What Do We Really Know About Risk Preferences for Binary Lotteries?*

Camila Farres
Caltech

Ted O'Donoghue
Cornell

Charles D. Sprenger
Caltech

June 3, 2026

Abstract

Choices over binary lotteries have been used both to document expected utility (EU) anomalies and to motivate features of non-EU alternatives. However, the parameter space for such problems has not been comprehensively explored; experiments rely heavily on canonical examples, leaving an important gap in knowledge. We provide this exploration. For the narrow regions previously studied, our data reproduce the typical patterns and thus are consistent with leading non-EU models. However, we find new patterns in the broader parameter space that are inconsistent with those models. It turns out that "upside potential" (McGranaghan et al., 2025) can provide a credible rationalization.

JEL Codes: C91, D81, D91

Keywords: Risky Choice, Lotteries, Prospect Theory, Upside Potential

*Farres: Division of the Humanities and Social Sciences, California Institute of Technology, email: cfarresr@caltech.edu; O'Donoghue: Department of Economics, Cornell University, email: edo1@cornell.edu; Sprenger: Division of the Humanities and Social Sciences, California Institute of Technology, email: sprenger@caltech.edu. For helpful comments and suggestions, we thank seminar participants at the Caltech Experimental Coffee Hour, the 2025 ESA Meetings, and the University of Iowa. The experiment reported in this paper was preregistered in the AEA RCT Registry in April 2025, under the ID AEARCTR-0015734. The design was reviewed and granted an exemption by the Institutional Review Board at the California Institute of Technology under protocol number IR24-1475.

1 Introduction

Choices over binary lotteries are a proving ground for developing behavioral theories of risk preferences. In this basic setting, choices inconsistent with expected utility (EU) provide the foundation for models such as prospect theory (Kahneman and Tversky, 1979) and disappointment aversion (Bell, 1985). However, some of the key patterns incorporated into these models are based on experimental evidence covering a very narrow range of the possible parameter space. We attempt to fill this gap by exploring choices over binary lotteries with broad coverage of the possible parameter space. In doing so, we document new regularities, at odds with both EU and leading non-EU theories, that require explanation, and we consider one recent model that appears capable of rationalizing some of these new regularities.

We focus on binary lotteries in which there is some probability of winning a positive amount and otherwise one gets zero. In particular, we consider comparisons that take the following form:

$$\text{Safe Option: } (M, r) \qquad \text{vs.} \qquad \text{Risky Option: } (H, pr),$$

with $H > M$, $r \in (0, 1]$, and $p \in (0, 1)$.[1] In this parameterization, r is the chance of getting the moderate prize, M, in the safer lottery, while p is the relative chance of getting the high prize, H, in the riskier lottery. Individual risk attitudes in these comparisons can be captured by an indifference valuation $m_i(p, r)$ defined by

$$(m_i(p, r),\ r) \sim_i (H,\ pr)$$

and the corresponding *normalized risk premium*

$$RP_i(p, r) \equiv pH - m_i(p, r).$$

Positive values for $RP_i(p, r)$ reflect risk aversion, and negative values reflect risk tolerance.[2]

Within this setting, the literature has documented two key patterns that contradict EU and that have been used to motivate various non-EU models.

First, when researchers elicit certainty equivalents for binary lotteries, they reliably find a transition from risk aversion for large probabilities to risk tolerance for small probabilities—in our parameterization, fixing $r = 1$, individuals appear risk averse at high p and risk tolerant at low p. This finding was first documented by Preston and Baratta (1948) and made prominent by Tversky and Kahneman (1992), and thus we label it the *PB-TK effect*. The PB-TK effect is frequently used to motivate the assumption of an inverse-S-shaped probability weighting function in prospect theory or in models of rank-dependent expected utility.

The PB-TK effect has been confirmed in numerous papers.[3] Indeed, the most common way to

[1] Throughout, we suppress zero outcomes. For instance, the safe lottery (M, r) means one gets M with probability r and zero with the remaining probability $1 - r$.

[2] The normalized risk premium is the standard risk premium $rpH - rm_i(p, r)$ divided by r. We use this normalization to put $RP_i(p, r)$ on the same scale as our response variable $m_i(p, r)$. Moreover, with this normalization, $RP_i(p, r)$ should be invariant to r under EU.

[3] In addition to Preston and Baratta (1948) and Tversky and Kahneman (1992), prominent examples include

estimate a probability weighting function is to use data on certainty equivalents for binary lotteries. Interestingly, however, while the PB-TK effect is essentially a comparative static with respect to p while holding $r = 1$, as far as we know, there are no papers that study analogous comparative statics with respect to p while holding r constant at values smaller than 1. Such evidence is relevant because models of risk preferences make predictions for these comparative statics. Hence, our first goal is to investigate comparative statics with respect to p while holding r constant at a broad range of parameters.

