Choice tasks show only a twofold pattern: uniform risk aversion in gains and uniform risk seeking in losses. This failure of prospect theory to describe binary choice shows up in aggregate means and meta-analytic estimates, but it also shows up almost without exception in individual studies. As in our own data, we also find far weaker evidence of likelihood insensitivity in Choice than Valuation in previous estimates, meaning probability weighting itself is substantially overstated in Valuation relative to direct Choice.

This analysis suggests that we have in some sense "known" that the fourfold pattern (and thus prospect theory) doesn't really predict binary choice all along, but the literature has somehow been unaware of this fact. How could this be? The answer is interesting and suggests important methodological lessons for the field. Simply put, we think this occurred because of an emphasis in the literature on estimates of the individual *theoretical* components of prospect theory relative to the core *empirical* patterns that theory was designed to explain. Prospect theory consists of two theoretical components that *jointly* determine risk attitudes: an inverse-S shaped probability weighting function and an S-shaped utility function (analogous to a reference-dependent version of a standard utility function). Estimates on both Choice and Valuation data have produced structural estimates consistent with inverse-S shaped probability weighting (Wu and Gonzalez, 1996; Stott, 2006; Conte, Hey and Moffatt, 2011), but this is not sufficient to establish the fourfold pattern because of the confounding influence of utility curvature, which is typically estimated as much more severe in Choice than Valuation tasks.[2] The literature has typically not put these estimates together to assess their joint implications for the (fourfold) pattern the theory was designed to accommodate, or conducted reduced form analysis on the raw data that would have revealed the pervasive failure of this key pattern to arise in binary Choice.

Implication for prospect theory. Of course the reason for the literature's failure

[2]Indeed, in a technical sense probability weighting is also insufficient to yield the fourfold pattern because, as Wakker (2010) emphasizes, inverse-S shaped probability weighting, per se, is perfectly consistent with uniform risk aversion or risk seeking.

to notice this near-universal failure of the fourfold pattern in binary Choice suggests a natural rejoinder to our central claim (i.e., that prospect theory doesn't predict Choice behavior). In particular, mathematically, prospect theory functionals are flexible enough to accommodate the failure of the fourfold pattern since they can technically be parameterized to produce certainty equivalents consistent with a two-fold (or even a one- or three-fold!) pattern of risk. Indeed, this flexibility remains even if those functionals yield estimates of the inverse-S shaped probability weighting function described in orthodox prospect theory (i.e., because sufficient utility curvature can nonetheless "break" the pattern in such cases). From this perspective, one might be tempted to argue, the literature "missed" the failure of the fourfold pattern because a failure of the pattern is in fact (in a technical sense) perfectly consistent with prospect theory.

To this we have three responses. First, to the degree one claims this, prospect theory is a nearly content-less theory (at least in the domain of unmixed lotteries) since it "predicts" nearly any correspondence between probabilities, gain/loss framings and risk attitudes. To the degree this is true, prospect theory is nearly unfalsifiable and hardly a theory at all. Second (and unsurprisingly), the prospect theory literature *does not* in fact claim that prospect theory is content-less in this way. Rather, it claims distinctive, predictable empirical implications, the most important of which is the fourfold pattern itself.[3,4] If the theory has any empirical content, the failure of the fourfold pattern to appear in Choice is a failure of its central, falsifiable empirical claim regarding attitudes towards unmixed lotteries. Finally prospect theory is starkly inconsistent with and (despite its flexibility) cannot accommodate our (and, retrospectively, the literature's) finding of completely different patterns of risk taking in Choice than in Valuation.[5] This pattern clearly

[3]As mentioned above in one of the foundational papers in the literature, Tversky and Kahneman (1992) claim the fourfold pattern as prospect theory's "most distinctive implication." In another seminal paper in the literature, Wu and Gonzalez (1996) describe the fourfold pattern as one of the "critical empirical regularities that any good descriptive model should accommodate [...]: risk aversion for most gains and low probability losses, and risk seeking for most losses and low probability gains" (p. 1676).

[4]Of course, prospect theory also predicts likelihood insensitivity, even if that insensitivity fails to "flip" risk postures as described by the fourfold pattern. However it is important to emphasize that we find (in both our new data and the meta-analysis) that *likelihood insensitivity itself* is dramatically weaker in Choice than in Valuation (with a significant amount of likelihood *over*-sensitivity in the former case). This weakening is not predicted by and in fact is starkly inconsistent with the predictions of prospect theory.

[5]We can immediately exclude a natural prospect-theory-based explanation that may occur to readers: endogenous reference points (Hershey and Schoemaker, 1985). Explaining the difference between Choice and Valuation via reference dependence would require people to have systematically different loss aversion parameters across the probability spectrum. While endogenous reference points in probability equivalents as discussed by Hershey and Schoemaker (1985) may arguably have contributed to Tversky and Kahneman's (1992) choice to use Valuation tasks to measure PT functionals, it has recently been

demands explanation and we cannot claim that we in any sense *understand* decision under risk on the basis of any theory that fails to provide such an explanation.

Mechanism and experimental validation. If prospect theory cannot explain why the fourfold pattern appears in Valuation but not in Choice, what theory can? We propose and experimentally validate an explanation that builds on a groundswell of recent evidence suggesting that key patterns from prospect theory are a result of imperfections in humans' ability to precisely represent and aggregate information (so called "noisy coding models", e.g., Khaw, Li and Woodford, 2021; 2023; Vieider, 2023; Oprea, 2024*b*; Oprea and Vieider, 2024). Noisy coding explains probability weighting (and by extension the fourfold pattern) as arising not due to preferences (described by a probability weighting function), but rather due to the noisy way decision makers represent probabilities (or certainty equivalents) internally, and the way they compensate for this imprecision by distorting valuations towards a prior in a Bayesian fashion, leading to compressed responses to probabilities (Khaw, Li and Woodford, 2023; Vieider, 2023).

Our model argues that explicit Valuation (à la multiple price lists) requires decision makers to evaluate lotteries in a sequential manner that intensifies this cognitive noise relative to binary Choice. This results in an intensification of compression-driven likelihood insensitivity and an upward shift in the prior, that can generate and account for the appearance of a fourfold pattern under Valuation but not under Choice. It also makes secondary and tertiary predictions that are matched by our finding of higher likelihood insensitivity and noisier behavior in the former than the latter in our data.

The explanation also comes with a distinctive prediction – a recipe for causing the gap between binary Choice and Valuation to disappear that sets up a prospective test. Because the model is rooted in the hypothesis that Valuation forces decision makers to evaluate the lottery prior to searching through potential values, it predicts that forcing subjects to evaluate lotteries in a similar, sequential way in binary Choice will cause Choice and Valuation behaviors to converge. We run an experiment that does just this: subjects in Choice are shown the lottery ahead of time and told they will have to compare it to a sequence of later values in a series of binary decisions. Just as the theory prescribes and predicts, this almost completely removes the gap between Choice and Valuation.

shown that loss aversion cannot actually account for the discrepancy (Feldman and Ferraro, 2023).

Our results thus both strongly reinforce the hypothesis that both probability weighting and the fourfold pattern are a consequence of cognitive frictions (not prospect theory preferences) and explains an important mystery as to the circumstances under which they distort and describe choice. Our findings also suggest that the fourfold pattern is unlikely to be a description of the unusual factors shaping risk preferences, as in many standard interpretations of it. Instead, the pattern is a description of the effects of cognitive distortions that prevent decision-makers from revealing their true preferences for risk when asked to value lotteries. There is a long history of ambivalence in the literature on the welfare implications of lottery anomalies and the theories built to describe them. Our findings add to a chorus of recent findings suggesting that anomalies (like the fourfold patterns) and the theories built to describe them (like prospect theory) contain very little welfare-relevant content.

Relation to prior work. Our work is broadly related to a long literature in psychology and economics showing that decision under risk often fails *procedural invariance*: measured risk preferences often depend in systematic ways on the method of elicitation used. These effects have been documented at least since Slovic (1964), and have periodically resurfaced in the literature in many different contexts (e.g., Hershey and Schoemaker, 1985; Crosetto and Filippin, 2015; Mata et al., 2018; Friedman et al., 2017; Zhou and Hey, 2018; Friedman et al., 2022), often yielding evidence broadly consistent with ours. In a recent paper McGranaghan et al. (2024) show that Choice and Valuation generate often very different evidence in favor of the common ratio effect, and argue that this is driven by the differential effects of noise in the two settings. Relative to much of this literature, the procedural invariance violations we document are particularly striking because they appear in Valuation and Choice tasks that consist of an *identical* set of constituent decisions. What's more, unlike most of this literature, our study is designed around a set of tasks that allow us to directly measure the implications of procedural invariance for the core predictions of prospect theory.

Closely related is a literature on "preference reversals" which shows that apparent preferences over lotteries often flip when elicited via binary Choice versus willingness-to-accept (WTA) elicitations (i.e. selling prices; Lichtenstein and Slovic, 1971; Grether and Plott, 1979).[6] Modern versions of prospect theory allow researchers to explain such preference

[6]The early literature has also documented preference reversals using willingness-to-pay (buying

reversals via loss aversion, due to the differing implicit endowments in Choice versus WTA tasks (Schmidt, Starmer and Sugden, 2008).[7] By contrast, the violations of procedural invariance we document involve identical choices without variation in endowments, meaning they cannot be explained in this way. Indeed, we show empirically that even allowing for endogenous reference points cannot account for our results via prospect theory.

Substantively, the closest paper to ours is Harbaugh, Krause and Vesterlund (2010) who compare Valuations (elicited using the BDM mechanism) made by subjects to a collection of four binary Choices between (i) lotteries corresponding to each of the conditions of the fourfold pattern and (ii) their expected value. They recover the fourfold pattern in Valuation but find Choice behavior that is "indistinguishable from random choice" (p. 595). The paper is seminal for introducing evidence of differences between Valuation and Choice in expressions of the fourfold pattern, but its findings in Choice are difficult to interpret given the limited subject-level data and increasing evidence in the literature that noisy behavior should be *expected* in choices that are close to indifference. We gather significantly sharper and more interpretable evidence by (i) collecting large amounts of Choice data sufficient to actually estimate certainty equivalents and by (ii) explicitly designing our experiment to feature identical constituent choices in Choice and Valuation, facilitating particularly direct comparisons between the two settings. Our findings are broadly consistent with theirs, but our methodological improvements allow us to show that Choice behavior is in fact overall *more* stable than Valuation behavior, except in regions of the parameter space close to indifference.

Methodologically, the closest work to ours is a recent series of papers that compares Valuation and Choice behavior by (i) having subjects make decisions in multiple price lists and (ii) having subjects also make decisions in the "stacked" choices of the lists

prices). These preferences reversals are, however, much less frequent and less systematic than for willingness-to-accept, and have been ascribed to white noise in the subsequent literature (Loomes, 2005; Alós-Ferrer, Buckenmaier and Garagnani, 2020).

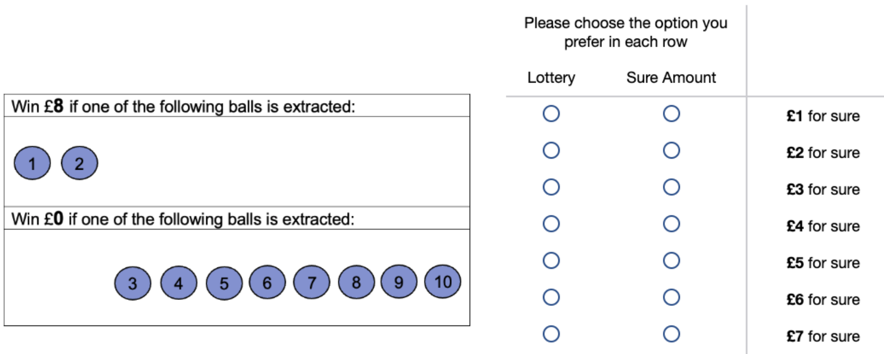
[7] The early literature explained preference reversals via the "compatibility" hypothesis that the weight attributed to a given descriptive aspect of a lottery (e.g., payoffs versus probabilities) increases in its similarity with the dimension in which preferences are elicited (Slovic, Griffin and Tversky, 1990; Tversky, Slovic and Kahneman, 1990). This hypothesis, however, cannot explain the large differences in reversals observed between willingness-to-pay and willingness-to-accept. More importantly for our purposes, it cannot explain our experimental results in Section 4, which suggest a very different explanation for failures of procedural invariance. Regardless, the compatibility hypothesis does not ease the problems with prospect theory we document: whatever the explanation, reversals violate procedural invariance and hence are inconsistent with the predictions of prospect theory.

individually and in a random order. Lévy-Garboua et al. (2012) compared choices in the task of Holt and Laury (2002) to the underlying binary decisions, and documented higher levels of risk aversion and increased noise in binary choices. Freeman, Halevy and Kneeland (2019) also compared risky choices obtained from a choice list to risky choices in a *single* binary choice, and found significantly more risk aversion in binary choice. They ascribed this effect to the random incentive mechanism (however, see Freeman and Mayraz, 2019, for an account contradicting this explanation). Revisiting this issue, Brown and Healy (2018) conclude that incentive-compatibility is guaranteed by randomly paying one of several choices presented in a binary choice mechanism in random order. These papers examine only one multiple price list and one collection of corresponding binary Choice problem. One of our main contributions relative to these papers is to study *multiple* choice lists in Valuation, and multiple corresponding collections of binary Choice tasks. This allows us to paint a much richer picture of the factors driving decisions, and to use the resulting patterns to test the predictive ability of prospect theory.

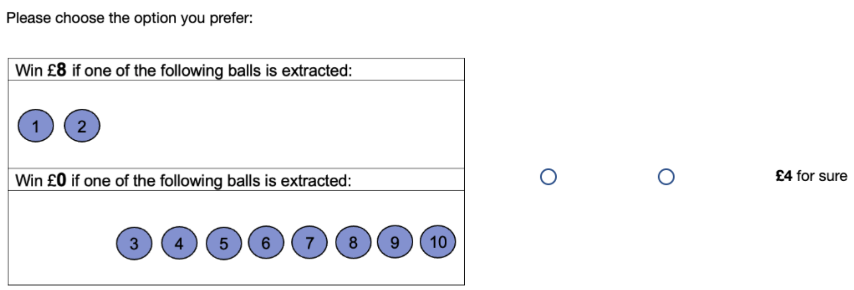
Our paper also contributes to a series of recent papers challenging prospect theory's ability to account for empirical paradoxes that were once considered to fall within its remit. Sydnor (2010) used a numerical calibration exercise to show that prospect theory parameters as measured in experiments are unable to account for the widespread overinsurance of modest risks. Oprea (2024*b*) showed that the fourfold pattern (measured via Valuation) is reproduced in evaluations of deterministic options that are represented with the same level of complexity as in choice lists under risk. Chapman et al. (2023*b*) document that different measures of loss aversion are unrelated to the gap between willingness-to-accept and willingness-to-pay in a large representative sample, contrary to the standard PT explanation. Chapman et al. (2024) present findings showing that more than half of the American population accepts mixed gain-loss gambles with negative expected value, thus casting doubt on the pervasiveness of loss aversion.

Finally, our paper relates to a growing movement in economics that attempts to reinterpret anomalies under the lens of domain-general cognitive frictions rather than domain-specific preferences or errors (Enke, 2024; Enke et al., 2024; Oprea, 2024*a*). Most relevant for our purposes is a growing literature on *noisy cognition*, which explains anomalies as growing out of the imprecise ways constrained brains represent information and the biases produced by Bayesian responses to this noise. These models can explain many aspects of prospect theory, including probability weighting (Vieider, 2023; Khaw, Li and Woodford, 2023; Frydman and Jin, 2023) and tests of these models have been highly empirically successful (Natenzon, 2019; Prat-Carrabin and Woodford, 2022; Oprea and Vieider, 2024). We use a model in this class to explain our main findings, and test that model using a novel experiment, adding to this accumulation of evidence.

2 Experiment



(a) *Valuation* Condition



(b) *Choice* Condition

Figure 1: Screenshots of the two treatments

Screenshots from the Gain treatments. Panel (a) shows a screenshot of a typical Valuation list for a lottery yielding £8 with a 20% probability, or else 0. Panel (b) shows one binary choice extracted from that same list as it was presented in the Choice treatment.

Our first contribution is a new experiment, designed to study how risk taking differs under *lottery choice* (which we call Choice tasks) versus *lottery valuation* (Valuation tasks). To accomplish this we study experimental subjects' revealed preferences over a collection of two-outcome lotteries that pay some high amount x with probability p, and some lower amount $y < x$ otherwise, using the two most often studied experimental protocols in the literature:

Valuation tasks: By far the most common method for directly eliciting valuations (certainty equivalents) for lotteries is via *multiple price lists* (MPLs). MPLs are (effectively) a series of binary choices "stacked" on top of one another; Figure 1 shows an example (a screenshot from our experiment). On the left is a description of the lottery being valued, and on the right is a series of binary choices between the lottery and an ascending sequence of certain monetary payments. The subject makes a choice in each "row". At the end of the experiment, one row from a random MPL is selected to count for payment. By examining at what sure payment amount the subject switches from preferring the lottery to the sure payment, the researcher obtains an interval estimate of the subject's certainty equivalent (monetary value) for the lottery. Each subject is assigned 21 MPLs, allowing us to assess the fourfold pattern of risk attitudes and some related diagnostic questions (see below).

Choice tasks: In our binary Choice tasks (shown in the bottom panel of Figure 1), subjects observe a lottery and a single sure amount, and simply choose the one they prefer. In the end, one of the choices is randomly selected for payment. Our Choice tasks always consisted of a choice between a lottery and a sure payment, as in the example at the bottom of Figure 1, allowing us to infer the certainty equivalent that rationalizes subjects' choices in a direct way. Each subject is assigned 274 Choice tasks, allowing us to assess the fourfold pattern of risk taking and some related diagnostics (see below).

Comparability: The key to our design is that we selected our Choice tasks to include *all* of the individual choices embedded in each of the 21 MPLs of our Valuation tasks. That is, we simply took the individual rows of each MPL, transformed each into a stand-alone Choice task (between the lottery and one of the sure amounts in each case), and assigned them all to subjects. Subjects were assigned these Choice tasks in an order that randomly mixed both the lottery and the certain payments. However, from a payoff perspective, subjects make identical decisions under an identical payoff scheme in our Valuation and Choice tasks. Under virtually any theory of risk preferences (including prospect theory) they should therefore yield identical patterns of behavior.

Parameters and Lotteries: The primary goal of the experiment is to contrast evidence of the fourfold pattern of risk under Choice and Valuation. The fourfold pattern of risk attitudes is a regularity in which subjects appear (i) risk averse for high probability gains (their certainty equivalents are lower than the lottery's expected value), (ii) risk seeking for low probability gains (certainty equivalents higher than expected value), (iii) risk averse for low probability losses, and (iv) risk seeking for high probability losses. Thus the fourfold pattern is a series of "flips" in risk posture as (i) probabilities go from low to high and (ii) the framing goes from gains to losses.

To measure the pattern, we vary both probability p and the gain/loss framing of payoffs x and y. We do this using a design with both between- and within-subjects variation. Within-subject, for every subject we vary p between 0.1 and 0.9. Between-subject we run both Gain treatments in which $x, y \geq 0$ and a Loss treatments in which $x, y \leq 0$. Parameters are provided for both treatments in Online Appendix Tables A.1 and A.2. Our Gain treatments are incentivized while we followed much of the literature (Wakker and Deneffe, 1996; Abdellaoui, 2000) by using hypothetical incentives in our Loss treatment – in order to avoid common concerns in the literature that integration of payoffs with the initial endowments required to incentivize losses might create distortions in measurement that would confound our inferences (a practice explicitly recommended by Etchart-Vincent and L'Haridon, 2011).[8] Results from our meta-analysis, below, suggest that expressions of the fourfold pattern are not different in hypothetical versus incentivized experiments in the literature. Finally, we varied x and y across tasks to allow us to evaluate relative risk aversion and verify that it varies with stakes in ways typically found in the literature.

In our Valuation tasks, we vary the certain payments against which lotteries were contrasted, s, in steps of £1 between y and x, meaning subjects made 274 "row" choices across the 21 lists. In the Choice tasks, subjects made the same 274 choices individually in a completely randomized order.

Understanding the Design: Relative to some related exercises from the literature, our design has three characteristics that (to our knowledge) have not been combined and that we believe allow for an especially crisp answer to our motivating questions:

- First, unlike the one other study we know of that contrasts the fourfold pattern un-

[8] The debate on whether hypothetical and real incentives produce the same results has recently received new impetus. Several high-profile papers have tested the effects of confronting subjects with real versus hypothetical incentives across a variety of decision-making tasks, including for high stakes. These new, well-powered experiments typically recommend using hypothetical incentives especially in situations that may be hard to incentivize properly (Brañas-Garza et al., 2021; 2023; Enke et al., 2023; Gneezy et al., 2024).

der Choice and Valuation (Harbaugh, Krause and Vesterlund, 2010), we designed the Choice and Valuation tasks to measure an identical object: the certainty equivalent. Instead of giving subjects one binary Choice task for each lottery, we gave them a rich set, sufficient to infer the certainty equivalent with the same precision afforded in our Valuation task. This produces a particularly strong test of our null hypothesis.

- Second, unlike that prior study (which uses a BDM mechanism), we use MPLs in our Valuation task, allowing us to make our Valuation and Choice tasks truly identical under standard theories. Subjects are literally making the exact same set of choices in the two environments, making our test again especially strong.
- Third, unlike recent papers that follow our basic design strategy (i.e., to contrast MPLs with individual binary choices that reproduce the individual rows of those MPLs; Lévy-Garboua et al., 2012; Freeman, Halevy and Kneeland, 2019; Freeman and Mayraz, 2019), we study a large number of distinct MPLs within-subject in our Valuation tasks and a number of decomposed MPLs in our Choice tasks, while implementing an *identical* payoff mechanism. This means we can, for the first time with such a design, really assess and contrast the performance of prospect theory predictions at the subject level. By following Brown and Healy (2018) and randomly selecting one choice to be paid (from a series of binary choices presented in random order), we make our Choice elicitation incentive-compatible.

These design elements give us an unusually interpretable answer to our main question – and one that arguably works against our hypothesis by maximizing the chances of finding similar behavior in Valuation and Choice.

Implementation Details: all experiments included in this paper were conducted online on Prolific UK within a short time span in the winter of 2022/23. Instructions were provided in short videos, which provided a machine-generated voice-over to slides illustrating the experimental tasks.[9] The main experiment is a between-subjects 2x2 design crossing $\{Valuation, Choice\} \times \{Gains, Losses\}$. In our Gains treatments 327 subjects signed up for the experiment, but we dropped 26 of them who failed to correctly answer some simple comprehension checks after watching the instructional video.

[9]The full video instructions are available at `https://www.youtube.com/@RislabUgent`. The slides are included in ONline Appendix D.

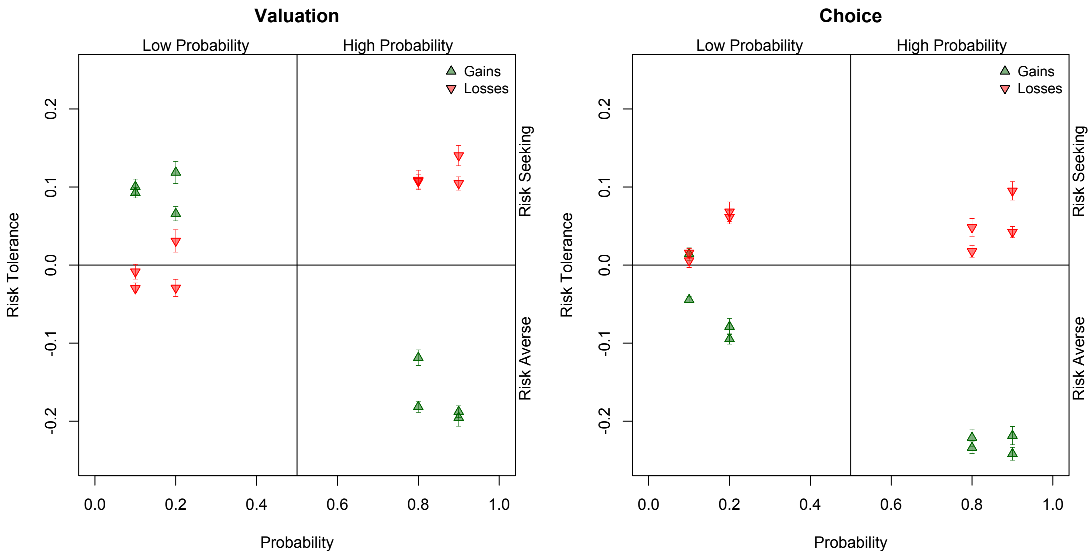


Figure 2: Nonparametric risk taking measures evaluating the fourfold pattern for Valuation and Choice.