Second, a stylized fact in the literature is that people exhibit a *common ratio effect (CRE)*: relative to their risk attitudes when comparing a binary lottery to a sure amount, scaling down the winning probabilities in each lottery by a common ratio increases risk tolerance. In our parameterization, this means $(M, 1) \sim_i (H, p)$ implies $(M, r) \prec_i (H, pr)$ for any $r < 1$, or equivalently that $RP_i(p, 1) > RP_i(p, r)$ for any $r < 1$. Allais (1953) first posited the CRE as a hypothetical example that contradicts EU, which implies that preferences should be invariant to such common ratio manipulations. The CRE was made prominent by Kahneman and Tversky (1979) when they used it as one of their main motivations for incorporating probability weighting into prospect theory.[4] Indeed, Kahneman and Tversky propose that probability weighting has an even stronger property that they label *subproportionality*, which says that $(M, r) \sim_i (H, pr)$ for any r implies $(M, r') \prec_i (H, pr')$ for any $r' < r$, or equivalently that $RP_i(p, r) > RP_i(p, r')$ for any $r' < r$.[5]

While a global CRE—and sometimes global subproportionality—is often assumed when people develop or analyze non-EU models, it is not clear that the evidence supports such strong assumptions. Blavatskyy et al. (2023) provides a meta-analysis of 143 CRE experiments, and whereas the CRE shows up more often than not, it does not show up universally. McGranaghan et al. (2025) reanalyze the experiments from Blavatskyy et al. and further document how those 143 experiments cover a rather limited portion of the parameter space. Moreover, to our knowledge, there are very few experiments that assess the stronger property of subproportionality.[6] On net, then, the evidence in support of the assumptions of a global CRE or global subproportionality is quite thin. Hence, our second goal is to investigate comparative statics with regard to r while holding p constant for a broad range of parameters, and to assess the extent to which we see subproportionality versus its opposite, which we label *superproportionality*.[7]

In Section 2, we present our experimental design. We conduct an online experiment for which we recruited 800 subjects from the online platform Prolific.co. For fixed values of H, p, and r, we

Tversky and Fox (1995), Gonzalez and Wu (1999), Bruhin et al. (2010), and Bernheim and Sprenger (2020).

[4] In terms of the label, Machina (1982) credits MacCrimmon and Larsson (1979), who discussed Allais' phenomenon as being delivered by a "common ratio" of probabilities.

[5] Interestingly, while the functional form for probability weighting assumed in Tversky and Kahneman (1992) generates a global CRE under its typically assumed parameter range, it does not generate global subproportionality. In particular, superproportionality can emerge for some r and $r' < r$ combinations where both are small. The functional form proposed in Prelec (1998) does generate global subproportionality (and thus a global CRE).

[6] Perhaps the main evidence comes from Kahneman and Tversky (1979), who found subproportionality when comparing $(p, r) = (0.5, 0.9)$ to $(p, r) = (0.5, 0.002)$.

[7] Formally, superproportionality means $(M, r) \sim_i (H, pr)$ for any r implies $(M, r') \succ_i (H, pr')$ for any $r' < r$, or equivalently that $RP_i(p, r) < RP_i(p, r')$ for any $r' < r$.

use multiple price lists (MPLs) to elicit subjects' indifference values $m_i(p,r)$ such that

$$(m_i(p,r),\ r) \sim_i (H,\ pr).$$

To systematically evaluate risk attitudes throughout the (p,r) parameter space, we consider all combinations of 9 values of $p \in \{0.1, 0.2, ..., 0.9\}$ and 21 values of $r \in \{0.01, 0.05, 0.10, ..., 0.95, 1\}$, yielding a total of 189 (p,r) combinations. Each subject provides 30 valuations that cover 25 different (p,r) combinations (with five valuations repeated), resulting in roughly 127 observations at each (p,r) combination. In addition, each subject's 25 (p,r) combinations are chosen such that we can conduct within-person comparative statics for how their $m_i(p,r)$ varies with p and r.

In Section 3, we present our aggregate-level results on how the mean normalized risk premium varies by p and r. Consider first comparative statics with respect to p. For the case where $r = 1$, we replicate the reliable finding that subjects display risk tolerance at low values of p and risk aversion at high values of p. Our novel (and perhaps surprising) finding is that this pattern turns out to be largely invariant to r—in other words, the PB-TK effect holds for all r. Moreover, the location of the transition from risk aversion to risk tolerance (around $p = 0.3$) is also consistent for all r. Indeed, for $p \geq 0.4$, 94% of (p,r) combinations reveal significant risk aversion; for $p \leq 0.2$, 93% indicate significant risk tolerance; and for $p = 0.3$, 95% are consistent with risk neutrality.