The figure plots net certainty equivalents, defined as $p - \frac{\widehat{s}-x}{y-x}$ for gains, and $\frac{\widehat{s}-x}{y-x} - p$ for losses, where $\widehat{s}$ is the certainty equivalent obtained from our data. The left-hand figure plots these measures for Valuation, and the right-hand figure for Choice.

We thus ended up collecting valid responses from 301 individuals (Valuation: N=156; Choice: N=145). The median subject took 40 minutes to finish the experiment (33 minutes in Valuation; 50 minutes in Choice). Each subject was compensated for their time according to Prolific regulations. In addition, each subject had a 1/10 chance to play one randomly selected choice for real money. In our Losses treatments, excluding 10 subjects who did not pass some very basic comprehension tests, we were left with 201 subjects providing valid responses (CE: N=98; BC: N=103). A typical subject took 42 minutes to complete the experiment (31 minutes for the Valuation treatment, 50 minutes for the Choice treatment).

2.1 Results

Figure 2 plots the main results, focusing on low (lower than 0.3) and high (higher than 0.7) probability lotteries, where we expect the fourfold pattern to obtain. On the y-axis we plot net certainty equivalents: the difference between subjects' normalized certainty equivalent and expected value so that positive values indicate risk seeking and negative values risk averse average choices. In particular, we plot the difference between the rate at which subjects choose the risky option (which under well-behaved preferences will approximate a normalized certainty equivalent[10]) and the normalized risk neutral benchmark (in this case simply p). Green upward-pointing triangles plot results from Gain tasks and red downward-pointing triangles results from Loss tasks.

The left hand panel plots results from our Valuation treatment. The results nearly perfectly (94% of the time) reflect the fourfold pattern. Except for one Loss MPL, subjects' certainty equivalents suggest subjects tend towards risk aversion for low probability losses (red downward arrows on the left side of the plot tend to be negative) but become risk seeking for high probability losses (on the right side they become positive). These patterns each flip for the Gains tasks (green upward arrows): at low probabilities (left side) subjects tend to be risk seeking (have positive net certainty equivalents) but at high probabilities (right side) are risk averse (have negative net certainty equivalents).

The right hand panel plots data from the Choice tasks which, recall, are exactly the same set of payoff-relevant decisions. Here, the fourfold pattern disappears almost entirely (in 94% of tasks), replaced by a twofold pattern. In particular, subjects in the Loss treatment are always risk seeking at both low and high probabilities, while subjects in Gain treatment are always risk averse.

Thus, raw analysis of our data suggests that the fourfold pattern is a phenomenon of Valuation but fails to appear in direct Choice. In Section 3.2 below we revisit these data structurally under the lens of prospect theory and show (i) that there is nothing unusual about our data in either case, in the sense that under both Choice and Valuation we estimate standard prospect theory parameters from this data (e.g., we get standard inverse-S shaped probability weighting) and (ii) that these prospect theory parameter estimates *themselves* imply exactly what our non-parametric results show directly: that the fourfold pattern does not occur in Choice as it does in Valuation.[11]

Likelihood insensitivity. We complement our primary analysis of the fourfold pattern with a second, subtler prediction about risky choice made by prospect theory: likelihood

[10]Thus by 'normalized certainty equivalent' we mean in general $\frac{\widehat{s}-y}{x-y}$, where $\widehat{s}$ indicates a 'stochastic switching point', or equivalently, the choice proportion of the risky option (since the sure amounts in our 'lists' are evenly distributed between the extremes of the lottery).

[11]We also find quite typical variations in choice patterns as stakes increase. In particular, we find evidence for increasing relative risk aversion (IRRA) in Gains in both Valuation and Choice tasks, although the level of risk aversion is more pronounced in Choice. In Loss tasks, we replicate the typical finding of constant relative risk aversion in Valuation (Fehr-Duda et al., 2011; Bouchouicha and Vieider, 2017), but we again find IRRA in the size of the loss in Choice. Online Appendix A.4 reports these findings in more detail.

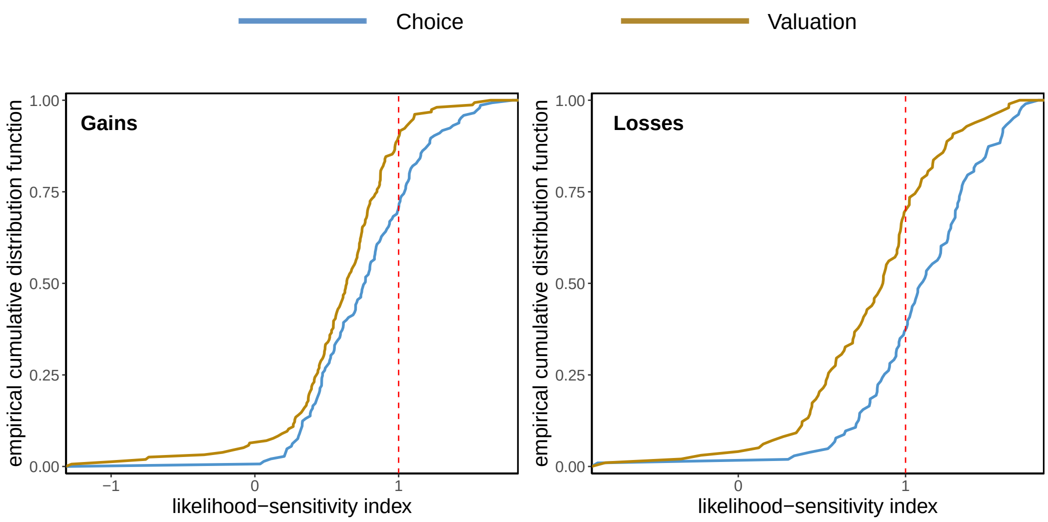


Figure 3: Measured likelihood sensitivity in Valuation and Choice

Empirical cumulative distribution function of the likelihood-sensitivity index, calculated as the average of the differences in normalized CEs for the pairs $(\pm 16, 0.9; \pm 4)$ and $(\pm 16, 0.1; \pm 4)$, $(\pm 16, 0.8; 0)$ and $(\pm 16, 0.2; 0)$, and $(\pm 16, 0.7; 0)$ and $(\pm 16, 0.3; 0)$. **Panel A** presents scatter plots of observations for gains. **Panel B** displays observations for the losses. Vertical solid lines indicate the point of risk neutrality.

insensitivity. While the fourfold pattern describes prospect theory's main prediction about how risk postures change with unmixed lottery characteristics, likelihood insensitivity describes how the intensity of preferences for or against risk change as probabilities do. Prospect theory predicts that subjects' apparent risk preferences respond to increases in probabilities sluggishly, with the intensity of preferences changing less quickly than probabilities themselves do. If a utility function is estimated assuming expected utility theory, the degree of curvature of that function will thus be dependent on the fixed probability used to estimate it (Hershey, Kunreuther and Schoemaker, 1982). Likelihood insensitivity is a necessary (but not sufficient) condition for the fourfold pattern to occur: if subjects are risk seeking for small probability gains (risk averse for small probability losses), their insensitivity to probabilities ensures that this preference 'flips' as probabilities get larger.

To gather a transparent, non-parametric measure of likelihood sensitivity, we calculate the change in the rate at which subjects select the risky lottery at probabilities "mirrored" around 0.5, e.g., 0.9 and 0.1, 0.8 and 0.2 etc. We then normalize these individual differences by the true difference in probabilities to put them on a common scale, and average them to get a subject-wise index of likelihood sensitivity. A subject who weights probabilities linearly (as in, e.g., expected utility theory) will have an index of 1. Subjects with indexes below 1 are likelihood-insensitive, so that their relative risk aversion increases in probabilities for gains (decreases in probabilities for losses), and above 1 likelihood-oversensitive (relative risk aversion decreasing in probabilities for gains, and increasing in probabilities for losses).

Figure 3 plots empirical CDFs of this index from the Valuation and Choice conditions, including separate plots for gains (left panel) and losses (right panel). In Gains, we find that subjects in both Valuation and Choice tend to be likelihood insensitive, but that subjects' behavior is far closer to the expected utility theory benchmark in Choice than Valuation. Indeed, sensitivities in Choice first order stochastically dominate sensitivities in Valuation, and subjects in Choice are over twice as likely to be *likelihood oversensitive* as subjects in Valuation. This same pattern intensifies further in losses where Choice sensitivities strongly first order stochastically dominate Valuation sensitivities. The median subject in Choice is in fact slightly likelihood *oversensitive*, in sharp contrast with the standard prospect theory account. In both cases, Wilcoxon tests suggest ($p < 0.001$) that subjects are significantly more sensitive to likelihoods in Choice than in Valuation.

Once again, then, we find that a core descriptive element of prospect theory is robust in Valuation but weakens (in gains) or disappears altogether (in losses) in direct Choice. As with the fourfold pattern, the likelihood insensitivity predicted by prospect theory is a substantially better description of how subjects explicitly *value* lotteries than how they directly choose between them.

2.1.1 Is Choice or Valuation More Internally Consistent?

Given the claims of the literature, it is natural to wonder whether these results are a consequence of lower data quality in Choice relative to Valuation. After all, Valuation tasks visually organize the same underlying tasks studied in Choice, grouping these tasks by lottery and monotonically by certain payment. Perhaps this leads to more internally coherent behavior under Valuation that better reflects true risk preferences (à la fourfold pattern) than Choice data does.

On net, this is a difficult interpretation to support. First, individual subjects in Choice tasks reveal certainty equivalents that suggest more consistent risk postures across evaluations of different lotteries than subjects do in Valuation tasks. In Figure 4 we plot the rate at which subjects decrease their choices of the lottery as its probability of paying out *increases*, a clear rationality violation, for Valuation and Choice. Clearly, subjects in Choice are less likely to exhibit such inconsistent behavior than subjects in Valuation. By this measure, subjects in Choice are thus clearly less noisy than subjects in Valuation.

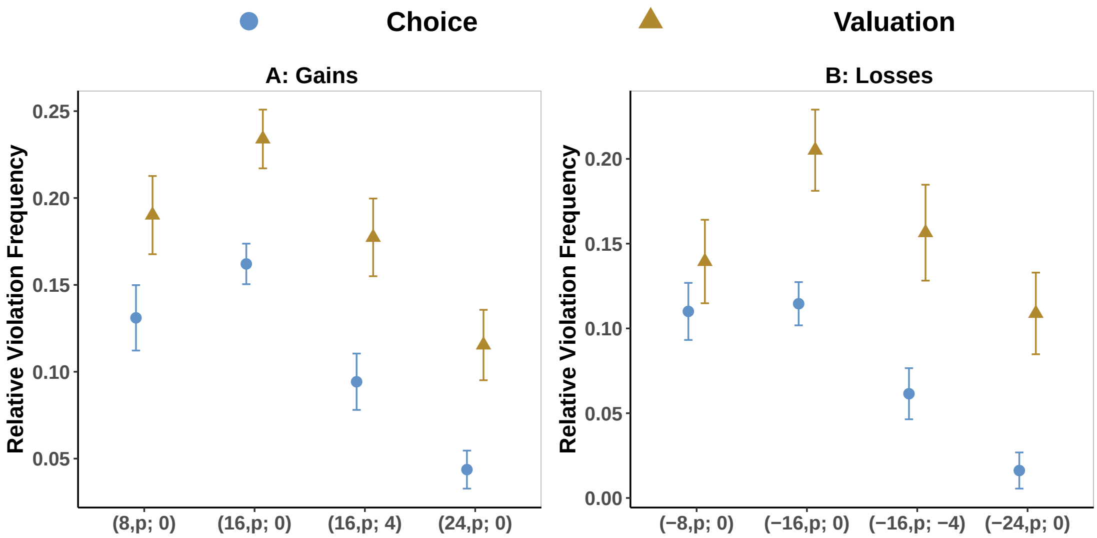


Figure 4: Between-lottery inconsistencies in Valuation vs Choice

Panels A and B examine consistency violations in levels of risk aversion between 'lists' as the probability increases for identical outcomes. These plots report the relative violation frequency, which is the violation frequency divided by the number of all possible violations.

Second, in addition to making choices that suggest more stable risk postures across lotteries, subjects in Choice also make individual decisions that are more consistent with one another when faced twice, revealing higher test-retest reliability (a common measure of the noisiness of behavior). In the experiment, we assigned subjects two entire 'choice lists' (again, presented as MPLs in Valuation and randomly ordered individual choice problems in Choice), $(8, 0.8; 0)$ and $(16, 0.8; 0)$ for gains and $(-8, 0.8; 0)$ and $(-16, 0.8; 0)$ for losses *twice* (at different points in the experiment) and calculated the proportion of the time the riskier option was selected. The test-retest reliability for Valuations on this metric is 0.69, with a 95% confidence interval of [0.598 , 0.831], a typical rate for MPL experiments. By comparison, the data points obtained using Choice tasks have a *much higher* test-retest reliability of 0.915 with a 95% confidence interval of [0.883 , 0.938].

Similar observations hold for losses.[12] Thus despite the fact that subjects face individual choices in a random order in Choice, they express substantially more stable preferences for risk than subjects do in the seemingly more orderly Valuation tasks.

Subjects are therefore more consistent in repetitions of the same lottery in Choice than Valuation. Subjects also show more consistent behaviour across lotteries in Choice, being more likely to increase their risk taking as the expected value of the lottery increases. The one sense in which subjects are plausibly less consistent is in their rate of "multiple switching" – i.e., in the rate at which subjects choose the risky option at a higher sure payment while also having chosen the sure option at a lower payment. In particular, in Choice subjects are more likely to switch multiple times "within list" than they are in Valuation. However, in Appendix A.5 we show that these errors (in both Valuations and Choice) are far from random errors, and instead strongly peak near expected value in a region matching subjects' own risk postures. This orderliness and the fact that they occur at points likely near subjects' own certainty equivalents is strongly consistent with documentation of noisiness occurring in "regions of indifference" by Cubitt, Navarro-Martinez and Starmer (2015) and Agranov and Ortoleva (2017). These inconsistences therefore (unlike instabilities in risk postures and noise in individual choices discussed above) seem inconsistent with confusion-driven random error and instead resemble standard psychometric trembles widely documented for (near-) indifferent subjects.

Summarizing, these data seem inconsistent with an explanation for our results rooted in lower data quality in Choice relative to Valuation tasks. At best, the evidence is mixed. But if anything the results seem more consistent with the opposite interpretation. This interpretation is indeed also supported by our structural estimations of prospect theory functionals, reported in section 3.1, which estimate much higher noise variance for Valuation compared to Choice, suggesting again that Choice behavior is overall more consistent than Valuation behavior in our data. Such errors are rarely reported and even less often discussed in the prospect theory literature, since they have little to do with the inner workings of the model. Examining errors, however, allows us to 'aggregate' the different types of errors that can occur in Choice and Valuation. Starting with Gains, the standard deviation of the residuals is 0.120 (SE 0.003) in Valuation, but much smaller at

[12] The test-retest reliability for Valuations in losses is 0.528, with a confidence interval of [0.368 , 0.657]. Once again, test-retest reliability is much larger for binary choice at 0.854, with a confidence interval of [0.791 , 0.899].

0.037 (SE 0.001) in Choice. The same holds for Losses, where we find errors of 0.127 (SE 0.004) for Valuation and 0.072 (SE 0.002) in Choice. Choice thus induces lower errors than Valuation in our data.

A natural interpretation of these results is that they show that subjects express highly internally consistent preferences towards lotteries in Choice but tremble near indifference (as documented in the vast literature on psychometrics in psychology). By contrast, Valuation tasks induce an artificial internal consistency within-list, causing them to therefore express artificial certainty equivalents that are highly unstable because of their more distant relationship to true preferences. If this interpretation is correct, it is the fourfold pattern (rather than the twofold pattern) that is an artifact of elicitation-driven errors in our data.

2.1.2 Endogenous reference points?

These results suggest that the key behavioral predictions of prospect theory fail (the fourfold pattern) or are seriously weakened (likelihood insensitivity) in Choice. This is a fundamental predictive failure of the theory *even if* prospect theory functionals were able to ex post rationalize the difference between Choice and Valuation. However, in Online Appendix A.3 we show that the machinery of prospect theory fails also at ex post rationalization. There, we rule out perhaps the most salient way prospect theory might be used to "explain" the surprising difference between Valuation and Choice in our data: endogenous reference points. If reference points are fixed at 0 in Valuation tasks like ours (as is often claimed in the literature, e.g., Hershey and Schoemaker, 1985), but vary across Choice tasks this will produce scope for the expression of loss aversion in the latter, driving a wedge between Choice and Valuation environments. However in the Appendix we show that this wedge is incapable of fitting our data – no single loss aversion parameter, λ, can organize the data and therefore ex post rationalize the differences we observe across the treatments. Thus, not only do prospect theory's core empirical *predictions* fail, the highly flexible functionals of prospect theory also cannot *ex post rationalize* the fundamental changes in behavior we observe in Choice.

3 Reassessing the Literature

The failure of the fourfold-pattern to appear in our Choice data – and, more generally, the very different behavior we observe in Choice relative to Valuation – is surprising for two reasons. First, prospect theory does not predict any difference in risk attitudes across these two settings. Second, a large literature has assessed prospect theory experimentally using data from both Choice and Valuation protocols in *dozens of studies* and has not to our knowledge (with the exception of Harbaugh, Krause and Vesterlund (2010)) reported a failure of the fourfold pattern to arise in Choice.

This raises the natural question: are our findings driven by some idiosyncratic feature of our design, or was this instead *always* true in the data but somehow unnoticed? In this Section we reassess the prior literature to answer this question. In Section 3.1 we collect all relevant prospect theory estimates from the prior literature and use these estimates to assess (using imputed certainty equivalents implied by these estimates) whether the fourfold pattern occurred in Valuation but not Choice as in our data in the past literature. We find that the literature has typically found exactly what we find in our experimental data: that the fourfold pattern does not occur in Choice but does occur in Valuation. In Section 3.2 we interpret this finding and discuss why the prospect theory literature has mostly failed to notice that its "most distinctive implication" does not actually occur in the decision environment (direct Choice) the theory was ultimately designed to predict.

3.1 Meta-analysis

In order to meta-analyze evidence on the fourfold pattern from the prior literature, we collect estimates of the parameters of prospect theory functionals from dozens of prior papers. Most of the prior literature relevant to the fourfold pattern summarize their data by reporting estimates of prospect theory parameters including those from (i) a probability weighting function and (ii) a utility function. Our approach is therefore to calculate certainty equivalents based on these past estimates and examine how often these calculations show evidence of the pattern in both Valuation tasks and Choice tasks.

Inclusion criteria. We began by collecting all prospect theory (PT) estimates from the prior literature that can be used to contrast the fourfold pattern in Valuation and Choice. Our inclusion criteria were that 1) estimates should stem either from a pure Valuation setup (e.g., multiple price lists) or a pure binary Choice setup (thus excluding hybrid elicitation mechanisms such as choice lists filled in based on a bisection procedure)[13]; and 2) for the paper to present an estimation of prospect theory parameters. The latter criterion ensures that we can use the reported PT parameters to infer a predicted CE for a given probability and outcome, and thus to have comparable quantities. We conducted a literature search in the spring of 2023. The search procedures followed closely those of the meta-analysis on loss aversion of Brown et al. (2024). We then read though the abstracts and excluded papers that clearly did not meet our criteria. We subsequently read all papers that had passed this initial stage, and encoded the PT parameters.

Analysis Approach. For each paper, we use prospect theory estimates to calculate a predicted certainty equivalent (CE), $\widehat{c} = u^{-1}[w(p)u(x)]$, where u designates the utility function, w the probability weighting function, and u^{-1} the inverse of the utility function. In what follows, we use $x = 100$ and probabilities $p = 0.1$ and $p = 0.9$ to calculate certainty equivalents at low and high probabilities, but the qualitative results do not change much for different monetary outcomes or even more extreme probabilities (see Online Appendix B).[14]

We obtain standard errors for the predicted CEs by applying a bootstrap procedure. We assume the parameters to be normally distributed around their mean estimate with variance equal to the squared standard error of the parameter (as costumary in meta-analysis). By drawing repeatedly from the uncertainty intervals surrounding the parameters and using the draws to calculate a vector of CEs, we can obtain an approximation of the standard errors surrounding our mean estimate of the predicted CE. Online Appendix B contains the details of the procedure and presents an overview of the imputed CEs for different studies.

Obtaining standard errors allows us to apply Bayesian meta-analytic procedures to the

[13]We decided to include papers that used a list format, but used 2 or more stages to zoom in on the precise CE. This allowed us to include e.g. the seminal papers of Tversky and Kahneman (1992) and Gonzalez and Wu (1999) in the analysis.

[14]The reason for using probabilities of 0.1 and 0.9 is that these are often the most extreme probabilities included in the studies. Extrapolations to more extreme probabilities should thus be consumed with some caution. Likewise \$100 was an amount often used in the early literature. Since we use CRRA coefficients in our computations, which are indeed estimated in the great majority of papers, the monetary amounts used do not have much influence on our conclusions.

data, weighing each observation by the inverse of its variance (see e.g. Brown et al., 2024). This procedure is of course only as good as the data we feed into it. The bootstrapped standard errors especially could be affected by differences in estimation methods, and functional and modelling assumptions across studies. Our calculation procedure also implicitly assumes that errors across parameters are independent, and may over-estimate the standard errors of the calculated CEs if that assumption does not hold. To counteract these issues, we will present a number of robustness checks. In particular, we will present simple averages of the point estimates, and approximate standard errors by making them proportional to the inverse of the squared sample size.

Results. As expected, we find robust evidence of the fourfold pattern in prior Valuation studies. Figures 5 and 6, Panel A, show estimates for Gains for $p = 0.1$ and $p = 0.9$ respectively. At $p = 0.1$ all but a handful of estimates fall far to the right of 0.1, indicating risk seeking, with an average meta-analytic normalized certainty equivalent of 0.180, with a 95% credible interval of $[0.160, 0.201]$. At $p = 0.9$ we find the opposite, with all estimates falling to the left of 0.9 with an average meta-analytic normalized certainty equivalent of 0.721, with a 95% credible interval of $[0.693, 0.748]$. Panel A in Figures 7 and 8 shows that each of these risk postures "flip" in Valuation studies under Losses. At $p = 0.1$ (pictured in Figures 7), the meta-analytic mean for the 20 studies again falls far to the right of 0.1, which in losses now indicate *risk aversion* (mean: 0.212; CrI: [0.170 , 0.256]). At $p = 0.9$ (pictured in Figures 8), the meta-analytic mean for the 20 studies flips again, now falling far to the left of 0.9, indicating *risk seeking* in losses (mean: 0.764; CrI: $[0.736, 0.790]$). Thus, in prior Valuation studies, with few exceptions, we calculate certainty equivalents indicating risk seeking for low- and risk aversion for high-probablility gains, and the reverse in each case for losses, replicating the fourfold pattern observed in Valuation tasks in our experiment.

Our main finding is that this pattern collapses in binary Choice studies, just as it does in our experiment. Panel B in each Figure plots data from these studies. In Figure 5 we find that, in violation of the fourfold pattern, the meta-analytic mean is 0.040 with a credible interval of $[0.029, 0.051]$, indicating a certainty equivalent less than 0.1 and therefore *risk aversion*. Indeed, almost all studies yield *risk averse* certainty equivalents, with no statistically significant results indicating the fourfold pattern's predictions of risk seeking. This risk aversion remains at 0.9 in Figure 6 (mean: 0.746; CrI: $[0.701, 0.788]$]),

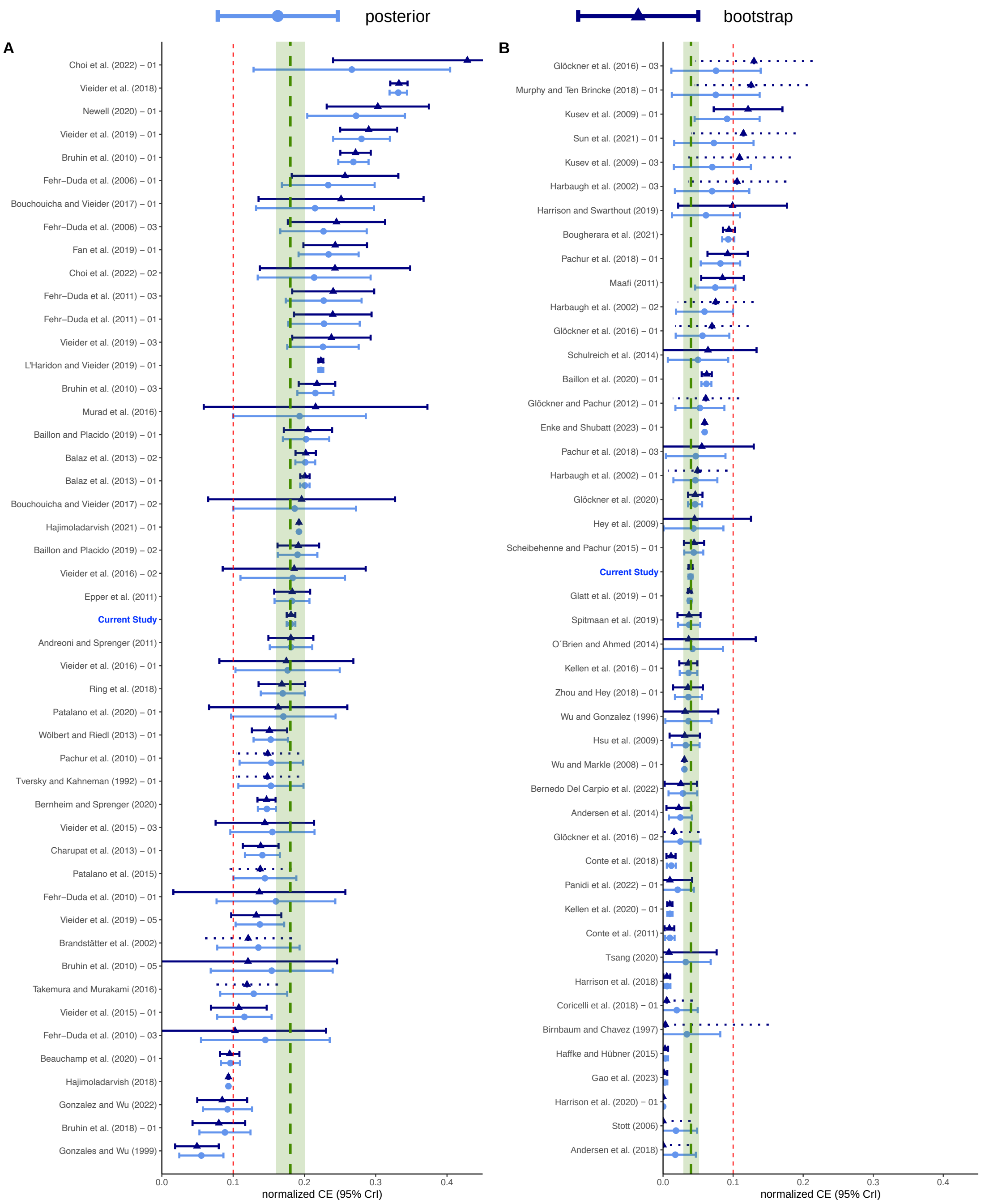


Figure 5: Forest plot of inferred CEs for a wager $(100, 0.1)$, gains
Forest plot of calculated CEs and 95% credible intervals. Panel A summarizes CEs inferred from Valuation studies using elicited certainty equivalents to measure risk attitudes, whereas panel B shows CEs inferred from studies using binary Choice to obtain PT parameters. The navy triangles indicate calculated data points, while the light blue squares indicate the meta-analytic posterior. The vertical dashed lines with the shaded green region surrouding it represents the meta-analytic average with its 95% credible interval. Dotted confidence intervals stem from studies which did not report statsitical information for the reported PT parameters, and for which we thus had to impute the SEs as missing data (see online appendix for details).

suggesting that unlike in Valuation subjects display *uniform risk aversion* in prior Choice experiments in the gains domain. In Losses, at $p = 0.1$ (Figure 7) the meta-analytic

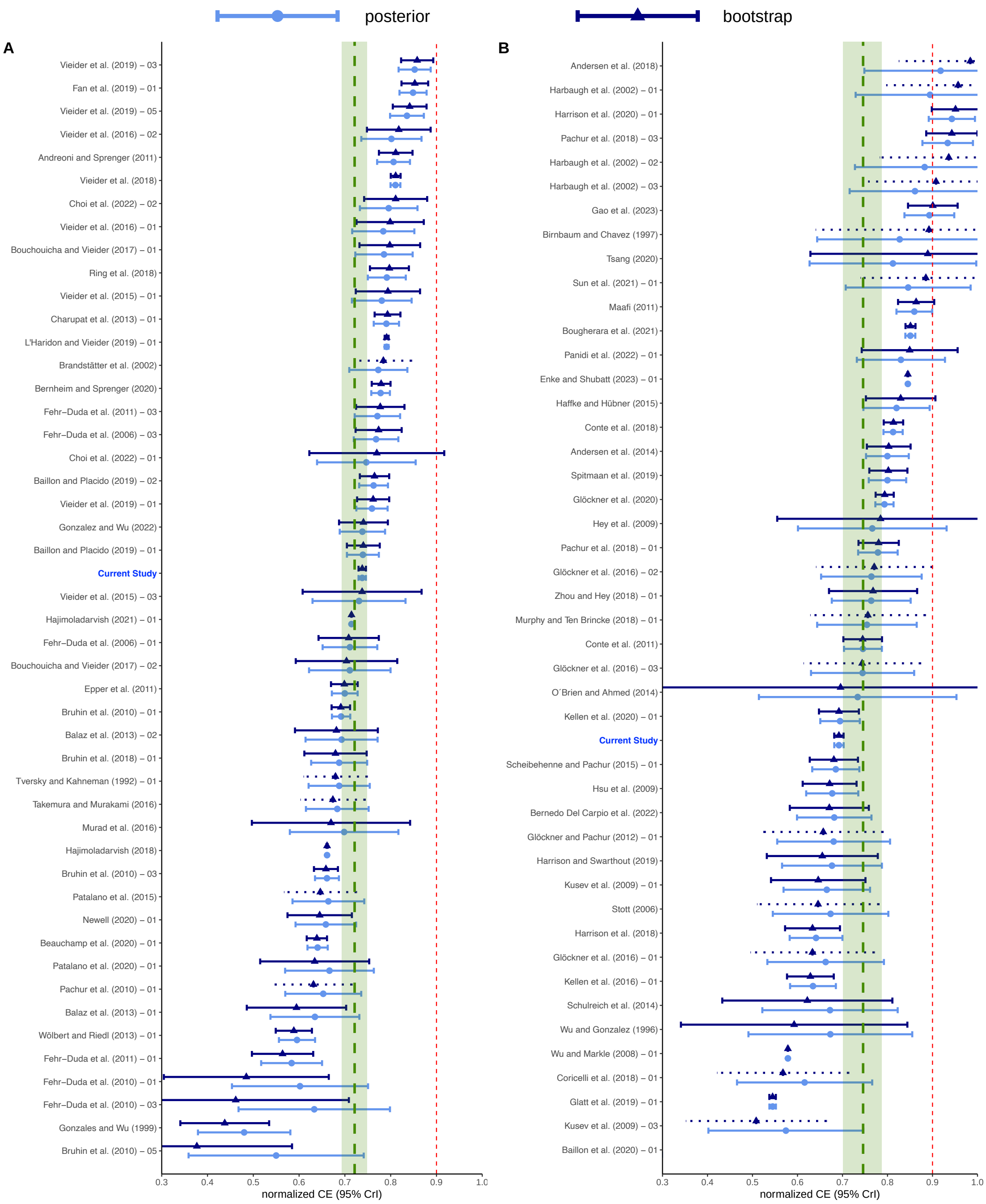


Figure 6: Forest plot of inferred CEs for a wager $(100, 0.9)$, gains

Forest plot of calculated CEs, normalized as $\frac{\widehat{c}}{100}$ to be directly comparable to the probability of winning, and 95% credible intervals. Panel A summarizes CEs inferred from Valuation studies using elicited certainty equivalents to measure risk attitudes, whereas panel B shows CEs inferred from studies using binary Choice to obtain PT parameters. The dark blue triangles indicate calculated data points, while the light blue squares indicate the meta-analytic posterior. The vertical dashed lines with the shaded green region surrouding it represents the meta-analytic average with its 95% credible interval. Dotted confidence intervals stem from studies which did not report statsitical information for the reported PT parameters, and for which we thus had to impute the SEs as missing data (see online appendix for details).

mean is 0.074 with a credible interval of $[0.054, 0.097]$, indicating modest risk seeking. Though the data are more mixed here, only one study shows statistically significant

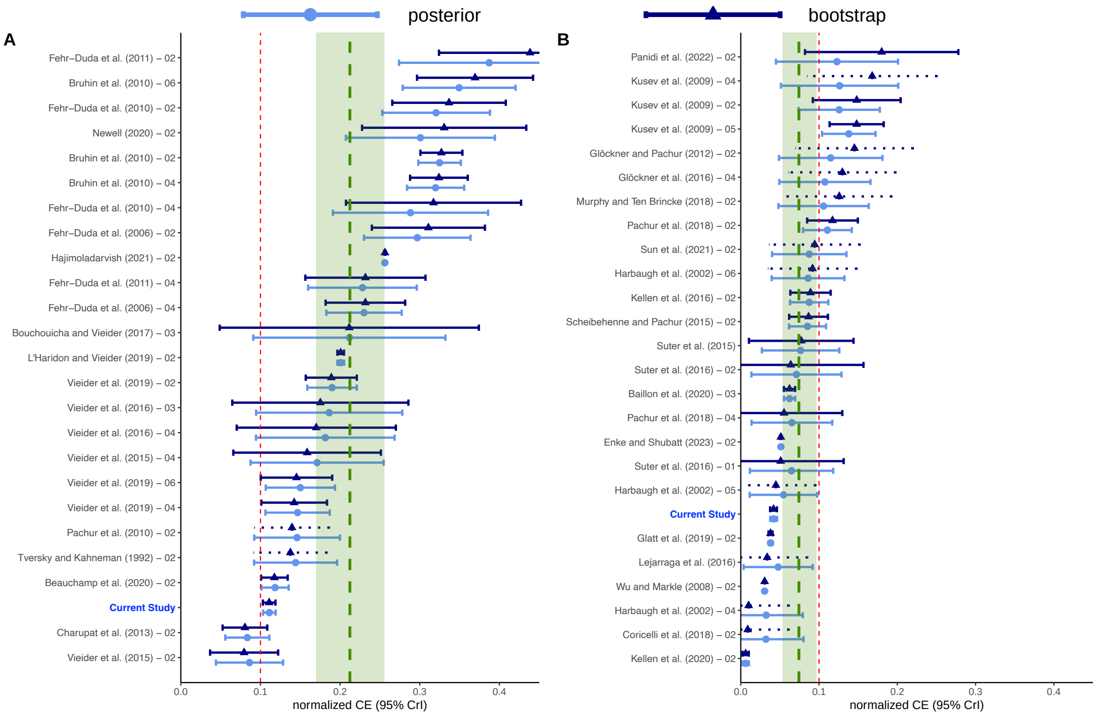


Figure 7: Forest plot of inferred CEs for a wager $(100, 0.1)$, losses
Forest plot of calculated CEs and 95% credible intervals. Panel A summarizes normalized CEs inferred from Valuation studies, whereas panel B shows CEs inferred from studies using binary Choice to obtain prospect theory (PT) parameters. The dark blue triangles indicate calculated data points, while the light blue squares indicate the meta-analytic posterior. The vertical dashed lines with the shaded green region surrounding it represents the meta-analytic average with its 95% credible interval. Dotted confidence intervals stem from studies which did not report statistical information for the reported PT parameters, and for which we thus had to impute the SEs as missing data (see online appendix for details).

evidence of the risk-aversion predicted by the fourfold pattern, whereas 8 are significant in the opposite direction. At $p = 0.9$ (Figure 8) we continue to find evidence of risk seeking behavior in losses in the meta-analytic mean (mean: 0.784; CrI: $[0.714, 0.842]$), suggesting that subjects are reliably risk seeking in losses regardless of probability.

Thus, overall – and as in our experimental data – the fourfold pattern collapses into a twofold pattern in Choice: consistent risk aversion in gains and consistent risk seeking in losses.

Robustness. Meta-analysis can suffer from data problems, meaning robustness checks are especially important to verify our statistical findings. Information on standard errors are not always reported or available in past studies. Even when they are available, the statistical evidence provided may be inaccurate. This holds all the more in our case, since we need to simulate the standard errors of the calculated CEs from a combination of prospect theory parameters. To test the robustness of the results, we can instead

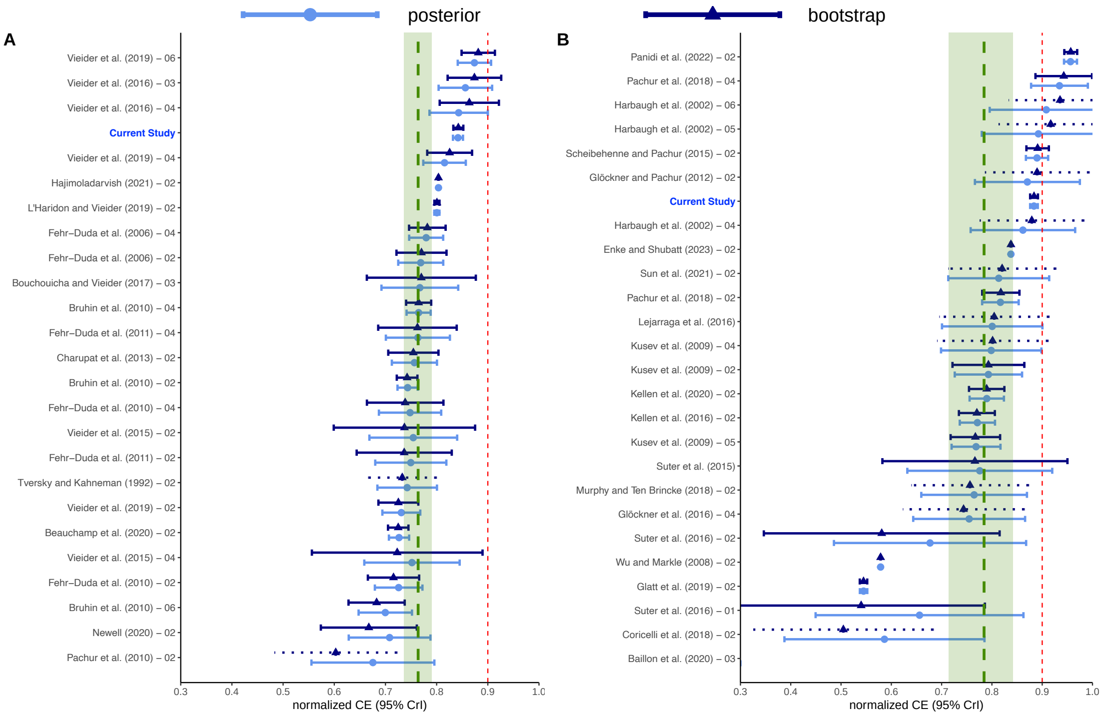


Figure 8: Forest plot of inferred CEs for a wager $(100, 0.9)$, losses

Forest plot of calculated CEs and 95% credible intervals. Panel A summarizes CEs inferred from Valuation studies using elicited certainty equivalents to measure risk attitudes, whereas panel B shows CEs inferred from studies using binary Choice to obtain PT parameters. The navy triangles indicate calculated data points, while the light blue squares indicate the meta-analytic posterior. The vertical dashed lines with the shaded green region surrounding it represents the meta-analytic average with its 95% credible interval. Dotted confidence intervals stem from studies which did not report statistical information for the reported PT parameters, and for which we thus had to impute the SEs as missing data (see online appendix for details).

calculate the simple means of the CEs calculated from the parameter point-estimates. The mean normalized CE calculated from Valuation studies for gains at $p = 0.1$ is 0.186 (median 0.182) and at $p = 0.9$ is 0.707 (median 0.726). The mean normalized CE from Valuation studies for losses is 0.231 (median 0.230) at $p = 0.1$ and 0.774 (median 0.765) at $p = 0.9$. The calculated CE for gains from binary Choice studies is 0.045 (median 0.036) at $p = 0.1$ and 0.752 (median 0.773) at $p = 0.9$. The mean relative CE for losses is 0.079 (median 0.069) at $p = 0.1$ and 0.765 (median 0.800) at $p = 0.9$. Thus our main findings continue to strongly hold. While there is overwhelming evidence for the fourfold pattern in PT estimates obtained from direct Valuation elicitation, estimates based on binary Choice indicate a twofold pattern – risk aversion for gains, and risk seeking for losses. In Online Appendix B we report two additional robustness checks. Following the recommendation of Furukawa et al. (2006), we use the average standard deviation in studies in which we can obtain it, and then divide this standard deviation by the square root of the sample size to derive standard errors. All of our results are robust to this additional robustness check. Furthermore, we conduct a series of meta-regressions, controlling for number of outcomes in the lottery, whether incentives are real or hypothetical, whether one of the two options consisted of a sure payment (in Choice), and a variety of study and estimation characteristics. Once again, our key results are unaffected when controlling for such potential differences between studies, showing that they are not driven by systematic differences in procedures or estimation techniques in Valuation versus Choice.

Likelihood insensitivity. Finally, our meta-analytic evidence suggests that, as in our experiment, likelihood insensitivity is sharply attenuated in past Choice data relative to Valuation data. Figure 9 plots empirical CDFs of the likelihood-sensitivity index (mirroring Figure 3), calculated as the *difference* in normalized certainty equivalents for the lottery pairs at $p = 0.9$ and at $p = 0.1$, at $p = 0.8$ and at $p = 0.2$, and at $p = 0.7$ and at $p = 0.3$, for Gains (left panel) and Losses (right panel) from previous studies. In both cases, perfect likelihood sensitivity would predict a difference of 1, but in Valuation estimates sensitivities are *universally* lower than this. By comparison, in Choice estimates sensitivities are substantially *higher*, with a significant number of studies showing the reversed pattern of likelihood *over-sensitivity*. The difference in likelihood-sensitivity across Valuation and Choice is highly statistically significant based on Wilcoxon tests ($p < 0.001$). Thus, as in our experiment, the prior literature suggests that not only the fourfold pattern but also the subtler phenomenon of likelihood insensitivity is much weaker in Choice than Valuation.

3.2 Why did this go largely unnoticed?

The results of our meta-analysis seem to paint a clear picture: the fourfold pattern was *always* a phenomenon of lottery Valuation, not a phenomenon of lottery Choice. Given the centrality of the pattern to prospect theory, why has the literature mostly failed to notice its absence from the very setting prospect theory was meant to ultimately explain? We think the answer is ultimately rooted in the fact that most of the literature on prospect theory is focused not on non-parametric assessment of risk postures, but instead on estimates of prospect theory parameters. This is especially true of binary Choice tasks in which it is somewhat more difficult to non-paramaterically assess risk postures than in Valuation tasks (where risk postures are readily available by comparing mean

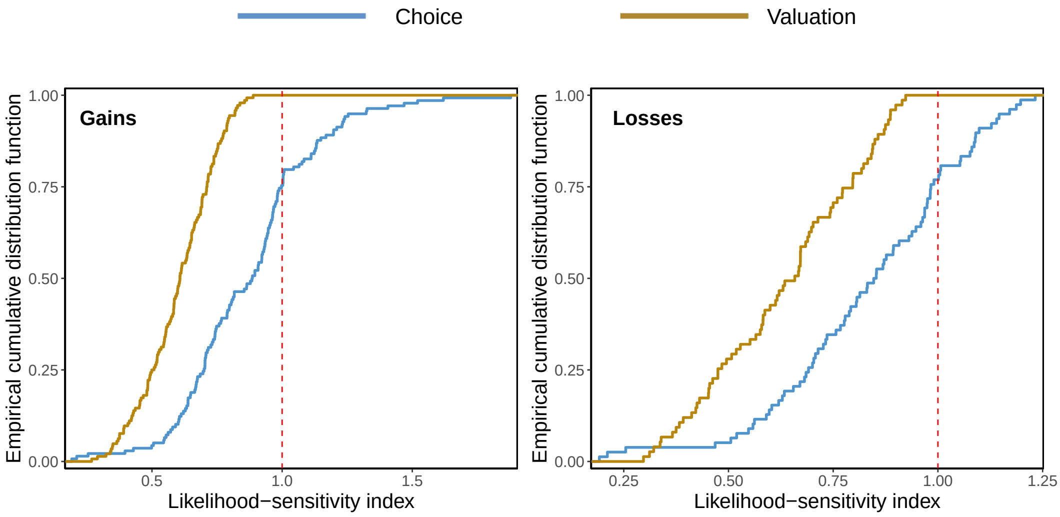


Figure 9: Empirical cumulative distribution function of likelihood-sensitivity
Empirical cumulative distribution function of the likelihood-sensitivity index, calculated as the average of the differences in normalized CEs for the pairs $(X, 0.9)$ and $(X, 0.1)$, $(X, 0.8)$ and $(X, 0.2)$, and $(X, 0.7)$ and $(X, 0.3)$. **Panel A** presents scatter plots of observations for the gain wager $X = 100$, derived from studies employing either certainty equivalents or binary choice methods. **Panel B** displays observations for the loss wager $X = -100$. Vertical solid lines indicate the point of risk neutrality.

elicited certainty equivalents to expected values), leading data to often be summarized via structural estimates.

Reliance on structural estimates, it turns out, makes it easy to miss failures of the fourfold pattern to appear, because such estimates invite the researcher to heuristically read parametric evidence of standard inverse-S shaped probability weighting as approximate evidence of the fourfold pattern. But this heuristic can badly fail because the probability weighting function is only one of the two elements of prospect theory that contribute to certainty equivalents and thereby risk attitudes. Sufficient curvature of the utility function (the second element of prospect theory, alongside the probability weighting function) can lead even a very standard inverse S-shaped probability weighting function to coincide with risk attitudes that are starkly inconsistent with the fourfold pattern.[15] Indeed, our meta-analysis suggests that in past Choice experiments this has *typically been the case*: while 76.1% of the results from prior Choice studies estimate inverse-S shaped probability weighting functions as described by prospect theory, only 15.2% produce certainty equivalents consistent with the fourfold pattern (none of which are

[15]The probability weighting function can be first concave and then convex and thus inverse-S shaped, but stay entirely below the 45 degree line (Wakker, 2010). Even a probability weighting function that crosses the 45 degree line and shows overweighting of small probabilities may actually occur in the presence of risk aversion for small probability gains (risk seeking for small probability losses) if utility is sufficiently concave (convex for losses).

significantly different from risk-neutrality)!

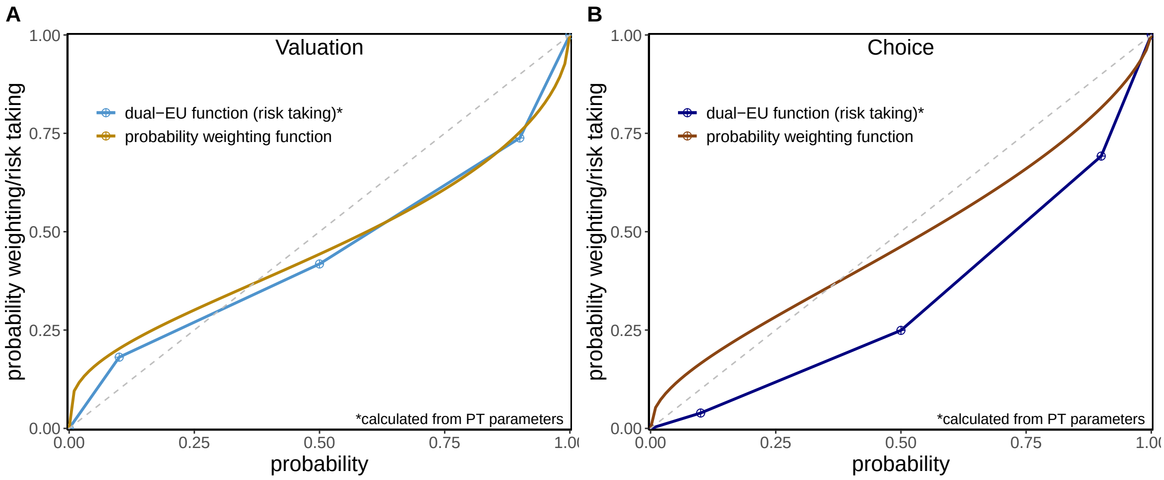


Figure 10: Probability weighting and risk taking

Panel A shows the parametric Prelec 1-parameter function estimated on our Valuation data, with a likelihood-sensitivity parameter of 0.560 (SE 0.007) (estimates using a Tversky-Kahneman function are similar). The blue points indicate CEs calculated based on our estimated functionals. Given that utility is estimated to be close to linear (with a power utility parameter of 0.934, SE=0.007), implied CEs closely track the probability weighting function. Panel B shows the probability weighting function obtained from our Choice data, with a likelihood-sensitivity parameter of 0.705 (SE=0.009). The function is thus inverse-S shaped, just like for Valuation. The implied CEs, however, paint a very different picture, due to the power utility parameter that now indicates much more risk aversion (0.556, SE=0.005).