In contrast, when we investigate comparative statics with respect to r, only limited variability in risk attitudes is observed. Across 1,890 potential comparisons in the data, 80.9% of them are statistically consistent with EU's prediction of stable risk attitudes as r changes. Our design allows us to combine multiple values of r within a given p to provide more highly powered statistical tests of trend. At high values of p, what emerges is statistically significant but quantitatively modest subproportionality (increasing risk aversion as r increases). At low values of p, we find either null effects or even some evidence of slight superproportionality (increasing risk tolerance as r increases).

In Section 4, we consider the implications of our aggregate-level results for models of risk preferences. Our data are clearly at odds with EU, and moreover they are also at sharp odds with typical models of probability weighting. In particular, the inverse-S-shaped probability-weighting function needed to explain the PB-TK effect for $r = 1$ actually implies that a person should be risk tolerant for most (p,r) combinations (as we document in Figure 4), in contrast to our finding that the same PB-TK effect holds for all r. In addition, standard models of probability weighting predict subproportionality to be both substantial in magnitude and prevalent throughout the parameter space, rather than the modest and variable r-effects that we find. We also describe in Section 4 two post-hoc variants of prospect theory that we were able to identify that can explain a PB-TK effect that is consistent across r, although both further imply that risk attitudes should be completely independent of r.

In Section 4, we also consider the model of upside potential (UP) that was recently proposed by McGranaghan et al. (2025). Within this model, individuals trade off the probability of winning something against an expected valuation of those winnings. McGranaghan et al. develop the UP model to explain the substantial preference for probabilistic mixtures that they see in their data.

Interestingly, we show that this model is actually able to capture the qualitative patterns that we see in our aggregate-level analysis. Indeed, the UP model predicts that there should be a tight connection between risk attitudes and subproportionality. Specifically, it predicts that when a person is risk averse, they should exhibit subproportionality, and when a person is risk tolerant, they should exhibit superproportionality—and thus can accommodate our observed aggregate-level pattern of risk aversion and mild subproportionality at high p combined with risk tolerance and possibly some superproportionality at low p.

In Section 5 we assess the quantitative performance of the various models we consider. For each model, we pursue two quantitative exercises. First, we estimate the model based on only the nine mean risk premia when $r = 1$ (analogous to what is typically done when people estimate probability weighting functions), and then assess the out-of-sample fit of the estimated model for the other 180 mean risk premia when $r < 1$. Second, we estimate the model based on all 189 mean risk premia, and assess in-sample fit.[8] In the first exercise, the probability-weighting model fits the $r = 1$ data well, but that estimated model predicts poorly out of sample because it predicts much more general risk tolerance in the remainder of the space, rather than the consistent PB-TK effect observed. In the second exercise, not surprisingly, the probability weighting model has a hard time fitting all of the data. In contrast, in the first exercise, the UP model also fits the $r = 1$ data well and additionally predicts quite well out of sample. Moreover, in the second exercise, the UP model also does quite well in fitting all the data. Finally, we note that our two post-hoc variants of prospect theory actually perform a little better than the UP model in both exercises due to the fact that they best capture the consistent PB-TK effect.

In Section 6, we investigate a key individual-level prediction that distinguishes the UP model from our two post-hoc variants of prospect theory. The latter models make the stark prediction that subjects should exhibit constant risk attitudes with respect to r. The UP model instead predicts that subjects might exhibit both subproportionality and superproportionality (even within subject but across p), and moreover that there should be a tight connection between risk attitudes for a p and whether a subject is sub- versus superproportional for that p. When we investigate behavior at the individual-p level, we indeed see a lot of heterogeneity in subjects exhibiting sub- versus superproportionality (including some that is quantitatively large), and moreover we also indeed document a strong connection between which one a subject exhibits and their risk attitudes.

In Section 7, we conclude with a discussion of how our analysis fits together to inform the literature on risk preferences and to develop new ground on which to theorize. We also describe some limitations of our analysis, and also some broader messages to take away.

[8] Each model that we estimate requires us to estimate a nonlinear function; to limit our flexibility, we use the same functional form for all models. For details, see Section 5.

2 Experimental Design and Analysis

Our experiment focuses on risk attitudes for binary lottery comparisons. Specifically, for fixed values of H, p, and r, we elicit the value $m_i(p, r)$ such that

$$(m_i(p, r),\ r) \sim_i (H,\ pr). \tag{1}$$

Throughout our experiment, we fix the value of H to be \$30. To deliver coverage of the (p, r) parameter space, we select 189 combinations of (p, r) as the product of 9 values of p and 21 values of r:

$$p \in \{0.1, 0.2, 0.3, ..., 0.9\} \qquad \text{and} \qquad r \in \{0.01, 0.05, 0.10, 0.15, ..., 0.95, 1\}$$

Our design calls for approximately equal data collection across these 189 locations with the objective of accumulating sufficient observations for statistical comparison between any two locations. We conducted our study with 800 subjects on the online platform Prolific.co on April 10, 2025.[9] The 800 subjects each provided 30 observations of $m_i(p, r)$, yielding a total of 24,000 observations or approximately 127 observations at each of the 189 locations in the (p, r) parameter space.