Returning to our Data. We can illustrate the problem by structurally estimating prospect theory parameters on our own data using the most standard parameterization: (i) a standard power-utility value function and (ii) a 1-parameter weighting function (though in Online Appendix B.4 we show similar findings for more flexible specifications). Focusing on Gains, we estimate a likelihood-sensitivity of 0.560 (SE 0.007) for our Valuation task, yielding a standard inverse S-shaped probability weighting function as pictured in the left hand panel of Figure 10. In Choice we estimate a sensitivity of 0.705 (SE 0.009) indicating substantially less insensitive behavior that comes much closer to EU benchmarks (matching our non-parametric findings in Section 2). Nonetheless, as the right hand panel of Figure 10 shows, in Choice we continue to find the expected inverse S-shaped weighting function. Were we to focus primarily on these estimates when assessing the fourfold pattern we would be lead to believe that the pattern arises in Choice just as it does in Valuation. However, as we've seen from our analysis of the raw data drawing this conclusion would be a significant mistake.

To see this, alongside the parameteric estimates in each panel we also plot (in blue) *certainty equivalents*, calculated from the equation $\widehat{c} = u^{-1}\left[w(p)u(x)\right]$, where $w(p)$ indicates the probability weighting function, $u(x)$ the utility function, and u^{-1} indicates the inverse of the utility function. Unlike probability weights, these certainty equivalents take utility curvature into account and so can be directly compared to the expected value (EV) of the wager, indicating risk aversion if $\widehat{c} \leq EV$, and risk seeking if $\widehat{c} \geq EV$. While these line up nicely with probability weighting estimates in Valuation, they imply fundamentally different risk postures in Choice. Instead of showing the characteristic "flip" in apparent risk preferences at low vs. high probabilities implied by the parametric estimates, behavioral data in Choice indicates that subjects are globally risk averse. The reason for this disconnect is that these identical estimation exercises yield almost no utility curvature for subjects in Valuation but substantial utility curvature in Choice. While power utility estimates for Valuation in our structural estimations are nearly linear (risk neutral) in Valuation with ρ =0.934 (SE 0.007), utility is extremely concave in Choice with an estimate of ρ =0.556 (SE 0.005).[16] To contextualize these results in terms of the literature, in Figures 5-8 we plot certainty equivalents inferred from our structural estimates alongside similarly computed estimates in our meta-analysis, and we recover exactly the same pattern our non-parametric analysis suggests: a fourfold pattern in Valuation and twofold pattern in Choice.

Similar patterns and opportunities for confusion have arisen throughout the literature's history, stretching back to the earliest efforts to estimate prospect theory functionals. To illustrate, consider two of the landmark studies in the early literature: Tversky and Kahneman (1992) (a Valuation study) and Wu and Gonzalez (1996) (a Choice study). As Figure 11 shows, estimated likelihood sensitivity is nearly identical in the two datasets, producing virtually identical inverse S-shaped probability weighting functions. However, as in our data, estimates suggest very little utility curvature in the Tversky and Kahneman (1992) Valuation study (a power estimate of $\rho = 0.88$) but pronounced curvature in the Wu and Gonzalez (1996) Choice study ($\rho = 0.5$). When we calculate certainty equivalents from these estimates as we did with our own data (plotted in blue in Figure 11), we find strikingly similar evidence to ours. Probability weighting estimates are a good approximation of certainty equivalents in the Tversky and Kahneman (1992)

[16] Similar findings also obtain for losses, where Valuation produces sensitivity 0.828 (SE 0.014) and utility 0.906 (SE 0.010), whereas Choice results in sensitivity 1.053 (SE 0.015) and utility curvature ρ =0.758 (SE 0.008). It has often been remarked in the literature that losses are more erratic than gains (Abdellaoui, 2000; L'Haridon and Vieider, 2019), so that the failure to find insensitivity in the loss domain is hardly remarkable.

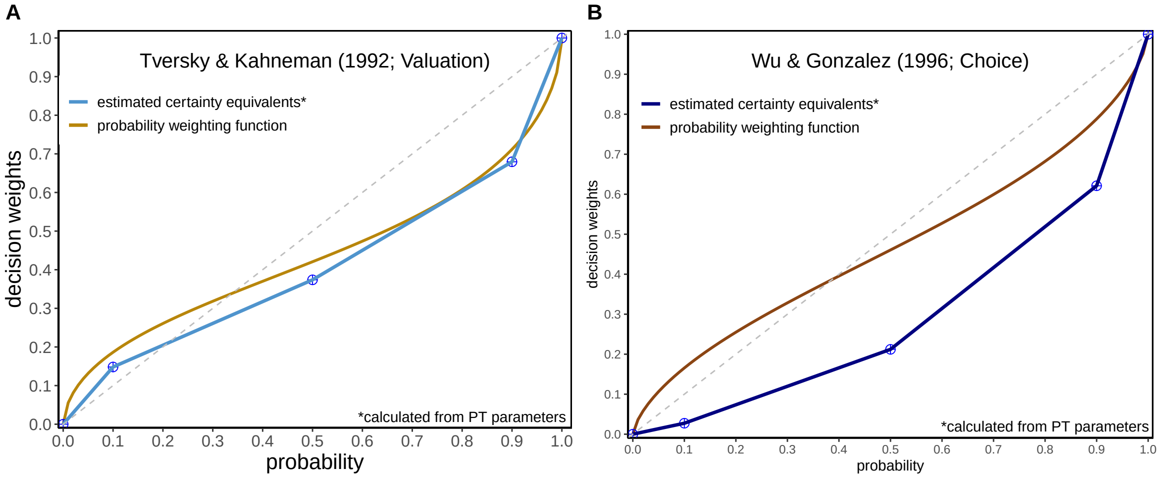


Figure 11: Probability weighting and risk taking

Panel A shows the parametric Tversky-Kahneman function estimated by Wu and Gonzalez (1996), with a curvature parameter $\gamma = 0.71$. The blue points indicate CEs calculated based on their estimated functions. While the parametric function shows overweighting of small probabilities, risk attitudes indicate risk *aversion* for small probability gains. The discrepancy is explained by their power utility coefficient, which at $\rho = 0.5$ indicates substantial risk aversion. Panel B shows the parametric probability weighting function estimated by Tversky and Kahneman (1992) from certainty equivalents. In this case, the overweighting of small probabilities of a gain indeed translates into risk seeking. The much smaller gap between the two functions can be traced back to utility curvature, which is much less pronounced at $\rho = 0.88$.

Valuation study and therefore of risk attitudes. But, the heuristic value of probability weighting for assessing risk attitudes falls apart in Choice: as in our data, there is a significant gulf between probability weights and certainty equivalents in the Wu and Gonzalez (1996) Choice study, where certainty equivalents are starkly inconsistent with the fourfold pattern.

Discussion. The striking similarity of Figures 10 and 11 strongly reinforces our main observation in this section: in an important sense our experimental findings are not new, but instead have been with us since the beginning. However, they have gone largely unnoticed probably because of the literature's intensive focus on structural parameters rather than their implications for risk attitudes. This is perhaps understandable because, as we've discussed, the distinctive character of the fourfold pattern is ultimately attributable to probability weighting so that it is perhaps natural to treat evidence of the latter as implicit evidence of the former. In order to identify the shortcomings of this heuristic inference, the literature would have had to calculate certainty equivalents by joining evidence from the probability weighting function with evidence from the utility function as we have above. The literature has rarely done this and so has missed the regularity we document. For instance Wu and Gonzalez (1996) emphasize the importance of the fourfold pattern as one of the "critical empirical regularities that any good descriptive model should accommodate [...]: risk aversion for most gains and low probability losses, and risk seeking for most losses and low probability gains" (p. 1676) but do not seem to have examined whether their estimates imply the pattern. If they had, they would have noticed that prospect theory fails this test for a good descriptive model![17]

Ultimately, the literature's failure to notice the collapse of the fourfold pattern in Choice is downstream of a larger point the literature has under-emphasized: the fact that Choice and Valuation behaviors tend to differ significantly, even in parametric estimates of prospect theory parameters. First, even when estimated probability weighting parameters are similar, estimated utility curvature tends to be substantially more severe in Choice than in Valuation. Indeed, in our meta-analysis we find that power utility parameters are far higher in *Valuation* tasks (mean = 0.86, median = 0.97) than in *Choice* tasks (mean = 0.64, median = 0.64) for gains, with similar divergences in losses.[18] Second, likelihood insensitivity – the core driver of probability weighting's distinctive shape – tends to be much less pronounced in Choice than Valuation in the prior data. Both of these patterns were probably under-noticed because very few studies actually conduct head-to-head comparisons of the two settings as we do in our experiment.

4 Cognitive Frictions and the Valuation-Choice Gap

The focus of our paper has been prospect theory's failure to *predict* Choice behavior. But its more fundamental failure is its inability to *explain* the significant differences between risk attitudes in Choice and Valuation – a failure that arises both in our experiment and in the prior literature. These differences can be summed up as follows:

1. Likelihood insensitivity, the core characteristic of the *probability weighting function*, tends to be far *less severe* in Choice than Valuation.

[17] The failure to detect the absence the fourfold pattern is particularly understandable in the case of Choice because (unlike with Valuations), certainty equivalents are difficult to directly infer non-parametrically from choice data and often have to be inferred indirectly based on certainty equivalents calculated from parameters. This is doubly true in Wu and Gonzalez (1996) who use a ladder design in which it is especially difficult to infer certainty equivalents based on raw choice patterns.

[18] In losses, power utility parameters are higher in *Valuation* tasks (mean = 1.02, median = 1.09) than in *Choice* tasks (mean = 0.75, median = 0.73). Note, indeed, the smaller values indicate increased *convexity* for losses, and are thus an indication of risk *seeking*.

2. Utility curvature, the core characteristic of the *utility function* tends to be far *more severe* in Choice than Valuation.[19]

3. Although less often measured, in datasets like ours, certainty equivalents tend to be *noisier* in Valuation than Choice.

A failure to explain these differences is hardly unique to prospect theory – *no* standard preference-based theories of risk-taking can account for patterns like these. This is because standard theories of risk preferences (prospect theory included) root risk postures entirely in the payoffs and probabilities underlying the lotteries being evaluated, which do not vary between the two choice environments. This is cast in particularly sharp relief in our experiment (Section 2), which was designed to feature *identical* menus, information and incentives in the two settings. Under the lens of preference-based theories, these tasks are identical.

Noisy coding. If preference-based theories cannot explain these regularities, what can? We consider the possibility that the widespread gulf we've documented between Choice and Valuation is an outgrowth of the way cognitive frictions differentially distort behavior in the two settings. Our starting point is a groundswell of recent evidence suggesting that probability weighting itself represents, not an expression of preferences, but instead a severe cognitive distortion in the evaluation of lotteries. In particular, a class of recent *noisy coding* models has proved remarkably successful in organizing and predicting distinctive anomalies surrounding probability weighting (Khaw, Li and Woodford, 2021; 2023; Vieider, 2023; Oprea, 2024*b*; Oprea and Vieider, 2024; Frydman and Jin, 2023). These models are rooted in the idea that (i) cognitive limitations cause decision makers to perceive or represent the descriptive primitives of lotteries (e.g., probabilities, payoffs) with *noise* and (ii) have prior beliefs about what these lottery primitives are. The key idea of noisy coding is that decision makers minimize the negative effects of their noisy perception by combining those perceptions with their prior belief in a standard, Bayesian manner. As a result, imperfections in the perception or representation of lottery primitives produce not just noise but systematic biases. In particular:

[19]Curvature is the core characteristic of the utility function when assessing unmixed lotteries, as we do in our paper. For mixed lotteries, the reference point and parameter of loss aversion are equally important. The evidence for loss aversion is currently also under review, with recent studies documenting stake-dependence of loss attitudes (Ert and Erev, 2013), increasing evidence that the parameter may be close to 1, or even fall below 1 (Chapman et al., 2024), and challenges to the explanatory power of the concept of loss aversion (Chapman et al., 2023*b*).

1. Bayesian shrinkage will produce likelihood insensitivity, systematically distorting the mapping between probabilities and values in a manner that exactly matches the standard inverse S-shape of standard probability weighting. The noisier perceptions are, the stronger the likelihood insensitivity.
2. Prior beliefs about the log-odds of the lottery paying an extreme amount will systematically distort the apparent risk attitudes of decision makers, producing apparent curvature in estimates of utility functions. Such apparent risk aversion can further be distorted by noise, with noise resulting in an uplift of the prior and hence an apparent increase in risk taking (for gains) or risk aversion (for losses).
3. Noisiness in decision-makers' perceptions will generate noise in their *decisions* – the noisier perceptions are, the noisier choices will be (at least over some, empirically plausible, ranges).

Given these three implications (and their correspondence to the three Choice-Valuation differences listed above), to whatever degree Choice and Valuation (i) generate different levels of noise in the perception of lottery primitives and (ii) distort prior beliefs about payoff distributions, we should expect likelihood sensitivity, utility curvature and behavioral noise to also differ across the two settings.

Procedural Differences Between Choice and Valuation. Noisy cognition models, like most cognitive models (and unlike standard models of risk preferences like EU and PT) make predictions that depend on the *procedure* the decision-maker uses to make decisions. In particular the way cognitive acts are sequenced and arranged by the DM to make decisions will have substantial impacts on the formation of beliefs and the way cognitive noise evolves. This is important because there are strong reasons to think that the way decision makers are primed to process information is procedurally different in Valuation than in Choice.

Intuitively, in binary Choice the most obvious procedure for choosing between a lottery and a sure payment is for the DM to separately and independently evaluate the relative worth of each, and then choosing whichever seems more valuable. By contrast, when *valuing* a lottery (i.e., in Valuation), the decision maker *first* must assess the value of the lottery and only after doing so search over many possible sure payments for one that is equivalent to that lottery value. These two choice procedures are equivalent for decision-makers who suffer no cognitive frictions, but not for decision makers who suffer from the kinds of frictions described in noisy coding models.

In particular, in Choice, noise in the independent imprecise evaluations DMs make of lotteries and rewards will tend to cancel one another out when the two are compared to make a choice, limiting (or even eliminating) the effects of cognitive noise and sharply attenuating the three implications of noisy coding sketched above. The result will be weak likelihood dependence, high risk aversion (high apparent utility curvature in estimation) and relatively low-noise decisions. By contrast, the sequential procedure of first evaluating the lottery and then iteratively searching for an equivalent value, as required in Valuation, will tend to produce the opposite. Evaluating potential outcomes by comparison to the initially valued lottery will introduce an additional Bayesian bias to the later evaluation of outcomes that will *intensify* rather than attenuate cognitive noise. Following the three implications of noisy coding models discussed above, this will produce strong likelihood dependence, low apparent risk aversion and relatively noisy decisions. Thus, noisy coding, when applied to the different procedures induced by Valuation and Choice, produce exactly the three differences between the two we summarized at the beginning of this section.

The technical arguments underlying these implications are subtle and require significant notation, so we defer them to Online Appendix C. There we offer a noisy coding model based on Vieider (2023), which uses technical assumptions suggested by recent neuroscience. However, the implications we draw (as sketched above) do not depend on the idiosyncracies of our model, but rather seem to be rather general implications of noisy coding applied to Choice-like and Valuation-like evaluation procedures. For instance Khaw, Li and Woodford (2023), in contemporaneous work, draw similar conclusions about Valuation using a noisy coding setup building on the somewhat different modeling choices used in Khaw, Li and Woodford (2021).

Implications. The most important implication of this is that coding models like ours *predict* the three key distinctions between Valuation and Choice that we have documented in our experiment and in the prior literature, and therefore ultimately explain why the fourfold pattern appears in Valuation but not in Choice. Such models predict that the *reason* the fourfold pattern does not appear in Choice is ultimately that the pattern itself is a consequence of cognitive errors that are intensified in Valuation relative to Choice.

In particular, the fourfold pattern is a consequence of the fact that noisy Valuation produces *both* intense likelihood insensitivity and artificial apparent risk neutrality (with the corollary of highly inconsistent behavior between tasks), the combination of which are necessary for the fourfold pattern to arise. Thus, in addition to explaining the Valuation-Choice gap, noisy coding offers an interpretation of the key patterns of prospect theory as growing not out of risk preferences, but of cognitive errors.

Of course, the measure of an explanation like ours is its *excess explanatory power* – its ability to explain further phenomena that it wasn't designed to account for. One crucial implication of our explanation is that the difference between Valuation and Choice actually isn't caused by the MPL format or the orderliness of typical MPL tasks etc., but instead to something subtle in the decision-making procedure the task of Valuation tends to induce. In particular, it is the fact that Valuation requires information to be processed *sequentially* that our model suggests generates the difference in behavior. To value a lottery, the DM must first form an assessment of the lottery and then subsequently assess certain payments to find one that equalizes with this value. Simply put, this is a more difficult and noisier process than direct Choice.[20]

This suggests a distinctive test for our explanation of the gap between Choice and Valuation. If, as our model suggests, it is the order of evaluation of information that generates the gap, we should be able to cause Choice behavior to converge towards Valuation behavior simply by forcing subjects to evaluate Choice problems in a Valuation-like way. In particular, if we induce subjects to evaluate a lottery up front and inform the subjects that they will be asked to subsequently contrast this with certain payments in a series of binary choices, then our explanation suggests that we should see some degree of convergence in likelihood dependence, valuation curvature and noise and the emergence of distinctive features of the fourfold pattern. We report just such a test next.

4.1 An Experimental Test

To test the hypothesis discussed at the end of the previous subsection, we conducted an experiment on Prolific UK using a similar design to the experiment reported in Section

[20]Notably, this explanation is highly consistent with the fact that the fourfold pattern *does* arise under the BDM mechanism, an alternative mechanism to the MPL for eliciting Valuations (see e.g. Chapman et al., 2023*a*, for evidence of very high correlations between valuations elicited in choice lists and valuations obtained from BDM mechanisms). This suggests that the Valuation-Choice gap is driven not by details of the elicitation mechanism, but instead by the cognitive act of valuation itself.

2, but with a simpler set of lotteries.[21] In the experiment we repeated our Valuation and Choice designs, and introduced a new treatment we call Sequential-Choice. Approximately 50 subjects participated in each condition, resulting in 150 participants for the entire experiment.

In our novel Sequential-Choice experiment, we attempt to induce subjects to reason about binary Choice in a Valuation-like, sequential way. For each probability, subjects are *first* shown the lottery and asked to evaluate it. They are then given a series of binary choices between the lottery they have just evaluated and certain payments in a random order. Thus the only real differences relative to our Choice treatment is (i) binary choices are temporally grouped by lottery and (ii) subjects are asked to consider the lottery prior to making those choices.

Results. Figure 12 shows the results. Panel A shows the choice proportions for each treatment with the 45 degree line plotting the risk neutral benchmark: choice proportions above (below) this line reveal evidence of apparent risk seeking (aversion). The Valuation and Choice conditions replicate the main results from the Gains treatment of the experiment in Section 2 above. In particular, subjects in Valuation are risk seeking at low and risk averse at high probabilities, replicating the gains component of the fourfold pattern. However in Choice, we find uniform risk aversion, consistent with the twofold pattern we've documented in our experiment and the prior literature. What's more, we see non-parametric evidence of the three key differences we opened this section with and which our model predicts. Panel A shows that choice proportions change much more gradually in Valuation than Choice (evidence of greater likelihood insensitivity), and the lower risk choice at $p = 0.5$ in Choice than Valuation is consistent with far greater utility curvature. In Panel B, we show that as in earlier work, behavior is considerably more noisy in Valuation than Choice (as measured by between-task inconsistencies in observed certainty equivalents).

Our main finding, however, is that these gaps in behavior almost entirely disappear in our Sequential-Choice treatment. Simply by forcing subjects to evaluate the lottery intensively prior to making binary choices, we cause the fourfold pattern to re-appear and indeed for choice proportions to become *virtually identical* to our Valuation treatment.

[21] In particular, subjects evaluated lotteries that paid £24 with probabilities of 0.1, 0.3, 0.5, 0.7, or 0.9, and £0 otherwise.

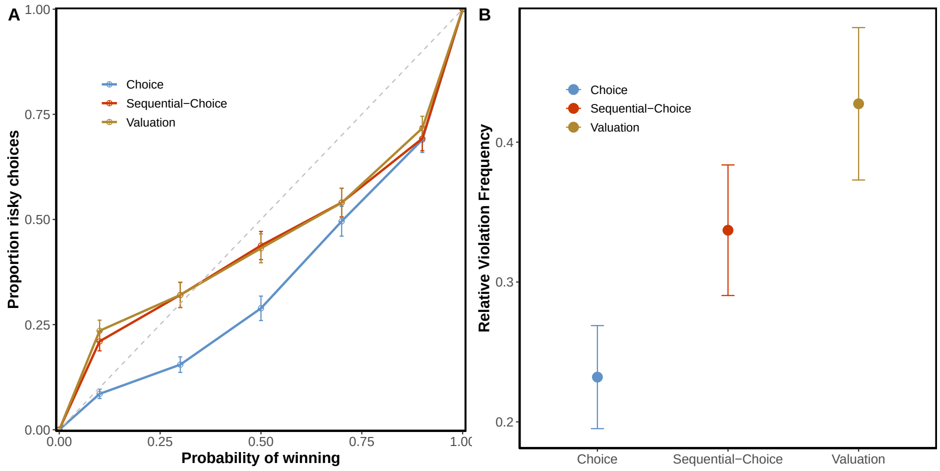


Figure 12: Choice patterns in Valuation, Choice, and Sequential-Choice.
Panel A plots certainty equivalents from each of the three treatments. The certainty equivalents are derived from stochastic switching points or choice proportions for the risky options, as throughout the paper. Panel B plots inconsistency in choice proportions across lotteries, defined as the choice proportion of the lottery declining as its expected value increases (i.e., as the probability of winning increases, since all else is kept constant).

We also find that choice inconsistency rises in this treatment relative to Choice, evidence for a distinctive claim of our model – that sequential evaluation produces noisier cognition and behavior.

Figure 13 shows the results of structural estimations of PT parameters, based on the same specification as in section 3.1 above (details about the estimation procedures can be found in Online Appendix B). In Valuation, the estimated probability weighting function and the inferred CEs track each other closely. Probability weighting indicates considerable likelihood insensitivity (0.649, SE 0.028), but utility is close to linear with a power parameter of 0.857 (SE 0.023). In Choice, shown in Panel B, likelihood insensitivity is again significantly less strong, but clearly present at 0.750 (SE 0.024). However, the strong concavity of the utility function (power parameter 0.571, SE 0.012) drives a wedge between the probability weighting function and the calculated CEs. This replicates and confirms our previous results. By contrast, the results for Sequential-Choice, shown in panel C are very similar to Valuation, rather than Choice: likelihood insensitivity is very strong at 0.530 (se 0.019), whereas utility is close to linear at 0.926 (se 0.017). These two facts together mean that the calculated CEs once more closely track the probability weighting function – mirroring typical Valuation rather than Choice behavior. Finally, the residual errors also line up with our hypothesis: 0.139 (se 0.010) in Valuation, 0.033 (se 0.001) in Choice, and 0.119 (se 0.007) in Sequential-Choice. Thus we can make Choice behavior nearly as noisy as Valuation behavior simply by inducing subjects to process information in a Valuation-like way.

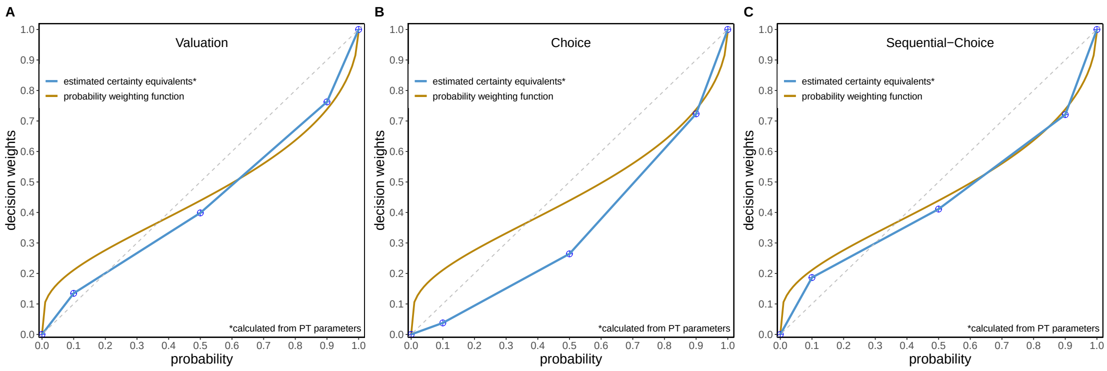


Figure 13: Choice patterns in CE, BC, and sequential evaluation
The figure shows probability weighting functions estimated in a PT model, jointly with certainty equivalents calculated from the PT parameter estimates. Panel A shows the estimates for Valuation tasks, panel B for Choice tasks, and panel C for Sequential-Choice tasks.

Summarizing then, our experimental results provide strong evidence in favor of a highly distinctive implication of our explanation. The highly likelihood-insensitive probability weighting and near-linearity of utility that drives the fourfold pattern in Valuation are not actually expressions of subjects' risk preferences. Rather they are outgrowths of the way decision makers procedurally approach the particularly difficult task of Valuation, which requires a sequential mode of evaluation that tends to amplify perceptual and evaluative biases. Simply inducing this style of evaluation in binary Choice is sufficient to induce these artificialities and thereby cause binary Choices to obey the predictions of prospect theory.

5 Discussion

We show that the "most distinctive implication" of prospect theory (Tversky and Kahneman, 1992) does not actually arise in direct lottery choice (Choice tasks). Instead the *fourfold pattern of risk attitudes* is an apparent artifact of explicit elicitations of certainty equivalents (Valuation tasks) that simply does not occur in direct choice. We first show this using a novel experiment that is designed to maximize comparability between binary Choice and overt Valuation tasks. We then show via meta-analysis that the same has always been true in the literature, but has gone mostly unnoticed because of the way the literature tends to summarize data. Thus prospect theory fails its most diagnostic test (at least according to the theory's creators) in the very setting the theory was ultimately designed to explain.

This fundamental *predictive failure* of prospect theory is downstream of a deeper *descriptive failure* of the theory that we likewise identify with new experiments and show by meta-analysis has always been latent in the literature. The estimated parameters of prospect theory functionals are systematically different in Valuation and Choice. In particular, the sharp likelihood insensitivity at the heart of prospect theory shrinks substantially or disappears in Choice relative to Valuation. Conversely, the near linear utility typically estimated in Valuation is replaced by sharp curvature in Choice. It is the combination of these two fundamental descriptive divergences that generate the predictive failures of the fourfold pattern. Crucially, despite prospect theory's multi-parameter flexibility, it cannot account for these divergences.[22] Indeed, in our experiment Valuation and Choice are *identical* from the perspective of the theory, yet nonetheless produce radically different behaviors.[23]

Importantly, we argue that these predictive and descriptive failures are not special to prospect theory, but are instead bound to arise in *any* attempt to rationalize classic lottery anomalies on the basis of non-neoclassical preferences. In order to account for the rift between the two types of decisions, we must instead understand the way cognitive imperfections express differently in the two settings. We show that simple models that account for the cognitive frictions that differentially arise in the acts of Valuation and Choice can account for the key divergences that perennially arise between the two settings. Under the guidance of this model, we conduct another round of experiments and show that we can cause Choice and Valuation behaviors to converge simply by artificially forcing these frictions to be similar. These results not only explain the gap, but also establish that the anomalies inspiring prospect theory are themselves driven not by

[22] For instance, as we show in Online Appendix A.3, invoking differences in reference points across the two cases and attempting to explain the difference via loss aversion cannot "explain" the differences we measure.

[23] Our data shows basic failures of prospect theory when interpreted both as a predictive and a descriptive theory. A third interpretation of prospect theory is as a collection of decision-theoretic axioms that themselves have testable implications. We do not test these axioms but instead the predictive and descriptive implications the theory articulates for lottery Choice and Valuation. One rather extreme possible response to our findings is that, although we cast doubt on both the predictive and descriptive adequacy of the theory, because our tests are not axiomatic we have in some deep sense not in fact falsified the theory. Of course important tests have already falsified prospect theory's axioms in prior work, some conducted prior to the theory's creation (see e.g. Tversky, 1969, for violations of transitivity). Nonetheless, to whatever degree we agree to define prospect theory merely as a body of true axioms that make no falsifiable predictions for lottery Choice or Valuation, we are happy to stipulate that the theory is metaphysically alive and well.

novel preferences but rather by cognitive frictions that distort risky choice in systematic ways.

Ultimately, our results point not merely to a failure of prospect theory, but more generally a failure of any theory that attempts to explain lottery anomalies using descriptions of preferences. Because of this, our paper adds to a growing body of evidence suggesting that lottery choice is fundamentally shaped by human cognitive costs and constraints and the often idiosyncratic ways humans adapt to these limitations when making decisions. This in turn points to two primary implications of our work. First, substantively, because cognitive frictions seem to be first order drivers of risky choice, it is highly unlikely that we can infer much welfare-relevant information from risky choices. They simply contain little information about true preferences and therefore are normatively uninformative. Second, methodologically, it is likely that the most promising approach for explaining and predicting risky choice is models rooted in the structure of human cognition rather than in descriptions of novel preferences like those offered by prospect theory.

References

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

Agranov, Marina, and Pietro Ortoleva. 2017. "Stochastic choice and preferences for randomization." *Journal of Political Economy*, 125(1): 40–68.

Alós-Ferrer, Carlos, Johannes Buckenmaier, and Michele Garagnani. 2020. "Stochastic choice and preference reversals." Working Paper.

Bouchouicha, Ranoua, and Ferdinand M. Vieider. 2017. "Accommodating stake effects under prospect theory." *Journal of Risk and Uncertainty*, 55(1): 1–28.

Brañas-Garza, Pablo, Diego Jorrat, Antonio M Espín, and Angel Sánchez. 2023. "Paid and hypothetical time preferences are the same: Lab, field and online evidence." *Experimental Economics*, 26(2): 412–434.

Brañas-Garza, Pablo, Lorenzo Estepa-Mohedano, Diego Jorrat, Victor Orozco, and Ericka Rascón-Ramírez. 2021. "To pay or not to pay: Measuring risk preferences in lab and field." *Judgment and Decision Making*, 16(5): 1290–1313.

Brown, Alexander L, and Paul J Healy. 2018. "Separated decisions." *European Economic Review*, 101: 20–34.

Brown, Alexander L., Taisuke Imai, Ferdinand M. Vieider, and Colin F. Camerer. 2024. "Meta-Analysis of Empirical Estimates of Loss Aversion." *Journal of Economic Literature, forthcoming*, 63(3): 485–616.

Bruhin, Adrian, Helga Fehr-Duda, and Thomas Epper. 2010. "Risk and Rationality: Uncovering Heterogeneity in Probability Distortion." *Econometrica*, 78(4): 1375–1412.

Chapman, Jonathan, Erik Snowberg, Stephanie Wang, and Colin F. Camerer. 2024. "Looming large or seeming small? Attitudes towards losses in a representative sample." *Review of Economic Studies*, forthcoming.

Chapman, Jonathan, Mark Dean, Pietro Ortoleva, Erik Snowberg, and Colin Camerer. 2023*a*. "Econographics." *Journal of Political Economy Microeconomics*, 1(1): 115–161.

Chapman, Jonathan, Mark Dean, Pietro Ortoleva, Erik Snowberg, and Colin Camerer. 2023*b*. "Willingness to Accept, Willingness to Pay, and Loss Aversion." National Bureau of Economic Research.

Conte, Anna, John D. Hey, and Peter G. Moffatt. 2011. "Mixture models of choice under risk." *Journal of Econometrics*, 162(1): 79–88.

Crosetto, Paolo, and Antonio Filippin. 2015. "A theoretical and experimental appraisal of four risk elicitation methods." *Experimental Economics*, 1–29.

Cubitt, Robin P, Daniel Navarro-Martinez, and Chris Starmer. 2015. "On preference imprecision." *Journal of Risk and Uncertainty*, 50(1): 1–34.

Enke, Benjamin. 2024. "The Cognitive Turn in Behavioral Economics."

Enke, Benjamin, Thomas Graeber, Ryan Oprea, and Jeffrey Yang. 2024. "Behavioural Attenuation." Mimeo.

Enke, Benjamin, Uri Gneezy, Brian Hall, David Martin, Vadim Nelidov, Theo Offerman, and Jeroen Van De Ven. 2023. "Cognitive biases: Mistakes or missing stakes?" *Review of Economics and Statistics*, 105(4): 818–832.

Ert, Eyal, and Ido Erev. 2013. "On the descriptive value of loss aversion in decisions under risk: Six clarifications." *Judgment and Decision making*, 8(3): 214–235.

Etchart-Vincent, Nathalie, and Olivier L'Haridon. 2011. "Monetary Incentives in the Loss Domain and Behavior toward Risk: An Experimental Comparison of Three Reward Schemes Inclusing Real Losses." *Journal of Risk and Uncertainty*, 42: 61–83.

Fehr-Duda, Helga, Thomas Epper, Adrian Bruhin, and Renate Schubert. 2011. "Risk and rationality: The effects of mood and decision rules on probability weighting." *Journal of Economic Behavior & Organization*, 78(1): 14–24.

Feldman, Paul J, and Paul J Ferraro. 2023. "A Certainty Effect for Preference Reversals Under Risk: Experiment and Theory." *Mimeo.*

Freeman, David J, and Guy Mayraz. 2019. "Why choice lists increase risk taking." *Experimental Economics*, 22(1): 131–154.

Freeman, David J, Yoram Halevy, and Terri Kneeland. 2019. "Eliciting risk preferences using choice lists." *Quantitative Economics*, 10(1): 217–237.

Friedman, Daniel, R. Mark Isaac, Duncan James, and Shyam Sunder. 2017. *Risky curves: on the empirical failures of expected utility.* Routledge.

Friedman, Daniel, Sameh Habib, Duncan James, and Brett Williams. 2022. "Varieties of risk preference elicitation." *Games and Economic Behavior*, forthcoming.

Friedman, Milton, and L. J. Savage. 1948. "The Utility Analysis of Choices Involving Risk." *Journal of Political Economy*, 56(4): 279–304.

Frydman, Cary, and Lawrence J. Jin. 2023. "On the Source and instability of probability weighting." Working Paper.

Furukawa, Toshi A, Corrado Barbui, Andrea Cipriani, Paolo Brambilla, and Norio Watanabe. 2006. "Imputing missing standard deviations in meta-analyses can provide accurate results." *Journal of clinical epidemiology*, 59(1): 7–10.

Gneezy, Uri, Yoram Halevy, Brian Hall, Theo Offerman, and Jeroen van de Ven. 2024. "How Real is Hypothetical? A High-Stakes Test of the Allais Paradox."

Gonzalez, Richard, and George Wu. 1999. "On the Shape of the Probability Weighting Function." *Cognitive Psychology*, 38: 129–166.

Grether, David M, and Charles R Plott. 1979. "Economic theory of choice and the preference reversal phenomenon." *The American Economic Review*, 69(4): 623–638.

Harbaugh, William T, Kate Krause, and Lise Vesterlund. 2010. "The fourfold pattern of risk attitudes in choice and pricing tasks." *The Economic Journal*, 120(545): 595–611.

Hershey, John C., and Paul J. H. Schoemaker. 1985. "Probability versus Certainty Equivalence Methods in Utility Measurement: Are They Equivalent?" *Management Science*, 31(10): 1213–1231.

Hershey, John C., Howard C. Kunreuther, and Paul J. H. Schoemaker. 1982. "Sources of Bias in Assessment Procedures for Utility Functions." *Management Science*, 28(8): 936–954.

Holt, Charles A., and Susan K. Laury. 2002. "Risk Aversion and Incentive Effects." *American Economic Review*, 92(5): 1644–1655.

Kahneman, Daniel, and Amos Tversky. 1979. "Prospect Theory: An Analysis of Decision under Risk." *Econometrica*, 47(2): 263 – 291.

Khaw, Mel Win, Ziang Li, and Michael Woodford. 2021. "Cognitive imprecision and small-stakes risk aversion." *The Review of Economic Studies*, 88(4): 1979–2013.

Khaw, Mel Win, Ziang Li, and Michael Woodford. 2023. "Cognitive imprecision and stake-dependent risk attitudes." NBER Working Paper 30417.

Lévy-Garboua, Louis, Hela Maafi, David Masclet, and Antoine Terracol. 2012. "Risk aversion and framing effects." *Experimental Economics*, 15(1): 128–144.

L'Haridon, Olivier, and Ferdinand M. Vieider. 2019. "All over the map: A Worldwide Comparison of Risk Preferences." *Quantitative Economics*, 10: 185–215.

Lichtenstein, Sarah, and Paul Slovic. 1971. "Reversals of preference between bids and choices in gambling decisions." *Journal of experimental psychology*, 89(1): 46.

Loomes, Graham. 2005. "Modelling the stochastic component of behaviour in experiments: Some issues for the interpretation of data." *Experimental Economics*, 8: 301–323.

Markowitz, Harry. 1952. "The Utility of Wealth." *Journal of Political Economy*, 60(2): 151–158.

Mata, Rui, Renato Frey, David Richter, Jürgen Schupp, and Ralph Hertwig. 2018. "Risk Preference: A View from Psychology." *The Journal of Economic Perspectives*, 32(2): 155–172.

McGranaghan, Christina, Kirby Nielsen, Ted O'Donoghue, Jason Somerville, and Charles D Sprenger. 2024. "Distinguishing common ratio preferences from common ratio effects using paired valuation tasks." *American Economic Review*, 114(2): 307–347.

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

Oprea, Ryan. 2024*a*. "Complexity and its measurement." Mimeo.

Oprea, Ryan. 2024*b*. "Decisions Under Risk are Decisions Under Complexity." *American Economic Review, forthcoming.*

Oprea, Ryan, and Ferdinand M. Vieider. 2024. "Minding the Gap: On the Origins of Probability Weighting and the Description-Experience Gap." Mimeo.

Prat-Carrabin, Arthur, and Michael Woodford. 2022. "Efficient coding of numbers explains decision bias and noise." *Nature Human Behaviour*, 6(8): 1142–1152.

Schmidt, Ulrich, Chris Starmer, and Robert Sugden. 2008. "Third-generation prospect theory." *Journal of Risk and Uncertainty*, 36(3): 203–223.

Slovic, Paul. 1964. "Assessment of risk taking behavior." *Psychological Bulletin*, 61(3): 220–233.

Slovic, Paul, Dale Griffin, and Amos Tversky. 1990. "Compatibility effects in judgment and choice."

Stott, Henry P. 2006. "Cumulative Prospect Theory's Functional Menagerie." *Journal of Risk and Uncertainty*, 32: 101–130.

Sydnor, Justin. 2010. "(Over)insuring Modest Risks." *American Economic Journal: Applied Economics*, 2(4): 177–199.

Tversky, Amos. 1969. "Intransitivity of preferences." *Psychological review*, 76(1): 31.

Tversky, Amos, and Daniel Kahneman. 1992. "Advances in Prospect Theory: Cumulative Representation of Uncertainty." *Journal of Risk and Uncertainty*, 5: 297–323.

Tversky, Amos, Paul Slovic, and Daniel Kahneman. 1990. "The causes of preference reversal." *The American Economic Review*, 204–217.

Vickrey, William. 1945. "Measuring Marginal Utility by Reactions to Risk." *Econometrica*, 13(4): 319.

Vieider, Ferdinand M. 2023. "Decisions under Uncertainty as Bayesian Inference on Choice Options." *Management Science, forthcoming.*

Wakker, Peter, and Daniel Deneffe. 1996. "Eliciting von Neumann-Morgenstern Utilities When Probabilities Are Distorted or Unknown." *Management Science*, 42(8): 1131–1150.

Wakker, Peter P. 2010. *Prospect Theory for Risk and Ambiguity.* Cambridge:Cambridge University Press.

Wu, George, and Richard Gonzalez. 1996. "Curvature of the Probability Weighting Function." *Management Science*, 42(12): 1676–1690.

Zhou, Wenting, and John Hey. 2018. "Context matters." *Experimental economics*, 21: 723–756.

ONLINE APPENDIX

Is Prospect Theory Really a Theory of Choice?

Ranoua Bouchouicha Ryan Oprea Ferdinand M. Vieider Jilong Wu

Table of Contents

A Experiments: Additional results

A.1 Stimuli and results Experiment I

Table A.1 and Table A.2 present the choice proportions for the two treatment conditions by task in both the gain and loss experiments, respectively. These tables provide non-parametric tests on the differences in calculated choice proportions, which correspond

Table A.1: Choice proportions of risky option by treatment for CEs over gains

Task	Valuation	Choice	p-value	Task	Valuation	Choice	p-value
(16, 0.2; 0)	0.27	0.11	<0.01	(17, 0.5; 4)	0.43	0.34	<0.01
(16, 0.3; 0)	0.32	0.13	<0.01	(24, 0.1; 0)	0.19	0.06	<0.01
(16, 0.5; 0)	0.44	0.27	<0.01	(24, 0.5; 0)	0.42	0.26	<0.01
(16, 0.7; 0)	0.55	0.46	<0.01	(24, 0.9; 0)	0.71	0.66	<0.01
(16, 0.8; 0)	0.62	0.56	<0.01	(24, 0.4; 12)	0.44	0.31	<0.01
(16, 0.8; 0)	0.62	0.57	<0.01	(24, 0.6; 12)	0.47	0.43	<0.01
(16, 0.1; 4)	0.20	0.11	<0.01	(8, 0.2; 0)	0.32	0.12	<0.01
(16, 0.5; 4)	0.43	0.33	<0.01	(8, 0.5; 0)	0.50	0.32	<0.01
(16, 0.9; 4)	0.70	0.68	0.15	(8, 0.8; 0)	0.67	0.57	<0.01
(15, 0.5; 4)	0.45	0.35	<0.01	**(8, 0.8; 0)**	0.69	0.59	<0.01
(16, 0.5; 5)	0.42	0.35	<0.01				

List of choice tasks with choice proportions per treatment condition. The indicated p-values are based on two-sided Wilcoxon rank sum tests for difference between treatments.

Table A.2: Choice proportions of risky option by treatment for CEs over losses

Tasks	Valuation	Choice	p-value	Tasks	Valuation	Choice	p-value
(-16, 0.2; 0)	0.23	0.14	<0.01	(-17, 0.5; -4)	0.43	0.36	<0.01
(-16, 0.3; 0)	0.32	0.20	<0.01	(-24, 0.1; 0)	0.13	0.08	<0.01
(-16, 0.5; 0)	0.49	0.41	<0.01	(-24, 0.5; 0)	0.50	0.46	0.02
(-16, 0.7; 0)	0.64	0.72	<0.01	(-24, 0.9; 0)	0.80	0.86	<0.01
(-16, 0.8; 0)	0.69	0.78	<0.01	(-24, 0.4; -12)	0.34	0.26	<0.01
(-16, 0.8; 0)	0.69	0.78	<0.01	(-24, 0.6; -12)	0.49	0.44	0.05
(-16, 0.1; -4)	0.11	0.09	0.27	(-8, 0.2; 0)	0.17	0.13	0.05
(-16, 0.5; -4)	0.42	0.34	<0.01	(-8, 0.5; 0)	0.47	0.37	<0.01
(-16, 0.9; -4)	0.76	0.80	<0.01	(-8, 0.8; 0)	0.70	0.75	0.04
(-15, 0.5; -4)	0.44	0.34	<0.01	**(-8, 0.8; 0)**	0.68	0.75	<0.01
(-16, 0.5; -5)	0.40	0.34	<0.01				

List of choice tasks with choice proportions per treatment condition. The indicated p-values are based on two-sided Wilcoxon rank sum tests for difference between treatments.

to Figure 2. The tables further provide nonparametric tests for the proportions of risk taking across the treatment conditions.

A.2 Stimuli and results Experiment II

Table A.3 presents the choice proportions for the three treatment conditions by task (24, p; 0). This analysis employs nonparametric tests to assess the differences in calculated choice proportions. Notably, we observe that subjects' choices in Sequential-Choice and Valuation do not differ significantly (refer to Column *p-value (3)*), as illustrated in Figure 12. In contrast, it is evident that for small and moderate probabilities ($p = 0.1, 0.3, 0.5$), subjects answering the Choice task opted for fewer risky options compared to their counterparts in both the Sequential-Choice and Valuation conditions.

Table A.3: Choice proportions of risky option by treatment for CEs over (24, p; 0))

p	Choice	Sequential-Choice	Valuation	p-value (1)	p-value (2)	p-value (3)
0.1	0.085	0.210	0.235	<0.01	<0.01	0.368
0.3	0.155	0.320	0.321	<0.01	<0.01	0.799
0.5	0.289	0.438	0.431	<0.01	<0.01	0.780
0.7	0.496	0.540	0.540	0.434	0.648	0.688
0.9	0.690	0.692	0.717	0.906	0.311	0.308

List of choice tasks with choice proportions per treatment condition. The indicated p-values are based on two-sided Wilcoxon rank sum tests for difference between treatments. Specifically, Column *p-value (1)* indicates the difference between treatment Choice and Sequential-Choice, Column *p-value (2)* indicates the difference between treatment Choice and Valuation, and Column *p-value (3)* indicates the difference between treatment Sequential-Choice and Valuation.

A.3 Endogenous reference points

A key reason why Valuation tasks are the tool of choice to elicit PT parameters is that they allow to exogenously fix the reference point to 0. This is essential if one wants to separately identify all the different components of PT, since in the presence of endogenous reference points all wagers would become mixed. Hershey and Schoemaker (1985) claimed that varying a probability in a list while keeping the sure amount fixed could create a risk of endogenous reference dependence. Given that in such a case all wagers involve both gains and losses, PT can no longer be fully identified. In particular, it would no longer be possible to separately identify reference-dependence and rank-dependence (i.e., loss aversion and optimism/pessimism for gains and losses). One possibility is then that in binary choice the sure outcome may also act as an endogenous reference point, which would be troublesome for the identification of the full array of PT parameters. To test this, we rescale the PT equation following Hershey and Schoemaker (1985):

$$u(0) = w^{+}(p)u(x-s) - \lambda w^{-}(1-p)u(s),$$

where u is a reference-dependent utility function, w^{+} and w^{-} the probability weighting functions for gains and losses, respectively, and λ captures loss aversion. We have all the elements to identify utility function curvature and probability weighting from Valuation tasks for both gains and losses. This implies that we can infer the loss aversion coefficient that would explain the discrepancy between Valuation and Choice tasks from the following equation:

$$\lambda = \frac{w^{+}(p)}{w^{-}(1-p)} \frac{u(x-s)}{u(s)}. \tag{1}$$

The exact parameters will of course depend on assumptions made about functional forms and errors. We use a 'standard' PT implementation. That is, we estimate a simple aggregate PT model from the Valuation data, by letting $u(x) = \frac{1-exp(-\rho x)}{\rho}$, with different parameters for gains and losses, entered in terms of absolute amounts. Using an exponential utility function specification avoids issues in the identification of loss aversion when different utility function coefficients are estimated for gains and losses using CRRA functions (Köbberling and Wakker, 2005). The probability weighting function is $w(p) = \frac{\delta p^{\gamma}}{\delta p^{\gamma}+(1-p)^{\gamma}}$, again with different parameters for gains and losses. We specifically employ a flexible 2-parameter function to give the PT explanation its best possible shot. We estimate the model using Bayesian techniques in Stan (see section B.4 for an example of the code used). The priors used for the parameters are mildly regularizing, i.e. they are uninformative in the sense of being centred on neutral values ($\rho = 0$, $\delta = \gamma = 1$), and they are diffuse, in the sense that the standard deviation is chosen in a way as to include a large range of parameters into the possible range (e.g., for γ and δ, 95% of the probability mass is allocated to the interval between 0 and 7).

The estimations are executed based on a standard discrete choice Probit model, with a noise term defined on the value scale, and errors that are heteroscedastic depending on the length of a choice list. The choice probability is modeled as follows:

$$Pr[(x, p; y) \succ s] = \Phi \left[\frac{\pi u(x) + \tilde{\pi} u(y) - u(s)}{\sigma |x - y|} \right],$$

where Φ is the standard normal cumulative distribution function providing the 'link function', π indicates a decision weight, and $\tilde{\pi}$ is the decision weight associated to the complementary event. For gains and losses, $\pi = w(p)$ and $\tilde{\pi} = 1 - \pi$, whereas for mixed prospects $\pi = w^{+}(p)$ and $\tilde{\pi} = w^{-}(1-p)$, where + and − indicate the weighting function for losses respectively. The likelihood function is then constructed by mapping the choice patterns for the risky option into choice probabilities via a Bernoulli density.

Once we have obtained the PT parameters from the Valuation data for gains and losses, we calculate the choice objects p, $x - s$ and s for each of the three tasks in figure 2, panel A, in the main text. We then inject the PT parameters estimated from Valuations. Importantly, we do so using the entire vector of posterior draws for each of the parameters, which allows us to take the uncertainty in the parameter estimates into account, and thus to obtain credibility intervals for the estimates of loss aversion. Finally, we use the equations thus obtained to calculate the loss aversion parameter, λ, that would be needed to bridge the gap between Valuation and Choice for each of the three data points displayed in the figure (i.e., for $p = \{0.1, 0.5, 0.9\}$).