Subjects were from the United States or Western Europe, fluent in English, had completed at least high school education, were substantially experienced with the Prolific platform, and had a high prior approval rating. We excluded subjects who participated in the pilot for this experiment and recruited an equal number of male and female participants.

2.1 Elicitation and Task Selection

The elicitation of $m_i(p, r)$ is conducted via the multiple price list (MPL) methodology, where subjects make a series of decisions between two options. Table 1 illustrates the structure of an MPL. For each MPL, Option A is the riskier lottery (H, pr); this option remains fixed across all rows with $H =$ \$30. Option B is the safer lottery (M, r); this option changes across rows, where M increases from \$0 in the first row to \$30 in the last row. Our protocol enforces a single switching row. Specifically, clicking Option A in any row highlights Option A in that row and all preceding rows. Analogously, clicking Option B in any row highlights Option B in that row and all subsequent rows. Participants can adjust their responses with additional clicks; when they are satisfied, they click submit to register their responses.

As our measure of a participant's $m_i(p, r)$, we use the midpoint of their switching rows. For instance, if a participant chooses Option A in row 2 and Option B in row 3, we record that their $m_i(p, r) =$ \$1. We intentionally chose the values of M to take the form $\$x.50$ so that interior switching points yield whole numbers, which in turn means that it is always possible for a participant to have $m_i(p, r) = pH$—that is, it is always possible to measure a participant to be risk-neutral. In addition, note that choosing Option B in the first row or choosing Option A in the last row would reflect a violation of dominance. In other words, observations that are censored at the boundaries of

[9] The experiment was programmed and administered using oTree (Chen et al., 2016).

Table 1: Elicitation of $m_i(p, r)$

Option A		**Option B**
pr chance of \$30 $1 - pr$ chance of \$0	OR	r chance of \$0 $1 - r$ chance of \$0
pr chance of \$30 $1 - pr$ chance of \$0	OR	r chance of \$0.50 $1 - r$ chance of \$0
pr chance of \$30 $1 - pr$ chance of \$0	OR	r chance of \$1.50 $1 - r$ chance of \$0
...	...	...
pr chance of \$30 $1 - pr$ chance of \$0	OR	r chance of \$29.50 $1 - r$ chance of \$0
pr chance of \$30 $1 - pr$ chance of \$0	OR	r chance of \$30 $1 - r$ chance of \$0

Notes: Structure of the multiple price list (MPL) used to elicit $m_i(p, r)$. In each row, participants choose between Option A and Option B. Interface enforces a single switching row, and our measure of $m_i(p, r)$ is the midpoint of the switching rows—for instance, if participant switches between rows 2 and 3, we record $m_i(p, r) = \$1$.

our price lists reflect violations of dominance; we investigate the extent of such behavior in Appendix C.1 as a measure of data quality.

Each subject provided $m_i(p, r)$ at 25 different locations in the (p, r) parameter space that are chosen to facilitate individual analysis. Specifically, for each subject, we select 5 values of p with coverage of low, middle, and high values. For each selected p, we select 5 values of r that also provide coverage of low, middle, and high values. Finally, we select those values of r such that, for at least 3 values of r, all 5 p values are covered. The exact algorithm for selecting each participant's (p, r) values is described in Appendix A.

To evaluate stability of responses and noisiness in behavior, we randomly select five of a participant's 25 (p, r) combinations and collect repeat elicitations of $m_i(p, r)$. Hence, each participant provides a total of 30 observations of $m_i(p, r)$. Repetitions are separated temporally from initial elicitations by at least 3 tasks. For the analysis, we randomly label either the first or second elicitation as the "repeat" to avoid any order effects.

2.2 Experimental Protocol

The experiment began with an instruction phase where participants completed practice tasks, followed by comprehension and attention checks. They then completed their 30 $m_i(p, r)$ elicitations, with breaks after tasks 4, 8, 16, 20, 24, and 28, during which participants completed a brief visual attention task (finding a camouflaged animal in an image). After completing all tasks, participants answered two incentivized quiz questions, which allow us to assess their understanding. Finally, we determined whether they were eligible for a bonus payment; if they were, the corresponding bonus

was revealed. Screenshots of the experiment are provided in Appendix G.

All participants who completed the study received a $5 participation fee, and one in five was randomly selected for a bonus payment. If chosen for the bonus, either an $m_i(p, r)$ elicitation or a quiz question was randomly selected. If an $m_i(p, r)$ elicitation task was chosen, a random row from that task was drawn, and the lottery selected in that row was played and paid accordingly. If a quiz question was selected, participants received $5 for a correct answer and $0 otherwise. The initial comprehension questions and attention checks did not qualify for the bonus payment, as participants had to answer them correctly to proceed in the experiment. On average, participants took 29 minutes to complete the study and earned $7.78.

2.3 Analytical Methods

For our analysis, we convert each elicited valuation $m_i(p, r)$ into a *normalized risk premium*:

$$RP_i(p, r) \equiv pH - m_i(p, r), \qquad \begin{cases} RP_i(p, r) = 0 \;\Rightarrow\; \text{risk neutral}, \\ RP_i(p, r) > 0 \;\Rightarrow\; \text{risk averse}, \\ RP_i(p, r) < 0 \;\Rightarrow\; \text{risk tolerant}. \end{cases} \tag{2}$$

Measurement of $RP_i(p, r)$ at 189 balanced locations in the parameter space permits separate investigation of p and r comparative statics that we evaluate both parametrically and nonparametrically.

For the parametric evaluation of p comparative statics, we consider the linear regression

$$RP_i(p, \bar{r}) = \alpha_p - \beta_p(1 - p) + \epsilon_p \tag{3}$$

for different fixed values of $r = \bar{r}$. With this specification, $\beta_p > 0$ corresponds to increasing risk aversion as p increases (i.e., the direction of the PB-TK effect). In our aggregate-level analysis in Section 3, we run this regression separately for each of the 21 values of $\bar{r}$, using all observations at that $\bar{r}$ and assuming that α_p and β_p are the same for everyone. In our individual-level analysis in Section 6, we run this regression for each individual-$\bar{r}$ combination that has at least five observations, permitting individual-specific estimates for $\alpha_{p,i}$ and $\beta_{p,i}$.

Analogously, for the parametric evaluation of r comparative statics, we consider the linear regression

$$RP_i(\bar{p}, r) = \alpha_r - \beta_r(1 - r) + \epsilon_r \tag{4}$$

for different fixed values of $p = \bar{p}$. With this specification, $\beta_r > 0$ corresponds to subproportionality (i.e., the direction of the CRE). In our aggregate-level analysis in Section 3, we run this regression separately for each of the nine values of $\bar{p}$, using all observations at that $\bar{p}$ and assuming that α_r and β_r are the same for everyone. In our individual-level analysis in Section 6, we run this regression separately for each individual-$\bar{p}$ combination (where by construction each has at least five observations), permitting individual-specific estimates for $\alpha_{r,i}$ and $\beta_{r,i}$.

For non-parametric evaluations, we use Kendall's τ^a as a rank-based measure of monotone trend

(Kendall, 1938). Kendall's τ^a compares the prevalence of increasing versus decreasing outcomes across pairs ordered by an independent variable. To account for heterogeneity, we build from Kendall's τ^a at the individual level.

To illustrate, consider how we analyze the trend with respect to p for a fixed value of $r = \bar{r}$. We first calculate $\tau^a_{p,i}$ for each individual i who saw at least two distinct values of p for that $\bar{r}$. Specifically, we order individual i's observations by p, and for each ordered pair $p < p'$, we record whether the normalized risk premium increases, $RP_i(p', \bar{r}) > RP_i(p, \bar{r})$, decreases, $RP_i(p', \bar{r}) < RP_i(p, \bar{r})$, or is constant. Finally, letting $I_{p,i}$ denote the number of increasing pairs, $D_{p,i}$ the number of decreasing pairs, and $N_{p,i}$ the total number of pairs (including pairs with $p = p'$), we calculate:

$$\tau^a_{p,i} = (I_{p,i} - D_{p,i})/N_{p,i}. \tag{5}$$

This statistic lies in $[-1, 1]$, where $\tau^a_{p,i} = 1$ corresponds to a perfectly increasing relationship (all ordered pairs are increasing), $\tau^a_{p,i} = -1$ corresponds to a perfectly decreasing relationship, and values near zero indicate little systematic trend.

In our aggregate-level analysis in Section 3, for each of the 21 values of $\bar{r}$, we report τ^a_p as a weighted mean of the $\tau^a_{p,i}$ across all individuals who saw at least two distinct values of p for that $\bar{r}$, where the weight on individual i is proportional to $N_{p,i}$. For our individual-level analysis, we report the distribution of $\tau^a_{p,i}$ in Appendix F.

Our non-parametric analysis of the trend with respect to r is analogous. For each $\bar{p}$, we first calculate $\tau^a_{r,i}$ for each individual i who saw at least two distinct values of r for that $\bar{p}$. Then in Section 3 we report τ^a_r as a weighted mean of the $\tau^a_{r,i}$ across all individuals who provided responses for that $\bar{p}$, and we report the distribution of $\tau^a_{r,i}$ in Appendix F.