For wagers offering £24 with probability $p = \{0.1, 0.5, 0.9\}$ or else 0, we find that the loss aversion coefficient needed to explain the gap between Valuation and Choice for $p = 0.01$ is $\lambda_{0.1} = 3.05$ [2.93 ; 3.17]. The loss aversion coefficient for intermediate probabilities is $\lambda_{0.5} = 2.50$ [2.42 ; 2.55], and is thus significantly smaller. The loss aversion coefficient for large probabilities is $\lambda_{0.9} = 3.46$ [3.20 ; 3.75], which is larger than for both previous ones. A different coefficient is thus needed to close the gap for each probability, implying that loss aversion is not a viable explanation for the discrepancy between Valuation and Choice we observe. Feldman and Ferraro (2023) have furthermore shown that even the gap between certainty equivalents and probability equivalents cannot actually be organized by loss aversion.

A.4 IRRA over stakes

Panel A in Figure A.1 shows a measure of relative risk aversion in the Gain tasks, given by the choice proportion of the wager subtracted from the probability of winning $(p - \frac{\hat{c}-y}{x-y})$. Note that in the case of $y = 0$, a normalized certainty equivalent subtracted from the probability allows us to capture changes in relative risk aversion in the sense of Arrow-Pratt. We thus use tasks varying x, while keeping y fixed at 0 and p fixed at 0.5.

Figure A.1 shows the results. We see important level effects indicating generally higher levels of relative risk aversion in Choice than in Valuation. Patterns for Valuation tasks indicate the typical increasing relative risk aversion (IRRA) documented in the literature (Holt and Laury, 2002; Bouchouicha and Vieider, 2017; Di Falco and Vieider, 2022). A pattern of IRRA of similar magnitude is also observed in Choice tasks, thus pointing to the robustness of the phenomenon. Panel B shows changes in relative risk aversion with stake size in Loss tasks. Most measures are negative, indicating a tendency towards risk seeking. In Valuation we observe a pattern resembling the constant relative risk aversion documented for losses in previous studies (Fehr-Duda et al., 2010; Bouchouicha and Vieider, 2017). In binary choice, however, we observe clear evidence for increasing

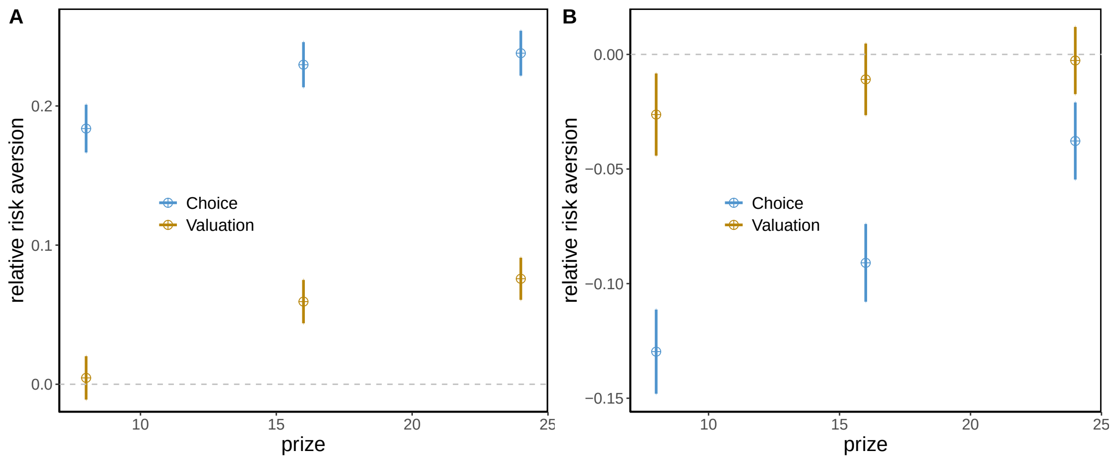


Figure A.1: Nonparametric measure of relative risk aversion by stake size

The figure plots a nonparametric measure of relative risk aversion, defined as $p - \frac{\widehat{c}}{y}$, where $\widehat{c}$ is the certainty equivalent resulting from the choices. The figure shows the evolution of the measure as the prize is increased from £8 to £24. Panel A shows the patterns for gains, and Panel B for losses.

relative risk aversion. In other words, whereas in Valuation tasks the patterns for losses tend to differ from the ones for gains, as also documented in the previous literature, in Choice we find convergent evidence for increasing relative risk aversion. Overall, increasing relative risk aversion is thus more robust in Choice than in Valuation.

A.5 Choice consistency and errors

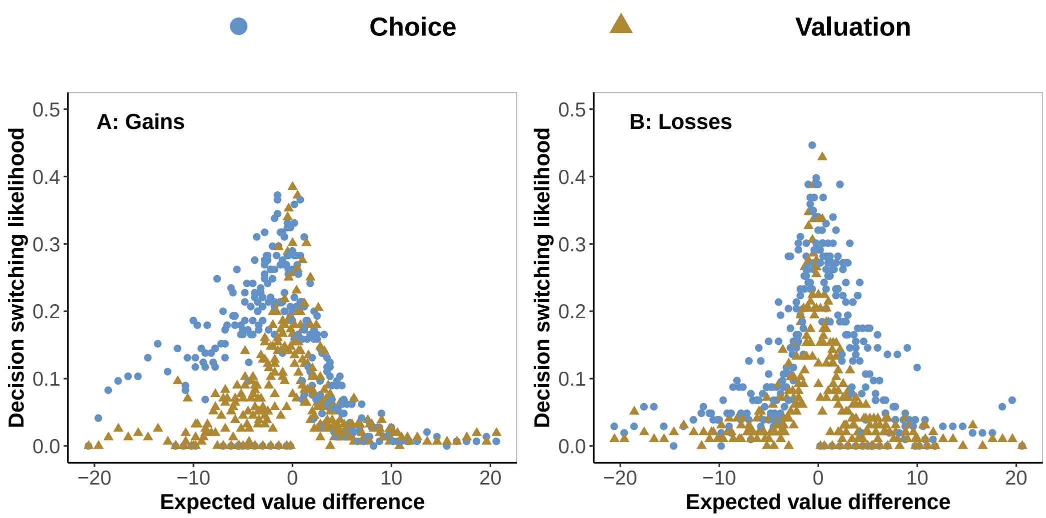


Figure A.2: Multiple Switching Likelihood for Valuation and Choice.

The figure compares subjects' tendency to switch from one option to another as the sure amount increases. The expected value difference on the x-axis is calculated as the sure amount minus the expect value of the prospect, and the y-axis shows switching frequencies with each increase in the sure outcome.

In the main text, we have shown how *violation* or *choice inconsistencies* across tasks are more severe in Valuation than Choice. Here, we show that within-list inconsistencies – something akin to 'multiple switching' – is more severe in Choice, but is nevertheless highly regular and concentrated around plausible points of indifference as shown in Figure A.2. As pointed out by Cubitt, Navarro-Martinez and Starmer (2015) and Agranov and Ortoleva (2017), such a pattern indicates those multiple switching is driven by indifference between two options.

B Meta-Analysis

B.1 Method

Paper selection. To substantiate our assertion that the absence of the four-fold pattern is not infrequent in the context of employing binary choices for the elicitation of risk preferences, we undertook a comprehensive meta-analysis. To ensure its unbiasedness, we rigorously selected relevant papers based on well-defined inclusion criteria. These criteria centered on the inclusion of experimental papers that estimated parameters of the probability weighting function under monetary risk, encompassing laboratory, lab-in-field, or online studies. Notably, this encompassed papers using pure binary choices or certainty equivalent choice lists as risk preference elicitation method in experiments. Regarding choice lists, specifically, we keep iterative certainty equivalent choice lists (e.g., Tversky and Kahneman, 1992) but drop those following bisection procedures (e.g., Abdellaoui, 2000). Our search for these papers primarily involved the scientific citation indexing database Web of Science. We initially screened titles and abstracts, and subsequently evaluated the remaining papers against our inclusion criteria, coding the relevant information. Additionally, we explored IDEAS/RePEc and Google Scholar for unpublished working papers to ensure a comprehensive review of the literature.

It is worth highlighting that certain studies encompass multiple estimations of prospect theory (or rank-dependent utility) parameters, owing to the adoption of various model specifications, the introduction of multiple treatment arms, or the examination of subsample variations. To ensure uniformity and facilitate comparative analysis across studies, we have employed the following filtering criteria:

- In the case of studies implementing diverse model specifications, characterized by varying utility functions and probability weighting functions, our initial step involves excluding those employing non-Constant Relative Risk Aversion (non-CRRA) utility functions. This step is imperative as our meta-analysis relies on the imputation of the certainty equivalent of a specific lottery $(100, 0.1; 0)$, and non-CRRA utility functions (e.g., exponential functions) can exert quantitative and even qualitative impacts on the calculated results. Subsequently, to align with the objectives of our meta-analysis, we prioritize the selection of estimations that either directly report the standard error or provide information that enables the approximate calculation of the standard error. Lastly, we opt for the estimation that is predominantly discussed or initially presented in the main body of the text.

- In instances where studies involve multiple treatments or explore specific population subsets, our focus is on selecting estimations associated with the control conditions or groups; Otherwise we are unable to know whether the presence or absence of the fourfold pattern is caused by the treatment or by population heterogeneity. For example, we do not incorporate observations stemming from treatments such as the sampling treatment in (Glöckner et al., 2016) or the outcome feedback treatment (Haffke and Hübner, 2014) in our analysis.

In terms of the availability of standard errors, a crucial element for our meta-analysis, it's noteworthy that 30.5% of estimations (N=36) within our dataset lack information that would be sufficient to calculate standard errors, including the seminal work by Tversky and Kahneman (1992). Among the remaining observations, a substantial majority, accounting for 76.8% (N=63), explicitly provide standard errors or other statistical metrics that allow for the precise calculation of standard errors, such as the inclusion of 95% confidence intervals. Additionally, there are 14 observations that present statistics such as 95% credible intervals, which, without certain assumptions, cannot be utilized directly in our analysis. In these cases, we adopt a conservative approach to ensure data retention while managing the potential inaccuracies. For example, when confronted with a study that exclusively reports the maximum and minimum values of estimated parameters, we calculate a conservative standard deviation as $(Max - Min)/4$. This method enables us to include the data while mitigating the impacts of imprecision. Apart from the estimates of our Experiment I, our final dataset includes 76 papers and 141 PT estimates, as listed in Subsection B.5.

Predictive certainty equivalents and associated standard error calculation

With the collected data, we first calculate the predicted certainty equivalent for each observation, $\widehat{c} = u^{-1}[w(p)u(x)]$, where u designates the utility function, w the probability weighting function, and u^{-1} is the inverse of the utility function. We use $x = 100$ and $p = 0.1$, but the qualitative results do not change much for different monetary outcomes or even smaller probabilities. To obtain the standard errors of such predicted certainty equivalents, we then apply a bootstrap procedure. Specifically, we assume that each PT parameter is normally distributed around their mean estimate with variance equal to the squared standard error of the parameter encoded from the papers. We then draw 4000 samples from these parameter distributions to obtain a vector of certainty equivalents of $(100, 0.1)$ for each study. We then use this vector to obtain the standard error associated with the predicted certainty equivalent.

Some studies that do not report any standard errors for their PT parameter estimates. nor any information from which tsuch errors could be reasonably approximated. As a result, we could not apply the bootstrap procedure to these studies. Given this situation, one option might be to drop these observations from the meta-analysis. This would, however,result in the loss of a substantial number of observations, including the seminal study of Tversky and Kahneman (1992). Also, it is possible that these studies which do not report standard errors are different than those indeed reporting. This implies that dropping these incomplete observations could lead to biased conclusions. Due to these reasons, we choose to impute the standard errors of the predicted certainty equivalent for those incomplete observations.

The approach we take involves estimating the parameters characterizing their distribution in the data from the equation $\log(se_o) \sim \mathcal{N}\left(\mu_{se}, \sigma_{se}^2\right)$, where using the log ensures that we only impute positive values. Utilizing these distributional parameters, we then imputed the missing values in SE by modeling $\log(se_m) \sim \mathcal{N}\left(\widehat{\mu}_{se}, \widehat{\sigma}_{se}^2\right)$, where the subscripts o and m denote *observed* and *missing*, respectively. The parameters $\left(\widehat{\mu}_{se}, \widehat{\sigma}_{se}^2\right)$ represent the estimated quantities. In implementing this estimation, we will initially obtain values for the missing observations in standard errors (SE) that maintain the same mean and variance. However, our approach can be significantly improved by identifying variables within our dataset that are strongly associated with SEs. The most effective predictors of SEs in our data include the predicted certainty equivalent, the dummy indicator whether the experiment is conducted in a lab, the dummy indicator whether the experiment is conducted in the field, the number of subjects, and the dummy indicator of whether the adopted PWF has two parameters. We thus conduct the imputation by defining $\mu_{se} = \alpha_{se} + \beta_{se} \times \mathbf{Z}$, where $\mathbf{Z}$ represents the vector of these optimal predictors.

Meta-analysis: Bayesian hierarchical estimation. Consider the dataset $(\widehat{c}_i, se_i)$, where $\widehat{c}_i$ is the imputed certainty equivalent of the ith observation in the dataset and the associated se_i quantifies the uncertainty around it. We assume that the $\widehat{c}_i$ is normally distributed around the parameter $\widetilde{c}_i$:

$$\widehat{c}_i \sim \mathcal{N}\left(\widetilde{c}_i, se_i^2\right), \tag{2}$$

where the variability of $\widehat{c}_i$ around the true but latent mean $\widetilde{c}_i^p$ is supposed to stem from sampling variation in small studies, as captured by the known standard error se_i. This is indeed a central feature of meta-analysis or indeed of measurement error models in general – see Vieider (2024) for a discussion and a tutorial.

Sampling variation contributes to the observed variability in $\widetilde{c}_i$, but it's not the only source; there may also be "genuine" heterogeneity across measurements, perhaps due to differing experimental settings, different subject pools, etc. To account for this, we assume that the study-level $\widetilde{c}_i$ follow a normal distributionacross all studies:

$$\widetilde{c}_i \sim \mathcal{N}\left(\mu, \tau^2\right), \tag{3}$$

where μ represents the meta-analytic mean of the imputed certainty equivalents, and τ represents the standard deviation of the true, latent certainty equivalents across studies. Incorporating variation across estimates due to observable characteristics, commonly known as meta-regression, can be achieved simply by defining $\mu = \boldsymbol{X\beta}$, where $\boldsymbol{X}$ is a matrix of study characteristics, including a column vector of 1s, and $\boldsymbol{\beta}$ is a vector of regression coefficients.

We estimate our models in Stan (Carpenter et al., 2017), executed from `R` (R Core Team, 2023) through `CmdStanR` (Gabry et al., 2024). Population-level parameter priors are selected to be mildly regularizing, providing informative yet broad ranges significantly larger than the expected estimates from data analysis. Lower-level parameter priors are derived from these estimated population-level parameters, ensuring a cohesive modeling framework.

B.2 Results

Meta-analysis Table B.1 presents the meta-analysis results for four groups. The absolute values of the average predictive certainty equivalent (CE) for the wager (100, 0.1) in the Valuation studies are above 0.1 for both gains and losses, with both results being significant within the 95% CrI. Conversely, the absolute values for the Choice studies are significantly below 0.1 for both gains and losses.

Table B.1: Meta-analysis results of normalized predictive CEs

Group	Mean	SD (100, 0.1)	2.5%	97.5%	Mean	SD (100, 0.9)	2.5%	97.5%
Valuation-gains	0.180	0.010	0.160	0.201	0.721	0.014	0.693	0.748
Choice-gains	0.040	0.006	0.029	0.051	0.746	0.022	0.701	0.788
Valuation-losses	0.212	0.022	0.170	0.256	0.764	0.014	0.736	0.790
Choice-losses	0.074	0.011	0.054	0.097	0.784	0.033	0.714	0.842

Figure B.1 and Figure B.2 present the funnel plots of the calculated predictive CEs for gains and losses, respectively. Our primary focus here is on Figure B.1, which indicates that there is no significant publication bias favoring the four-fold pattern for both groups of studies.

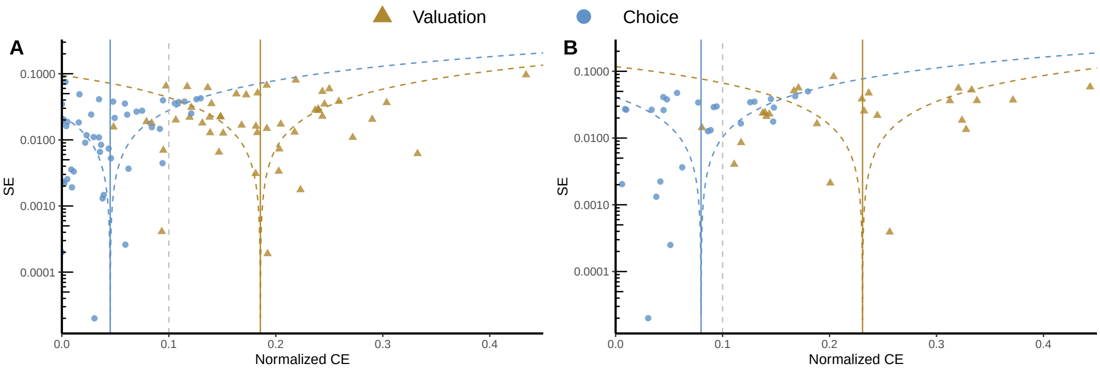


Figure B.1: Funnel plots of inferred CEs for a wager $(100, 0.1)$ and $(-100, 0.1)$
Funnel plots of calculated normalized CEs and associated standard errors. Panel A scatters CEs-SEs observations for the gain wager $(100, 0.1)$ inferred from studies using certainty equivalents or binary choices to measure risk attitudes, whereas panel B shows CEs-SEs observations for the loss wager $(-100, 0.1)$. The vertical solid lines mark the mean of normalized CEs from each condition, while the dashed gray line indicates the neutrality status (normalized CE $= p = 0.1$). Two dashed curves delineate the boundaries for a statistically significant deviation from the "true" CEs value. The y-axis is presented in log scale for improved visualization.

Meta-regression Our meta-regression aggregates all predictive certainty equivalents

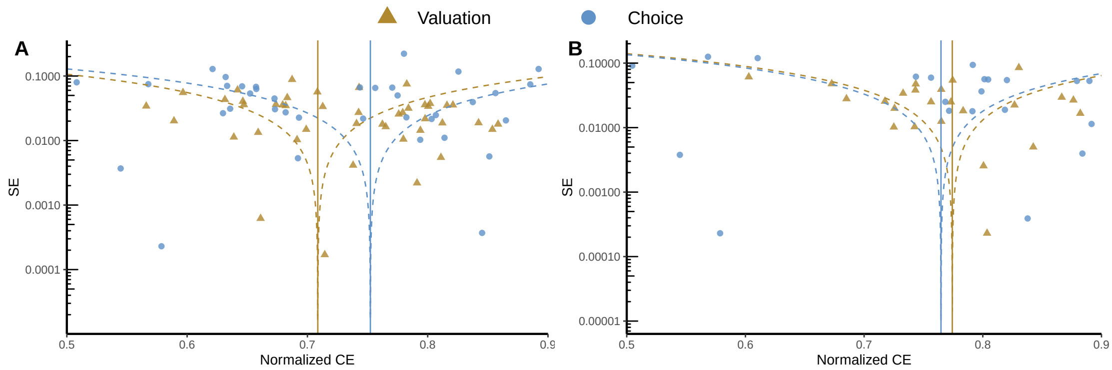


Figure B.2: Funnel plots of inferred CEs for a wager $(100, 0.9)$ and $(-100, 0.9)$
Funnel plots of calculated normalized CEs and associated standard errors. Panel A scatters CEs-SEs observations for the gain wager $(100, 0.9)$ inferred from studies using certainty equivalents or binary choices to measure risk attitudes, whereas panel B shows CEs-SEs observations for the loss wager $(-100, 0.9)$. The vertical solid lines mark the mean of normalized CEs from each condition, while the dashed gray line indicates the neutrality status (normalized CE $= p = 0.9$). Two dashed curves delineate the boundaries for a statistically significant deviation from the "true" CEs value. The y-axis is presented in log scale for improved visualization.

(CEs) and incorporates a range of covariates. These include indicators for whether the method used was valuation or choice (*Valuation*), whether the payoff domain was positive or negative (*Loss*), whether the incentive was real or hypothetical (*Hypothetical*), whether each choice included a sure amount option (*Sure*), and whether the experimental environment was a field experiment (*Field*). Additionally, the regression includes two interaction terms: one between the method dummy and payoff domain, and another between the payoff domain and incentive type. Finally, the regression incorporates other characteristics related to the stimuli, such as the largest amount and the highest probability in the lottery outcomes. For the wager (100, .1), Table B.2 indicates that the coefficient for the *Valuation* dummy variable is 11.171 and for the dummy variable *Loss* is 3.964, both showing significant positive impacts. No other variables significantly influenced the predictive CEs. The table provides the estimation results for the predictive CE of (100, 0.9). For this predictive CE, we do not see a significant difference between the two elicitation methods, namely Valuation and Choice.

B.3 Robustness

B.3.1 Different standard error imputation methods

To verify the robustness of our meta-analysis results concerning the calculation of predictive certainty equivalents' standard error (SE), this subsection introduces two alternative imputation methods to get standard errors for those incomplete observations. Specif-

Table B.2: Meta Regression of Predictive CE of (100, 0.1) and (100, 0.9)

Variable	Mean	SD	2.5%	97.5%	Mean	SD	2.5%	97.5%
		(100, 0.1)				(100, 0.9)		
Valuation	11.39	2.02	7.41	15.25	-4.23	3.25	-9.42	-1.24
Loss	3.24	1.44	0.39	6.11	-1.45	4.88	-6.67	4.70
Hypothetical	-2.10	1.85	-5.76	1.46	-6.14	2.86	-9.34	-1.80
sure_presence	2.25	1.96	-1.56	6.09	3.86	2.24	1.79	7.37
Log(Stiml_high_Amt)	-0.01	0.36	-0.70	0.72	0.27	0.56	-0.61	0.93
Log(Stiml_high_P)	7.93	4.77	-1.19	17.93	-4.32	13.25	-21.69	9.46
Stiml_Num_outcome	-0.81	0.75	-2.30	0.65	-0.32	1.60	-1.97	2.19
Field	1.97	2.35	-2.49	6.77	-0.67	4.45	-5.89	6.25
CE*Loss	1.06	2.17	-3.27	5.26	1.25	3.80	-4.32	5.08
Loss*Hypothetical	-2.47	2.63	-7.55	2.74	5.73	6.29	-5.09	10.54
Cons	7.12	2.27	2.62	11.68	80.84	1.91	77.85	83.08

Note: This table presents a meta-regression of predictive certainty equivalents (CEs), incorporating the following variables: *Valuation*: Indicator variable (1 = valuation method, 0 = choice method) used to distinguish whether certainty equivalents or binary choices were applied. *Loss*: Indicator variable (1 = negative payoff domain, 0 = positive payoff domain), representing whether the study dealt with losses or gains. *Hypothetical*: Indicator variable (1 = hypothetical incentives, 0 = real incentives), denoting whether monetary incentives were hypothetical or real. *Sure*: Indicator variable (1 = sure option included, 0 = no sure option), capturing whether a sure amount option was provided in the choice set. *Log(Stimul_high_Amt)*: The natural logarithm of the largest monetary amount presented in the stimuli (lottery). *Log(Stimul_high_P)*: The natural logarithm of the highest probability used in the lottery outcomes. *Stimul_Num_outcome*: The number of outcomes in the stimuli (lottery). *Field*: Indicator variable (1 = field experiment, 0 = laboratory experiment), distinguishing between field and laboratory settings. *Valuation*Loss*: Interaction term between valuation method and the payoff domain (loss or gain). *Loss*Hypothetical*: Interaction term between the payoff domain (loss or gain) and the incentive type (hypothetical vs. real).

ically, the first method uses Bruhin, Fehr-Duda and Epper (2010) as the benchmark study and calculate the SD for other studies as follows:

$$SE_i = SE^* \cdot \frac{\sqrt{N^*}}{\sqrt{N_i}}, \tag{4}$$

where SE^* and N^* represent the bootstrapped standard error and the number of subjects from the "*Zurich-03*" study in Bruhin, Fehr-Duda and Epper (2010), and N_i is the number of subjects in study i. Table B.3 reports the resulting estimates with using the newly derived standard errors for the wager (100, 0.1) and (100, 0.9), which are very close to those we report in the main text.

The second SE calculation method instead takes the average standard deviation of all studies from a same group (Choice or Valuation), denoted as $SD_{average}$. Then, for every single incomplete observation, we can derive its predictive CE's standard error as below:

$$SE_i = \frac{SD_{average}}{\sqrt{N_i}}, \tag{5}$$

Table B.3: Meta-analysis Results with the first different standard error calculation method

Group	Mean	SD	2.5%	97.5%	Mean	SD	2.5%	97.5%
	(100, 0.1)				(100, 0.9)			
Valuation-gains	0.180	0.010	0.160	0.201	0.725	0.014	0.697	0.753
Choice-gains	0.042	0.006	0.030	0.053	0.759	0.022	0.715	0.803
Valuation-losses	0.224	0.021	0.185	0.266	0.773	0.014	0.744	0.801
Choice-losses	0.076	0.011	0.056	0.097	0.805	0.028	0.744	0.856

Table B.4: Meta-analysis Results with the second different standard error calculation method

Group	Mean	SD	2.5%	97.5%	Mean	SD	2.5%	97.5%
	(100, 0.1)				(100, 0.9)			
Valuation-gains	0.182	0.010	0.162	0.202	0.727	0.014	0.699	0.754
Choice-gains	0.042	0.006	0.030	0.055	0.760	0.022	0.717	0.803
Valuation-losses	0.226	0.021	0.185	0.268	0.774	0.014	0.746	0.802
Choice-losses	0.077	0.011	0.056	0.099	0.804	0.029	0.742	0.856

where N_i is the number of subjects for study i. Table B.4 reports the results after adopting these differently derived standard errors for the wager (100, 0.1) and (100, 0.9). Again, these results are essentially similar to those reported in the main text, and thus showing that our findings of the meta analysis are robust to different standard error calculation approaches.

B.3.2 Different Wagers

To demonstrate the robustness of our conclusion with respect to the wager employed for the predictive CE calculation, we have re-estimated using an alternative wager (200, 0.05) and (20, 0.05) respectively. Tables B.5 presents these estimates, which are largely consistent with our primary conclusions detailed in the main text, namely the absence of the four-fold pattern in the Choice condition.

B.4 Structural Estimation of PT parameters

Methods and code We recover PT parameters from aggregate Bayesian estimations using Stan. In our preferred specification reported in the main text, we use a power utility function, $u(x) = x^{\rho}$. Following the majority of the papers in our meta-analysis, we estimate a 1-parameter specification of the probability weighting function proposed by Prelec (1998), which takes the form $w(p) = exp(-(-ln(p))^{\gamma})$, where values of $\gamma < 1$ cap-

Table B.5: Meta-analysis Results for the inferred CE of different wager

Group	Mean	SD	2.5%	97.5%	Mean	SD	2.5%	97.5%
	(200, 0.05)				(20, 0.05)			
Valuation-gains	0.130	0.009	0.112	0.149	0.133	0.009	0.116	0.152
Choice-gains	0.018	0.003	0.012	0.025	0.019	0.003	0.013	0.025
Valuation-losses	0.157	0.020	0.118	0.198	0.161	0.020	0.122	0.201
Choice-losses	0.038	0.007	0.025	0.051	0.038	0.007	0.026	0.052

ture likelihood insensitivity. Estimation results using alternative 1-parameter functions, such as the one of Tversky and Kahneman, produce very similar results. Estimations obtained using 2-parameter functions are discussed farther below.

We estimate the functionals using purpose-coded models in Stan, which we launch from R. Vieider (2024) provides a detailed tutorial on estimations of structural models in Stan. We use the following code:

```
data {
  int<lower=1> N;
  array[N] real high;
  array[N] real low;
  array[N] real sure;
  array[N] real p;
  array[N] int choice_risky;
}
parameters {
  real rho;
  real<lower=0> gamma;
  real<lower=0> sigma;
}
model {
  vector[N] pw;
  vector[N] pv;
  vector[N] udiff;

  rho ~ normal(1 , 0.5);
  gamma ~ normal(1 , 0.5);
  sigma ~ normal(0 , 0.5);

  for (i in 1:N) {
    pw[i] = exp(-(-log(p[i]))^gamma);
```

```
    pv[i] = pw[i] * pow(high[i], rho) + (1 - pw[i]) * pow(low[i], rho);
    udiff[i] = (pv[i] - pow(sure[i], rho)) / (sigma * (high[i] - low[i]));
  }

  choice_risky ~ bernoulli_logit(udiff);
}
```

We also estimate PT parameters with the 2-parameter version of probability weighting function, $w(p) = exp(-\delta(-ln(p))^{\gamma})$. The code looks as follows:

```
data {
  int<lower=1> N;
  array[N] real high;
  array[N] real low;
  array[N] real sure;
  array[N] real p;
  array[N] int choice_risky;
}
parameters {
  real rho;
  real<lower=0> gamma;
  real<lower=0> delta;
  real<lower=0> sigma;
}
model {
  vector[N] pw;
  vector[N] pv;
  vector[N] udiff;

  rho ~ normal(1 , 0.5);
  gamma ~ normal(1 , 0.5);
  delta ~ normal(1 , 0.5);
  sigma ~ normal(0 , 0.5);

  for (i in 1:N) {
    pw[i] = exp(-delta*(-log(p[i]))^gamma);
    pv[i] = pw[i] * pow(high[i], rho) + (1 - pw[i]) * pow(low[i], rho);
    udiff[i] = (pv[i] - pow(sure[i], rho)) / (sigma * (high[i] - low[i]));
  }

  choice_risky ~ bernoulli_logit(udiff);
```

}

Estimation results Table B.6 and B.7 report PT parameters' estimation results for Experiment I and II respectively. Table B.6 includes four groups: Choice/Valuation * Gains/Losses while Table B.7 includes the three treatments in the gain domain.

Table B.6: Prospect theory parameters estimation results - Experiment I

Group	ρ	SD_ρ	γ	SD_γ	δ	SD_{delta}	σ	SD_σ
				$Prelec-I$				
Valuation-gains	0.93	0.01	0.56	0.01			0.12	0.00
Choice-gains	0.56	0.01	0.71	0.01			0.04	0.00
Valuation-gains	0.91	0.01	0.83	0.01			0.13	0.00
Choice-gains	0.76	0.01	1.05	0.02			0.07	0.00
				LLO				
Valuation-gains	1.00	0.01	0.55	0.01	0.78	0.01	0.24	0.01
Choice-gains	0.81	0.01	0.89	0.01	0.47	0.01	0.15	0.01
Valuation-gains	1.13	0.02	0.86	0.02	0.69	0.02	0.42	0.03
Choice-gains	1.17	0.02	1.39	0.03	0.51	0.01	0.44	0.02

Table B.7: Prospect theory parameters estimation results - Experiment II

Parameter	Mean	SD	2.5%	97.5%	Mean	SD	2.5%	97.5%
	$Prelec-I$				$Prelec-II$			
Valuation								
ρ	0.857	0.023	0.814	0.901	0.626	0.069	0.487	0.763
γ	0.649	0.028	0.594	0.705	0.614	0.028	0.562	0.670
δ					0.594	0.107	0.385	0.816
σ	0.139	0.010	0.120	0.160	0.061	0.016	0.033	0.097
Sequential Choice								
ρ	0.926	0.017	0.894	0.959	0.818	0.064	0.688	0.949
γ	0.530	0.019	0.494	0.568	0.525	0.019	0.491	0.562
δ					0.798	0.112	0.583	1.035
σ	0.119	0.007	0.107	0.133	0.084	0.019	0.051	0.129
Choice								
ρ	0.571	0.012	0.546	0.595	0.596	0.048	0.505	0.694
γ	0.750	0.024	0.707	0.799	0.758	0.027	0.708	0.815
δ					1.070	0.121	0.842	1.322
σ	0.033	0.001	0.031	0.036	0.038	0.007	0.026	0.053

B.5 Collected Studies for Meta-analysis

Andersen, Steffen, Glenn W. Harrison, Morten I. Lau, and E. Elisabet Rutström. 2018. "Multiattribute Utility Theory, Intertemporal Utility, And Correlation Aversion." *International Economic Review*, 59: 537–555.

Andersen, Steffen, John Fountain, Glenn W. Harrison, and E. Elisabet Rutström. 2014. "Estimating subjective probabilities." *Journal of Risk and Uncertainty*, 48: 207–229.

Andreoni, James, and Charles Sprenger. 2011. "Uncertainty Equivalents: Testing the Limits of the Independence Axiom."

Baillon, Aurélien, and Lætitia Placido. 2019. "Testing constant absolute and relative ambiguity aversion." *Journal of Economic Theory*, 181: 309–332.

Baillon, Aurélien, Han Bleichrodt, and Vitalie Spinu. 2020. "Searching for the reference point." *Management Science*, 66: 93–112.

Baláž, Vladimír, Viera Bacova, Eva Drobna, Katarina Dudekova, and Kamil Adamik. 2013. "Testing Prospect Theory Parameters." *Ekonomický časopis*, 61: 655–671.

Beauchamp, Jonathan P., Daniel J. Benjamin, David I. Laibson, and Christopher F. Chabris. 2020. "Measuring and controlling for the compromise effect when estimating risk preference parameters." *Experimental Economics*, 23: 1069–1099.

Bernheim, B. Douglas, and Charles Sprenger. 2020. "On the Empirical Validity of Cumulative Prospect Theory: Experimental Evidence of Rank-Independent Probability Weighting." *Econometrica*, 88: 1363–1409.

Birnbaum, Michael H, and Alfredo Chavez. 1997. "Tests of Theories of Decision Making: Violations of Branch Independence and Distribution Independence." *Organizational Behavior and Human Decision Processes*, 71: 161–194.

Bouchouicha, Ranoua, and Ferdinand M. Vieider. 2017. "Accommodating stake effects under prospect theory." *Journal of Risk and Uncertainty*, 55: 1–28.

Bougherara, Douadia, Lana Friesen, and Céline Nauges. 2021. "Risk Taking with Left- and Right-Skewed Lotteries*." *Journal of Risk and Uncertainty*, 62: 89–112.

Brandstätter, Eduard, Anton Kühberger, and Friedrich Schneider. 2002. "A Cognitive-Emotional Account of the Shape of the Probability Weighting Function." *Journal of Behavioral Decision Making*, 15: 79–100.

Bruhin, Adrian, Helga Fehr-Duda, and Thomas Epper. 2010. "Risk and Rationality: Uncovering Heterogeneity in Probability Distortion." *Econometrica*, 78: 1375–1412.

Bruhin, Adrian, Luís Santos-Pinto, and David Staubli. 2018. "How do beliefs about skill affect risky decisions?" *Journal of Economic Behavior and Organization*, 150: 350–371.

Carpio, María Bernedo Del, Francisco Alpizar, and Paul J. Ferraro. 2022. "Time and risk preferences of individuals, married couples and unrelated pairs." *Journal of Behavioral and Experimental Economics*, 97: 101794.

Charupat, Narat, Richard Deaves, Travis Derouin, Marcelo Klotzle, and Peter Miu. 2013. "Emotional balance and probability weighting." *Theory and Decision*, 75: 17–41.

Choi, Syngjoo, Jeongbin Kim, Eungik Lee, and Jungmin Lee. 2022. "Probability Weighting and Cognitive Ability." *Management Science*, 68: 5201–5215.

Conte, Anna, John D. Hey, and Peter G. Moffatt. 2011. "Mixture models of choice under risk." *Journal of Econometrics*, 162: 79–88.

Conte, Anna, M. Vittoria Levati, and Chiara Nardi. 2018. "Risk Preferences and the Role of Emotions." *Economica*, 85: 305–328.

Coricelli, Giorgio, Enrico Diecidue, and Francesco D. Zaffuto. 2018. "Evidence for multiple strategies in choice under risk." *Journal of Risk and Uncertainty*, 56: 193–210.

Enke, Benjamin, and Cassidy Shubatt. 2023. "Quantifying Lottery Choice Complexity."

Epper, Thomas, Helga Fehr-Duda, and Adrian Bruhin. 2011. "Viewing the future through a warped lens: Why uncertainty generates hyperbolic discounting." *Journal of Risk and Uncertainty*, 43: 169–203.

Fan, Yuyu, David V. Budescu, and Enrico Diecidue. 2019. "Decisions with compound lotteries." *Decision*, 6: 109–133.

Fehr-Duda, Helga, Adrian Bruhin, Thomas Epper, and Renate Schubert. 2010. "Rationality on the rise: Why relative risk aversion increases with stake size." *Journal of Risk and Uncertainty*, 40: 147–180.

Fehr-Duda, Helga, Manuele De Gennaro, and Renate Schubert. 2006. "Gender, financial risk, and probability weights." *Theory and Decision*, 60: 283–313.

Fehr-Duda, Helga, Thomas Epper, Adrian Bruhin, and Renate Schubert. 2011. "Risk and rationality: The effects of mood and decision rules on probability weighting." *Journal of Economic Behavior and Organization*, 78: 14–24.

Gao, Xiaoxue Sherry, Glenn W. Harrison, and Rusty Tchernis. 2023. "Behavioral welfare economics and risk preferences: a Bayesian approach." *Experimental Economics*, 26: 273–303.

Glatt, Markus, Roy Brouwer, and Ivana Logar. 2019. "Combining Risk Attitudes in a Lottery Game and Flood Risk Protection Decisions in a Discrete Choice Experiment." *Environmental and Resource Economics*, 74: 1533–1562.

Glöckner, Andreas, and Thorsten Pachur. 2012. "Cognitive models of risky choice: Parameter stability and predictive accuracy of prospect theory." *Cognition*, 123: 21–32.

Glöckner, Andreas, Baiba Renerte, and Ulrich Schmidt. 2020. "Violations of coalescing in parametric utility measurement." *Theory and Decision*, 89: 471–501.

Glöckner, Andreas, Benjamin E. Hilbig, Felix Henninger, and Susann Fiedler. 2016. "The reversed description-experience gap: Disentangling sources of presentation format effects in risky choice." *Journal of Experimental Psychology: General*, 145: 486–508.

Gonzalez, Richard, and George Wu. 1999. "On the Shape of the Probability Weighting Function." *Cognitive Psychology*, 38: 129–166.

Gonzalez, Richard, and George Wu. 2022. "Composition rules in original and cumulative prospect theory." *Theory and Decision*, 92: 647–675.

Haffke, Peter, and Ronald Hübner. 2014. "Effects of different feedback types on information integration in repeated monetary gambles." *Frontiers in Psychology*, 5.

Hajimoladarvish, Narges. 2018. "How do people reduce compound lotteries?" *Journal of Behavioral and Experimental Economics*, 75: 126–133.

Hajimoladarvish, Narges. 2021. "Explaining Heterogeneity in Risk Preferences Using a Finite Mixture Model." *Journal of Money and Economy*, 16: 533–554.

Harbaugh, William T, Kate Krause, and Lise Vesterlund. 2002. "Risk Attitudes of Children and Adults: Choices Over Small and Large Probability Gains and Losses." *Experimental Economics*, 5: 53–84.

Harrison, Glenn W., and J. Todd Swarthout. 2019. "Eye-tracking and economic theories of choice under risk." *Journal of the Economic Science Association*, 5: 26–37.

Harrison, Glenn W., Andre Hofmeyr, Don Ross, and J. Todd Swarthout. 2018. "Risk Preferences, Time Preferences, and Smoking Behavior." *Southern Economic Journal*, 85: 313–348.

Harrison, Glenn W., Morten I. Lau, and Hong Il Yoo. 2020. "Risk attitudes, sample selection, and attrition in a longitudinal field experiment." *Review of Economics and Statistics*, 102: 552–568.

Hey, John D., Andrea Morone, and Ulrich Schmidt. 2009. "Noise and bias in eliciting preferences." *Journal of Risk and Uncertainty*, 39: 213–235.

Hsu, Ming, Ian Krajbich, Chen Zhao, and Colin F. Camerer. 2009. "Neural response to reward anticipation under risk is nonlinear in probabilities." *Journal of Neuroscience*, 29: 2231–2237.

Kellen, David, Markus D. Steiner, Clintin P. Davis-Stober, and Nicholas R. Pappas. 2020. "Modeling choice paradoxes under risk: From prospect theories to sampling-based accounts." *Cognitive Psychology*, 118.

Kellen, David, Thorsten Pachur, and Ralph Hertwig. 2016. "How (in)variant are subjective representations of described and experienced risk and rewards?" *Cognition*, 157: 126–138.

Kusev, Petko, Paul van Schaik, Peter Ayton, John Dent, and Nick Chater. 2009. "Exaggerated Risk: Prospect Theory and Probability Weighting in Risky Choice." *Journal of Experimental Psychology: Learning Memory and Cognition*, 35: 1487–1505.

Lejarraga, Tomás, Thorsten Pachur, Renato Frey, and Ralph Hertwig. 2016. "Decisions from Experience: From Monetary to Medical Gambles." *Journal of Behavioral Decision Making*, 29: 67–77.

l'Haridon, Olivier, and Ferdinand M. Vieider. 2019. "All over the map: A worldwide comparison of risk preferences." *Quantitative Economics*, 10: 185–215.

Maafi, Hela. 2011. "Preference reversals under ambiguity." *Management Science*, 57: 2054–2066.

Murad, Zahra, Martin Sefton, and Chris Starmer. 2016. "How do risk attitudes affect measured confidence?" *Journal of Risk and Uncertainty*, 52: 21–46.

Murphy, Ryan O., and Robert H.W.Ten Brincke. 2018. "Hierarchical maximum likelihood parameter estimation for cumulative prospect theory: Improving the reliability of individual risk parameter estimates." *Management Science*, 64: 308–326.

Newell, Anthony. 2020. "Is your heart weighing down your prospects? Interoception, risk literacy and prospect theory."

O'Brien, Megan K., and Alaa A. Ahmed. 2014. "Take a stand on your decisions, or take a sit: Posture does not affect risk preferences in an economic task." *PeerJ*, 2014.

Pachur, Thorsten, Michael Schulte-Mecklenbeck, Ryan O. Murphy, and Ralph Hertwig. 2018. "Prospect theory reflects selective allocation of attention." *Journal of Experimental Psychology: General*, 147: 147–169.

Pachur, Thorsten, Yaniv Hanoch, and Michaela Gummerum. 2010. "Prospects behind bars: Analyzing decisions under risk in a prison population." *Psychonomic Bulletin and Review*, 17: 630–636.

Panidi, Ksenia, Alicia Nunez Vorobiova, Matteo Feurra, and Vasily Klucharev. 2022. "Dorsolateral prefrontal cortex plays causal role in probability weighting during risky choice." *Scientific Reports*, 12: 16115.

Patalano, Andrea L., Alexandra Zax, Katherine Williams, Liana Mathias, Sara Cordes, and Hilary Barth. 2020. "Intuitive symbolic magnitude judgments and decision making under risk in adults." *Cognitive Psychology*, 118.

Patalano, Andrea L., Jason R. Saltiel, Laura Machlin, and Hilary Barth. 2015. "The role of numeracy and approximate number system acuity in predicting value and probability distortion." *Psychonomic Bulletin and Review*, 22: 1820–1829.

Ring, Patrick, Catharina C. Probst, Levent Neyse, Stephan Wolff, Christian Kaernbach, Thilo van Eimeren, Colin F. Camerer, and Ulrich Schmidt. 2018. "It's all about gains: Risk preferences in problem gambling." *Journal of Experimental Psychology: General*, 147: 1241–1255.

Scheibehenne, Benjamin, and Thorsten Pachur. 2015. "Using Bayesian hierarchical parameter estimation to assess the generalizability of cognitive models of choice." *Psychonomic Bulletin and Review*, 22: 391–407.

Schulreich, Stefan, Yana G. Heussen, Holger Gerhardt, Peter N.C. Mohr, Ferdinand C. Binkofski, Stefan Koelsch, and Hauke R. Heekeren. 2014. "Music-evoked incidental happiness modulates probability weighting during risky lottery choices." *Frontiers in Psychology*, 4.

Spitmaan, Mehran, Emily Chu, and Alireza Soltani. 2019. "Salience-driven value construction for adaptive choice under risk." *Journal of Neuroscience*, 39: 5195–5209.

Stott, Henry P. 2006. "Cumulative prospect theory's functional menagerie." *Journal of Risk and Uncertainty*, 32: 101–130.

Sun, Qingzhou, Evan Polman, and Huanren Zhang. 2021. "On prospect theory, making choices for others, and the affective psychology of risk." *Journal of Experimental Social Psychology*, 96: 104177.

Suter, Renata S., Thorsten Pachur, and Ralph Hertwig. 2016. "How Affect Shapes Risky Choice: Distorted Probability Weighting Versus Probability Neglect." *Journal of Behavioral Decision Making*, 29: 437–449.

Suter, Renata S., Thorsten Pachur, Ralph Hertwig, Tor Endestad, and Guido Biele. 2015. “The neural basis of risky choice with affective outcomes.” *PLoS ONE*, 10.

Takemura, Kazuhisa, and Hajime Murakami. 2016. “Probability Weighting Functions Derived from Hyperbolic Time Discounting: Psychophysical Models and Their Individual Level Testing.” *Frontiers in Psychology*, 7.

Tsang, Ming. 2020. “Estimating uncertainty aversion using the source method in stylized tasks with varying degrees of uncertainty.” *Journal of Behavioral and Experimental Economics*, 84.

Tversky, Amos, and Daniel Kahneman. 1992. “Advances in Prospect Theory: Cumulative Representation of Uncertainty.” *Journal of Risk and Uncertainty*, 5(4): 297–323.

Vieider, Ferdinand M., Abebe Beyene, Randall Bluffstone, Sahan Dissanayake, Zenebe Gebreegziabher, Peter Martinsson, and Alemu Mekonnen. 2018. “Measuring Risk Preferences in Rural Ethiopia.” *Economic Development and Cultural Change*, 66: 417–446.

Vieider, Ferdinand M., Clara Villegas-Palacio, Peter Martinsson, and Milagros Mejía. 2016. “Risk Taking For Oneself And Others: A Structural Model Approach.” *Economic Inquiry*, 54: 879–894.

Vieider, Ferdinand M., Peter Martinsson, Pham Khanh Nam, and Nghi Truong. 2019. “Risk preferences and development revisited.” *Theory and Decision*, 86: 1–21.

Vieider, Ferdinand M., Thorsten Chmura, Tyler Fisher, Takao Kusakawa, Peter Martinsson, Frauke Mattison Thompson, and Adewara Sunday. 2015. “Within- versus between-country differences in risk attitudes: implications for cultural comparisons.” *Theory and Decision*, 78: 209–218.

Wu, George, and Richard Gonzalez. 1996. “Curvature of the Probability Weighting Function.” *Management Science*, 42: 1676–1690.

Wu, J. George, and Alex B. Markle. 2008. "An empirical test of gain-loss separability in prospect theory." *Management Science*, 54: 1322–1335.

Wölbert, Eva, and Arno Riedl. 2013. "Measuring Time and Risk Preferences: Reliability, Stability, Domain Specificity."

Zhou, Wenting, and John Hey. 2018. "Context matters." *Experimental Economics*, 21: 723–756.

C A noisy coding model of Choice versus Valuation

Here, we formalize the intuition presented in the main text by sketching a stylized noisy coding model. However, we stress that the intuition described in the main text is considerably more general than many of the specific modelling choices and assumptions we make here. Indeed, any noisy coding model in which sequential evaluation results in an accumulation of coding noise will deliver key insights like those we derive here. For instance, Khaw, Li and Woodford (2023) present a closely related model of valuation built on somewhat different detail-level modeling choices, but which shares some of the key predictions we derive here.

Lottery evaluation

We discuss choices between a lottery $(x, p; y)$ and a sure outcome c, following the setup in all our experiments. We start by describing the evaluation of the lottery, i.e. of the log-odds in favour of winning a prize. To introduce noisy cognition, we assume that the log-odds are not directly accessible to the DM but are noisily coded, i.e. they are mentally represented by a signal r_p. On average, this signal will be unbiased, i.e. it will reflect the true log-odds shown to the decision maker. In any specific instance, however, the signal may be affected by some noise, which we assume to be normally distributed. We thus obtain the following likelihood function encoding a given probability p:

$$r_p \sim \mathcal{N}\left(ln\left(\frac{p}{1-p}\right), \nu_p^2\right),$$

where ν_p^2 is the variance of the coding errors.