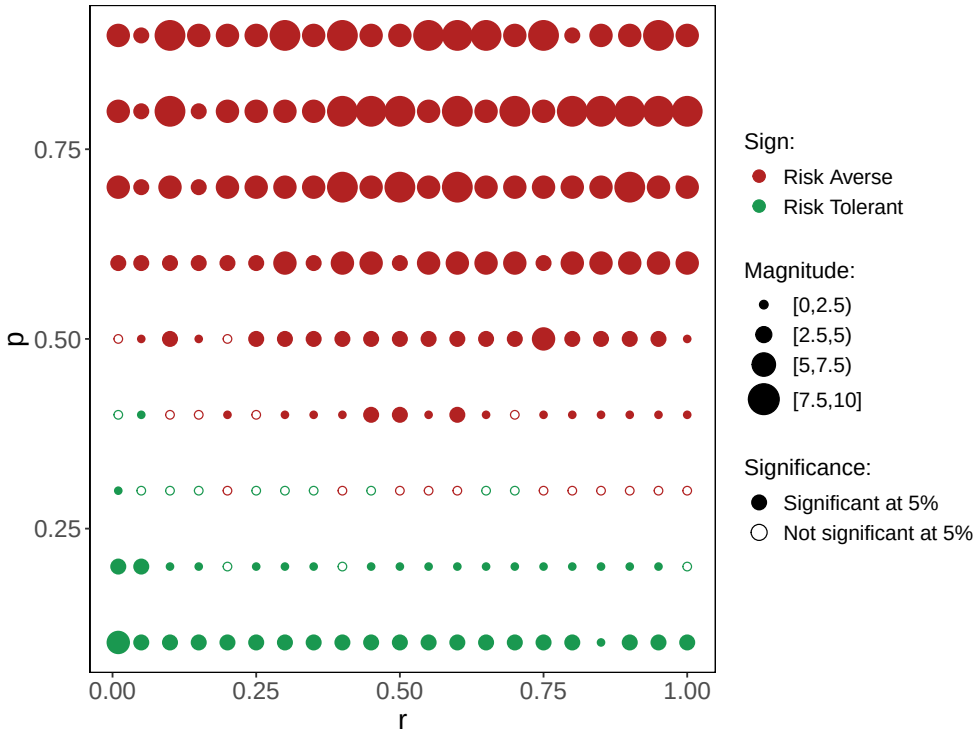
3 Aggregate-Level Experimental Results

In this section, we analyze our data at the aggregate level. Recall that our 800 subjects each completed 30 valuation tasks, giving a total of 24,000 observations of $m_i(p, r)$ and $RP_i(p, r)$ across 189 different (p, r) combinations, and thus an average of 127 observations at each (p, r) combination. We begin by analyzing patterns in the mean normalized risk premium across the full (p, r) grid, and we then conduct aggregate-level versions of the parametric and nonparametric comparative static analyses described in Section 2.3.

For each (p, r), we define the mean valuation $m(p, r)$ to be the mean of $m_i(p, r)$ across all observations for that (p, r), including repeats. We analogously define the mean normalized risk premium $RP(p, r)$ to be the mean of $RP_i(p, r)$, and note that mechanically $RP(p, r) = pH - m(p, r)$. In Appendix B, we present all 189 values for $m(p, r)$ and for $RP(p, r)$ along with their standard errors in Appendix Tables B.1 and B.2.[10]

[10] In our pre-analysis plan, we outlined five checks for evaluating data quality. Because our data met all five checks (see details in Appendix C.1), we proceed by using the entire data set. As also specified in our pre-analysis plan, in Appendices C.2 and C.3 we assess the robustness of our main aggregate-level results in two pre-specified high-quality subsamples; the results are largely unchanged.

Figure 1: Mean Normalized Risk Premium $RP(p, r)$ by (p, r)