To decode the signal – to rein in the errors incurring in the coding process – the signal is decoded by combination with a learned prior, capturing the statistics of the environment. Note that this prior does not necessarily correctly reflect those statistics, either because few stimuli have been encountered as of yet (as is the case at the beginning of an experiment, where the full stimulus distribution will only be known once the experiment is over), or because of the influence of a hyperprior summarizing experiences across different environments. Another possibility is that positive versus negative experiences may have asymmetric impacts on learning (see Vieider, 2023*a*, for a formal model of learning in the present context, and a discussion of why learning will necessarily be affected by noise). In particular, if the prior mean is "too pessimistic" relative to the stimuli presented, then this will results in risk averse choices.

Decoding the log-odds signal by a conjugate normal prior, which takes the form $\mathcal{N}(ln(\eta), \sigma_p^2)$, will yield the following posterior distribution:

$$ln \frac{p}{1-p} \,|\, r_p \sim \mathcal{N}\left(\gamma\, r_p + (1-\gamma)\, ln(\eta)\,,\, \frac{\nu^2 \sigma_p^2}{\nu^2 + \sigma_p^2}\right). \tag{6}$$

For simplicity, we will henceforth normalize the prior SD to 1 by dividing all parameters by σ_p^2. The upshot is that we reduce the system by one parameter without losing any information (see Natenzon, 2019, for an analogous simplification). The coding noise becomes $\widehat{\nu}_p = \frac{\nu_p}{\sigma_p}$, thus capturing the coding noise relative to the prior variance of the log-odds. The Bayesian evidence weight is then defined as $\gamma \triangleq \frac{1}{\widehat{\nu}_p^2+1} = \frac{1}{\nu_p^2/\sigma_p^2+1}$.

To make this expression observable to the econometrician, we further condition the posterior mean on many repetitions of the same stimulus p. The variation of responses across repeated presentations of the same stimulus yields the following observable expression, referred to as the *response distribution* (Ma, Kording and Goldreich, 2023):

$$\mathbb{E}\left[\gamma\, r_p + (1-\gamma)\, ln(\eta) \,\middle|\, p\right] \;\sim\; \mathcal{N}\left(\gamma\, ln\left(\frac{p}{1-p}\right) + (1-\gamma)\, ln(\eta)\,,\, \gamma^2 \nu_p^2\right). \tag{7}$$

Proof. Let $z \sim \mathcal{N}(\widehat{z}\,,\, \tau^2)$. From the well-known properties of the normal distributions it follows that $bz + a \sim \mathcal{N}(b\widehat{z} + a\,,\, b^2\tau^2)$. The result above obtains by letting $b = \gamma$, $z = r_p$, $a = (1-\gamma)\, ln(\eta)$, $\widehat{z} = ln\left(\frac{p}{1-p}\right)$, and $\tau = \nu_p$. □

The average inference on the log-odds displayed above is now systematically shaded or shrunk towards the mean of the prior. Intuitively, this happens because the signal is only taken into account in proportion to the 'confidence' the DM has in the signal, as captured by its precision (the inverse of the coding noise variance, $\widehat{\nu}^{-2}$, since we can alternatively define $\gamma = \frac{\widehat{\nu}^{-2}}{1+\widehat{\nu}^{-2}}$). This results in *systematic bias* in the evaluation of the log-odds, which as we will see shortly is at the origin of probability distortions (see Oprea and Vieider, 2024, for an experimental test of this mechanism). Indeed, substituting $\delta \triangleq \eta^{1-\gamma}$ into the expectation of the normal distribution above yields a linear in log-odds probability weighting function that is frequently used in the prospect theory literature (Gonzalez and Wu, 1999; Bruhin, Fehr-Duda and Epper, 2010).

C.1 Binary Choice

In binary Choice, the posterior for the log-odds derived above is simply compared to the posterior for the comparative outcomes, given by the log cost-benefits, $ln\left(\frac{c-y}{x-c}\right)$ (this derives from a choice rule that maximizes expected value – see Vieider, 2023, for a detailed discussion). Importantly, we assume that the log-odds of the lottery and the log-cost benefits are evaluated simultaneously and independently. This seems indeed natural, inasmuch as there is no reason in binary Choice why the evaluation of one dimension should be conditioned on the other. Given this independence in evaluations, the signal for the log cost-benefits will once again be unbiased, having as its mean the true log cost-benefits, but potentially deviating from them in any single draw from the following distribution:

$$r_o \sim \mathcal{N}\left(ln\left(\frac{c-y}{x-c}\right),\, \nu_o^2\right),$$

where ν_o is the standard deviation of outcome coding noise.

The evaluation of the log cost-benefits then proceeds much like for the log-odds above (see Vieider, 2023, for details). For simplicity, we assume that costs and benefits are expected to be equal in the prior, so that the mean of the log cost-benefits is 0. This seems a natural assumption, and it is made without loss of generality, as the results presented below will generalize to setups with a non-degenerate prior for outcomes. The response distribution will now take the following form:

$$\mathbb{E}\left[\alpha\, r_o \,|\, c, y, x\,\right] \;\sim\; \mathcal{N}\left(\alpha\, ln\left(\frac{c-y}{x-c}\right),\, \alpha^2 \widehat{\nu}_o^2\right), \tag{8}$$

where similarly to above $\widehat{\nu}_o = \frac{\nu_o}{\sigma_o^2}$ is the normalized outcome coding noise, and $\alpha \triangleq \frac{1}{\widehat{\nu}_o^2+1}$

is the outcome discriminability parameter.

Following an EV maximization rule that is often employed in signal detection theory (Green, Swets et al., 1966; Gold and Shadlen, 2001), we assume that the log-odds are traded off against the log cost-benefits to reach a decision. Given that the objective quantities are not available to the decision maker, she decides instead based on her posterior inferences on those quantities. On average, this will result in the following stochastic choice rule observable to the econometrician:

$$Pr[(x,p;y) \succ c] = \Phi\left[\frac{\alpha^{-1}\left[\gamma\, ln\left(\frac{p}{1-p}\right) + (1-\gamma)\, ln(\eta)\right] - ln\left(\frac{c-y}{x-c}\right)}{\sqrt{\alpha^{-2}\gamma^2\widehat{\nu}_p^2 + \widehat{\nu}_o^2}}\right]. \tag{9}$$

Proof. The proof proceeds as in Vieider (2023). Re-arrange the threshold equation in (10) by multiplying both sides by α^{-1}. The rest of the proof proceeds as usual. □

Given that $\alpha^{-1} > 1$ for any $\nu_o > 0$, noise in outcome assessments counteracts noise in probability assessments. Likelihood insensitivity is driven by $\gamma/\alpha < 1$, and will thus appear whenever probability coding noise exceeds outcome coding noise, $\widehat{\nu}_p > \widehat{\nu}_o$, resulting in $\gamma < \alpha$. A second implication is that – as long as $\eta < 1$, indicating a "risk-averse prior" as typically found in experiments (Vieider, 2023; Oprea and Vieider, 2024) – any $\alpha < 1$ (any noise in outcome perceptions) will reduce the upweighting of $\eta < 1$ towards 1 (towards 0 for $ln(\eta)$) produced by the power $1-\gamma$. An interesting special case occurs when coding noise is exactly equal for probabilities and outcomes: In this case, there will be no probability distortions, and decisions will be an expression of the 'true' prior mean $ln(\eta)$. Interestingly, this case can occur even in the presence of considerable coding noise, as long as the level of that noise is the same for probabilities and outcomes.

C.2 Valuation

In Valuation tasks the decision situation is quite different. DMs are confronted with a list of varying sure outcomes, which is compared to an unchanging lottery, which is prominently displayed. This makes it natural to assume that the lottery is evaluated first, and that the point of indifference is found in a second stage conditional on the evaluation of the lottery. The first stage will then result in an evaluation of the log-odds just as described above. It is in the second-stage evaluation – which now consists in finding an indifference value – that the differences with Choice situations will emerge.

The second stage evaluation now consists in finding the value of the sure amount that equalizes the log cost-benefits with the posterior inference on the log-odds, call it c^*. This fundamentally changes the decision problem, since the DM now no longer tries to infer the true log-cost benefits from an unbiased signal as in Choice, but rather tries to identify the (noisy signal for) the sure amount that produces indifference between the posterior log-odds and the signal for the log-cost benefits. As we will see shortly, this implies that the outcome signal is now no longer unbiased as in Choice, but is itself affected by systematic bias, since it centered on the posterior for the log-odds, which is be definition a biased quantity in the presence of coding noise.

Technically, the Valuation task now takes the form of a search for the noisy outcome signal (out of the vector of signals $\boldsymbol{r_o}$ in the list) that minimizes the absolute difference between the log cost-benefit signal and the posterior of the log odds, i.e.

$$r_o^* \,|\, c^* = \arg\min_{\boldsymbol{r_o}|\boldsymbol{c}} \left| \boldsymbol{r_o} - [\gamma\, r_p + (1-\gamma)\, ln(\eta)\,] \right|. \tag{10}$$

This minimization problem will result in the selection of r_o^* such that the average difference with the posterior mean of the lottery evaluation is 0. At the point of indifference, we will thus observe the following relation on average:

$$r_o^* - [\gamma\, r_p + (1-\gamma)\, ln(\eta)] \sim \mathcal{N}\left(0\,,\, \nu_o^2 + \frac{\widehat{\nu}_p^2}{\widehat{\nu}_p^2+1}\right), \tag{11}$$

where the variance is the sum of the coding noise variance of the outcome signal and the variance of the posterior distribution of the log-odds in (6). It follows that

$$r_o^* \sim \mathcal{N}\left(\gamma\, r_p + (1-\gamma)\, ln(\eta)\,,\, \nu_o^2 + \frac{\widehat{\nu}_p^2}{\widehat{\nu}_p^2+1}\right). \tag{12}$$

An important insight results from this equation. Instead of an outcome signal that is an unbiased estimator of the true log cost-benefits, as in choice, the DM now chooses a signal that is centered on the posterior inference of the log-odds, which is a systematically biased quantity. This, in turn, will result in the accumulation of probability coding noise with outcome coding noise in the process of finding an indifference value.

To see this technically, we start again by deriving the posterior distribution. To simplify notation, we now define $\widetilde{\nu}_o^2 \triangleq \nu_o^2 + \frac{\nu_p^2}{\nu_p^2+1}$. Conditional on the noisy signal r_o^*, the posterior distribution takes the following form:

$$ln\left(\frac{c^* - y}{x - c^*}\right) \,|\, r_o^* \sim \mathcal{N}\left(\beta\, r_o^*\,,\, \frac{\widetilde{\nu}_o^2}{\widetilde{\nu}_o^2 + 1}\right), \tag{13}$$

where $\beta \triangleq \frac{1}{\widetilde{\nu}_o^2+1}$. This equation assumes again implicitly (and without loss of generality) that costs and benefits are equal in the prior on average, so that the prior mean drops out of the equation (or equivalently, that the cost-benefit prior has mean 0).

To make the equation above observable to the experimenter, we can reformulate it in terms of the *response distribution*, i.e. the distribution conditional on repeated presentations of the same choice stimulus (given that r_p and r_o^* are stochastic due to coding noise). To achieve this, we condition on two quantities: 1) the expectation of the probability response distribution in (7), which we omit from the notation below to avoid clutter; and 2) on the vector of sure amounts:

$$\mathbb{E}\left[\mathbb{E}\left(ln\left(\frac{c^* - y}{x - c^*}\right) \,|\, r_o^*\right) |\, \boldsymbol{c}, x, y\right] \sim \mathcal{N}\left(\beta\left[\gamma\, ln\left(\frac{p}{1-p}\right) + (1-\gamma) ln\,(\eta)\right],\, \widetilde{\nu}_o^2 + \beta^2\gamma^2\widehat{\nu}_p^2\right). \tag{14}$$

This results in an empirically estimable equation describing valuations for lotteries (here described as the density around the indifference value, as typically modelled when estimating valuation data, see e.g. Gonzalez and Wu, 1999; Bruhin, Fehr-Duda and Epper, 2010; L'Haridon and Vieider, 2019).

Proof. From (12), $\mathbb{E}\left[r_o^* \,|\, \mathbb{E}\left[ln\left(\frac{p}{1-p}\right) \,|\, r_p\right]\right] = \gamma\, r_p + (1-\gamma)\, ln(\eta)$. We next make this observable step by step. We first condition on the probability response distribution in (7) to obtain $\mathbb{E}\left[r_o^* \,|\, \mathbb{E}\left[ln\left(\frac{p}{1-p}\right) \,|\, p\right]\right] = \gamma\, ln\left(\frac{p}{1-p}\right) + (1-\gamma)\, ln(\eta)$, with notation in (14) simplified by directly conditioning on p. This brings the variance to $\widetilde{\nu}_o^2 + \gamma^2\widehat{\nu}_p^2$. Finally, we exploit the posterior distribution of the log cost-benefits in (13) and condition on the vector of sure amounts $\boldsymbol{c}$ to obtain the expectation in (14). The proof for the variance proceeds as above and is not repeated here. □

Equation (14) produces the key elements distinguishing Valuation from Choice. The outcome discriminability or 'shrinkage' weight β is now applied to the posterior of the log-odds (instead of to the true log cost-benefits, as in Choice), implying that

1. the discriminability weight attributed to the true log-odds, γ, is now attenuated in the presence of outcome coding noise, since the log-odds now receive a weight $\beta\gamma$, which is smaller than the weight γ/α in Choice – i.e., likelihood insensitivity will the more pronounced in Valuation than in Choice;

2. The weight attributed to the prior is now also attenuated, since it is given by $\beta(1-\gamma)$, which is smaller than $(1-\gamma)/\alpha$ in Choice. Any value η will thus be "compressed" towards 1 (any prior mean $ln(\eta)$ will be compressed towards 0), which in the presence of risk aversion in the prior (captured by $\eta < 1$) implies an increase in risk taking under Valuation relative to Choice (a *decrease* in risk taking in the presence of an optimistic prior for losses);

3. Given that $\widetilde{\nu}_o > \widehat{\nu}_o$ for any $\nu_p > 0$, we can see that the coding noise at the indifference point will be larger than the coding noise for log cost-benefits in binary Choice. This, jointly with errors arising mainly at the list level, implies that between-task consistency will be lower in Valuation compared to Choice.

D Experimental materials

For each specific experiment survey, please refer:

- Experiment I
  - Valuation and Choice - Gains
  - Valuation and Choice - Losses
- Experiment II
  - Benchmark-Valuation and Benchmark-Choice
  - Treatment-Sequential

Below we show the experiment materials of Experiment I with the condition - BCs in gains - as an example.

[Page 1: Attention Pledge]

Attention Pledge

We ask that you give us your full attention throughout the study. Please refrain from all other activities, including using your phone and browsing the internet. If we find that you are not paying attention or are violating any rules you will be dismissed without being paid.

Specifically, by continuing with the study you declare:

- I will be available for the full-time of the study
- I will devote my full attention to the experiment and will not engage in other activities, such as browsing the internet
- I will put my mobile devices in silent mode and I will not use them during the study
- I will not communicate with any other participants during the session

I accept these requirements and wish to participate this study.

I reject these requirements and wish to leave this study.

[Page break here]

Welcome to our study. This study will take around 45 minutes. You will be asked to complete **several different tasks and a short questionnaire**. The answer to the questionnaire and all your decisions on the tasks will be private, and cannot be traced back to you personally.

If you complete the study you will be paid a fixed participation fee of **£7**. On top of that, you may earn additional money based on the decisions that you take during the study. Your participation in this study is voluntary. You have the right to withdraw at any point during the study. However, **if you do not complete the study, you will not be paid.**

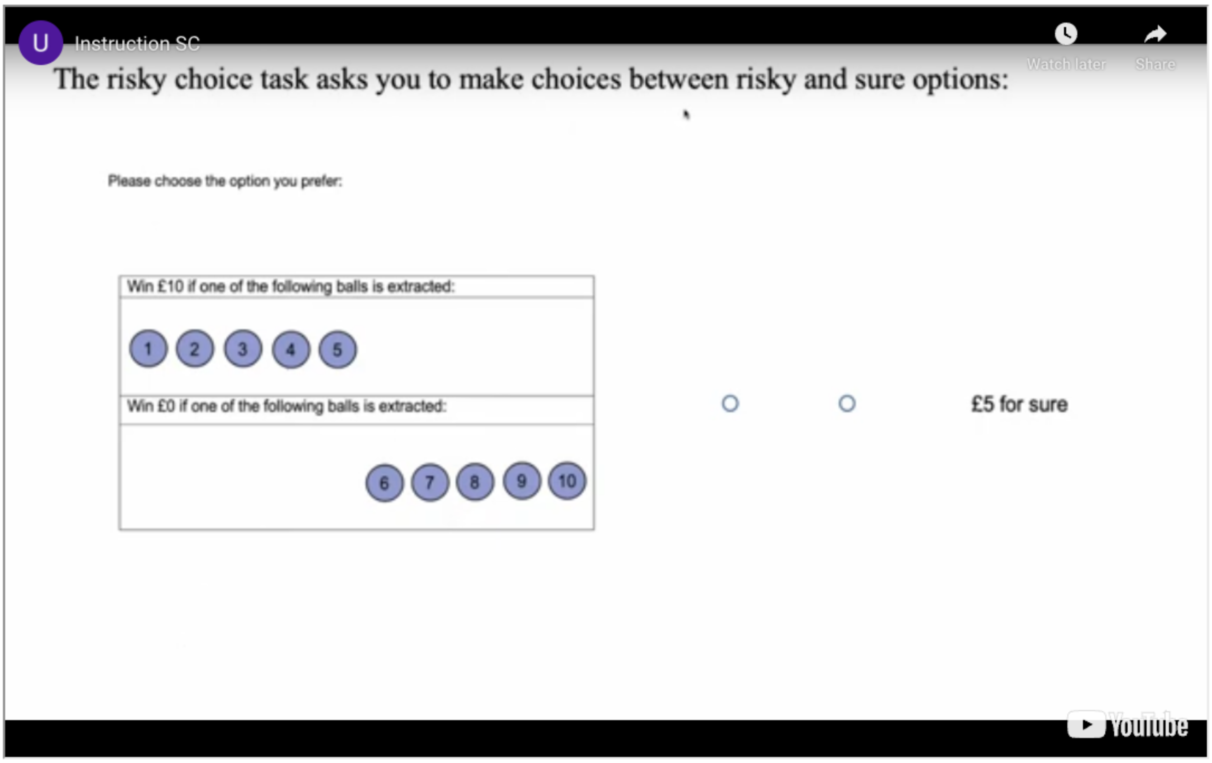
Please consider each decision carefully. Remember that your final payoffs from this experiment depend on the decisions you make.

By clicking the next button below, the instruction video of the risky choice task will be played.

Next

[Page 2: Video Playing]

Please watch the instruction video:



If you have any uncertainty about the content of the video, please feel free to watch it again.

[Page 3: Comprehension Check]

Comprehension Test

Before starting the experimental tasks, please answer the three questions to confirm your comprehension of the study. Please note that if you fail to answer the three questions correctly, the study will be terminated, and you will have to leave the study with being only paid £0.5.

Q1: Please choose the correct option:

You will only get paid with a fixed participation fee.

Even if you withdraw your participation halfway, you would still get the participation fee and additional payouts from two tasks respectively.

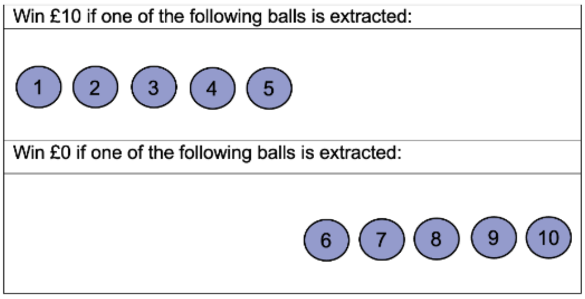
After completing the study, you will be paid an immediate participation fee. In addition, you might get additional payout(s) if you are selected by the system.

Q2: Please choose the correct option:

For the additional payout of the risky task, if you are selected by the system:

- The amount is randomly determined, irrespective of your responses.
- One of your choices is randomly picked by the system to play for real money, so it is important to carefully make every single choice.
- The amount is determined by a specific choice which you know in advance, so you only need to answer that question seriously.

Q3: For the below example choice,



£9 for sure

what outcomes would you have if you selected the lottery option?

Please choose the correct option:

- You have 5/10 chance of winning £10, and 5/10 chance of winning £0
- You have 9/10 chance of winning £9, and 1/10 chance of winning £1
- You will receive £9 for sure.

By clicking the NEXT button, you confirm and submit your answer. If all questions are correctly answered, you will proceed to answer the risky tasks. Good luck!

Appendix Reference

Agranov, Marina, and Pietro Ortoleva. 2017. "Stochastic choice and preferences for randomization." *Journal of Political Economy*, 125(1): 40–68.

Bouchouicha, Ranoua, and Ferdinand M. Vieider. 2017. "Accommodating stake effects under prospect theory." *Journal of Risk and Uncertainty*, 55(1): 1–28.

Bruhin, Adrian, Helga Fehr-Duda, and Thomas Epper. 2010. "Risk and Rationality: Uncovering Heterogeneity in Probability Distortion." *Econometrica*, 78(4): 1375–1412.

Carpenter, Bob, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. "Stan: A probabilistic programming language." *Journal of Statistical Software*, 76(1): 1–32.

Cubitt, Robin P, Daniel Navarro-Martinez, and Chris Starmer. 2015. "On preference imprecision." *Journal of Risk and Uncertainty*, 50(1): 1–34.

Di Falco, Salvatore, and Ferdinand M. Vieider. 2022. "Environmental Adaptation of Risk Preferences." *Economic Journal*, 132(648): 2737–2766.

Fehr-Duda, Helga, Adrian Bruhin, Thomas F. Epper, and Renate Schubert. 2010. "Rationality on the Rise: Why Relative Risk Aversion Increases with Stake Size." *Journal of Risk and Uncertainty*, 40(2): 147–180.

Feldman, Paul J, and Paul J Ferraro. 2023. "A Certainty Effect for Preference Reversals Under Risk: Experiment and Theory." *Mimeo.*

Gabry, Jonah, Rok Češnovar, Andrew Johnson, and Steve Bronder. 2024. "cmdstanr: R Interface to 'CmdStan'." R package version 0.8.1, https://discourse.mc-stan.org.

Glöckner, Andreas, Benjamin E Hilbig, Felix Henninger, and Susann Fiedler. 2016. "The reversed description-experience gap: Disentangling sources of presentation format effects in risky choice." *Journal of Experimental Psychology: General*, 145(4): 486.

Gold, Joshua I, and Michael N Shadlen. 2001. "Neural computations that underlie decisions about sensory stimuli." *Trends in cognitive sciences*, 5(1): 10–16.

Gonzalez, Richard, and George Wu. 1999. "On the Shape of the Probability Weighting Function." *Cognitive Psychology*, 38: 129–166.

Green, David Marvin, John A Swets, et al. 1966. *Signal detection theory and psychophysics.* Vol. 1, Wiley New York.

Hershey, John C., and Paul J. H. Schoemaker. 1985. "Probability versus Certainty Equivalence Methods in Utility Measurement: Are They Equivalent?" *Management Science*, 31(10): 1213–1231.

Holt, Charles A., and Susan K. Laury. 2002. "Risk Aversion and Incentive Effects." *American Economic Review*, 92(5): 1644–1655.

Khaw, Mel Win, Ziang Li, and Michael Woodford. 2023. "Cognitive imprecision and stake-dependent risk attitudes." NBER Working Paper 30417.

Köbberling, Veronika, and Peter P. Wakker. 2005. "An index of loss aversion." *Journal of Economic Theory*, 122(1): 119 – 131.

L'Haridon, Olivier, and Ferdinand M. Vieider. 2019. "All over the map: A World-wide Comparison of Risk Preferences." *Quantitative Economics*, 10: 185–215.

Ma, Wei Ji, Konrad Paul Kording, and Daniel Goldreich. 2023. *Bayesian Models of Perception and Action: An Introduction.* MIT press.

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

Oprea, Ryan, and Ferdinand M. Vieider. 2024. "Minding the Gap: On the Origins of Probability Weighting and the Description-Experience Gap." Mimeo.

Prelec, Drazen. 1998. "The Probability Weighting Function." *Econometrica*, 66: 497–527.

R Core Team. 2023. "R: A Language and Environment for Statistical Computing." Vienna, Austria, R Foundation for Statistical Computing.

Tversky, Amos, and Daniel Kahneman. 1992. "Advances in Prospect Theory: Cumulative Representation of Uncertainty." *Journal of Risk and Uncertainty*, 5: 297–323.

Vieider, Ferdinand M. 2023*a*. "Cognitive Foundations of Delay-Discounting." *Working Paper.*

Vieider, Ferdinand M. 2023*b*. "Decisions under Uncertainty as Bayesian Inference on Choice Options." *Management Science, forthcoming.*

Vieider, Ferdinand M. 2024. "Bayesian Estimation of Decision Models." RISLab.