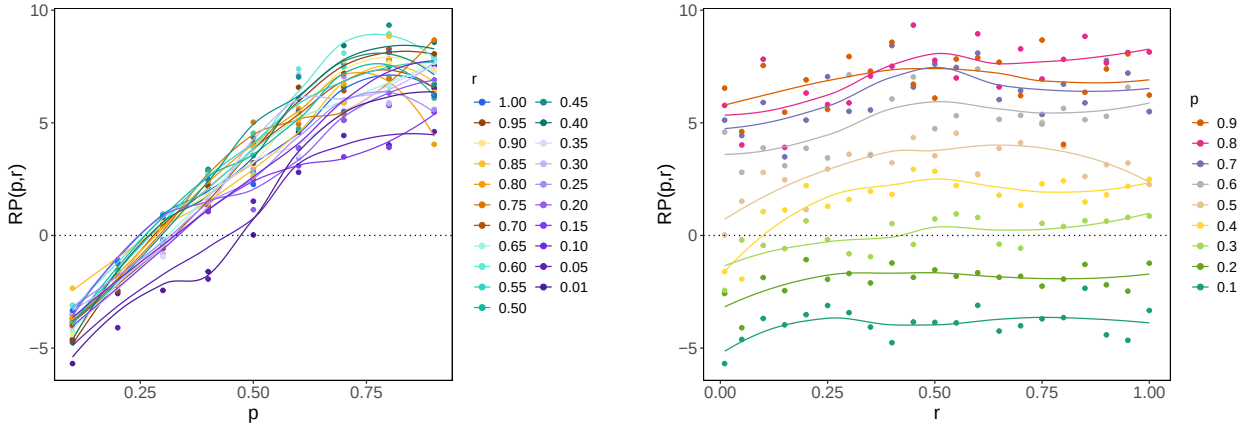
Notes: Figure presents mean normalized risk premium $RP(p, r)$ for each of the 189 (p, r) combinations. For each (p, r), $RP(p, r)$ is the mean of $RP_i(p, r)$ across all observations for that (p, r), including repeats. For each (p, r), color of circle captures sign of $RP(p, r)$, with red indicating risk aversion and green indicating risk tolerance, while size of circle captures the magnitude of $RP(p, r)$. Filled points correspond to (p, r) for which $H_0 : RP(p, r) = 0$ is rejected at the 5% level. See Appendix Table B.2 for exact quantities and standard errors for each circle. Based on 24,000 observations from all 800 participants, or roughly 127 observations for each circle.

Figure 1 presents a visualization of how the mean normalized risk premium $RP(p, r)$ varies by p and r, where each circle reflects the sign and magnitude of $RP(p, r)$ for that (p, r). Red circles represent positive (risk-averse) values, while green circles represent negative (risk-tolerant) values. The size of the circle captures the magnitude. Filled circles reflect (p, r) combinations where we reject the null hypothesis $H_0 : RP(p, r) = 0$ at the 5% level in a two-sided test.[11]

Figure 1 reveals two striking patterns. First, across all values of r, there is a clear transition from risk aversion at high p to risk tolerance at low p (i.e., as we move down each column, the circles clearly transition from red to green). For $r = 1$, this result replicates the standard PB-TK effect that reliably appears in prior work; the novel result here is that this finding seems to hold for all values of r. Moreover, the location of the shift from risk tolerance to risk aversion appears consistently around $p = 0.3$. For $p \geq 0.4$, 94% (118 of 126) of $RP(p, r)$ values indicate statistically significant

[11] An alternative way to present the data is to report, for each (p, r) combination, the relative proportion of observations with $RP(p, r) > 0$ versus observations with $RP(p, r) < 0$. We do this in Appendix Figure B.1, and the message is virtually identical to what emerges from Figure 1.

Figure 2: $RP(p, r)$ Variation Across p and r



(a) $RP(p, r)$ vs. p (by r)

(b) $RP(p, r)$ vs. r (by p)

Notes: Panel (a) presents mean normalized risk premium $RP(p, r)$ against p, showing a separate line for each value of r. Panel (b) plots $RP(p, r)$ against r, showing a separate line for each value of p. Lines in both panels correspond to LOESS fits. See Appendix Table B.2 for exact quantities and standard errors at each point. Based on 24,000 observations from all 800 participants, or roughly 127 observations for each (p, r).

risk aversion; for $p \leq 0.2$, 93% (39 of 42) of $RP(p, r)$ values indicate statistically significant risk tolerance; and for $p = 0.3$, 95% (20 of 21) of $RP(p, r)$ values are consistent with risk neutrality.

Second, patterns of sub- or superproportionality are not visually salient in the mean data. Recall that subproportionality means that, holding p fixed, the normalized risk premium $RP(p, r)$ increases with r. Superproportionality means that it decreases with r. Figure 1 shows relatively little systematic change as we move across rows.

We next explore both comparative statics a bit more rigorously. Consider first comparative statics with respect to p. Panel (a) of Figure 2 provides a better visualization of this comparative static by plotting the mean $RP(p, r)$ against p separately for the 21 different values of r. Again, the comparative statics with respect to p are notably consistent across values of r. $RP(p, r)$ indicates risk tolerance for low p, then generally increases with p, crossing from risk tolerance to risk aversion at around $p = 0.3$, and reaching a maximum at around $p = 0.8$ before flattening. In terms of magnitude, the normalized risk premium increases from roughly -4 when $p = 0.1$ to roughly 7 when $p = 0.8$.[12]

The left-hand panel of Table 2 assesses the statistical significance of this comparative static by using the β_p and τ_p^a measures defined in Section 2.3. First, we estimate the regression specification in equation (3) separately for each value of r, and report the 21 estimated coefficients for β_p. For all 21 values of r, the estimated β_p is positive and statistically significant, and moreover the magnitudes are reliably quite large. Second, we calculate the non-parametric measure of trend τ_p^a. Again, the calculated τ_p^a is positive and statistically significant for all 21 values of r. Risk aversion at high p transitioning to risk tolerance at low p seems to be a remarkably robust and meaningfully large phenomenon across all r.

[12]Note that this corresponds to $m(p, r)$ increasing from roughly 7 when $p = 0.1$ to roughly 17 when $p = 0.8$.

Table 2: Parametric and Non-Parametric p and r Comparative Statics

	p **Comparative Statics When** $\bar{r} = \bar{s}$				r **Comparative Statics When** $\bar{p} = \bar{s}$			
$\bar{s}$	β_p (SE)	N	τ_p^a (SE)	m	β_r (SE)	N	τ_r^a (SE)	m
0.01	15.939 (1.236)	1127	0.408 (0.042)	190				
0.05	12.331 (1.079)	1099	0.299 (0.042)	176				
0.10	14.840 (1.019)	1115	0.404 (0.040)	179	0.599 (0.494)	2537	-0.060 (0.019)	422
0.15	11.073 (0.939)	1182	0.360 (0.031)	190				
0.20	12.396 (0.923)	1106	0.301 (0.041)	179	0.613 (0.467)	2658	-0.031 (0.019)	445
0.25	12.450 (0.883)	1242	0.331 (0.036)	207				
0.30	14.371 (0.912)	1142	0.476 (0.035)	190	1.685 (0.444)	2775	0.026 (0.019)	463
0.35	14.612 (0.887)	1094	0.456 (0.037)	168				
0.40	16.442 (0.848)	1114	0.508 (0.038)	178	2.482 (0.478)	2564	0.065 (0.020)	428
0.45	15.603 (0.843)	1142	0.501 (0.032)	181				
0.50	13.781 (0.807)	1199	0.419 (0.036)	181	1.874 (0.471)	2774	0.070 (0.019)	467
0.55	14.654 (0.824)	1155	0.498 (0.034)	190				
0.60	15.817 (0.834)	1096	0.522 (0.031)	167	2.243 (0.503)	2683	0.063 (0.019)	444
0.65	14.787 (0.816)	1216	0.452 (0.033)	196				
0.70	14.958 (0.806)	1218	0.504 (0.031)	187	1.819 (0.535)	2628	0.044 (0.021)	436
0.75	14.880 (0.819)	1138	0.477 (0.036)	173				
0.80	12.971 (0.832)	1116	0.457 (0.033)	182	3.115 (0.568)	2606	0.132 (0.023)	433
0.85	13.075 (0.748)	1232	0.448 (0.032)	189				
0.90	15.609 (0.862)	1187	0.477 (0.032)	184	0.700 (0.561)	2775	0.069 (0.020)	462
0.95	16.707 (0.919)	1076	0.510 (0.032)	165				
1.00	12.641 (0.931)	1004	0.366 (0.043)	145				
Total		24000				24000		

Notes: Entries reflect separate comparative-static measures. In each row, column (2) reports estimated coefficient β_p from regression specification in equation 3 using all observations with $\bar{r} = \bar{s}$ (number of observations N reported in column (3)); standard error reported in parentheses is clustered by participant using CR2. In each row, column (4) reports weighted mean τ_p^a of individual $\tau_{p,i}^a$ calculated from equation 5 using all individuals who saw at least two distinct values of p for $\bar{r} = \bar{s}$ (number of individuals m reported in column (5)); weight for individual i is $w_i = N_{p,i} / \sum_j N_{p,j}$; standard error in parentheses is corresponding weighted empirical standard error $\hat{\sigma}_w \cdot \sqrt{\sum_i w_i^2}$, where $\hat{\sigma}_w^2 = (1/(1 - \sum_j w_j^2)) \sum_i w_i (\hat{\tau}_{p,i}^a - \tau_p^a)^2$. Columns (6) and (7) are analogous to columns (2) and (3), reporting estimated coefficient β_r from regression specification in equation 4 using all observations with $\bar{p} = \bar{s}$. Columns (8) and (9) are analogous to columns (4) and (5), reporting calculated τ_r^a using all individuals who saw at least two distinct values of r for $\bar{p} = \bar{s}$.

Next consider comparative statics with respect to r. Panel (b) of Figure 2 provides a better visualization of this comparative static by plotting the mean $RP(p,r)$ against r separately for different values of p. The curves are all relatively flat, although this visualization suggests perhaps some subproportionality.

The right-hand panel of Table 2 assesses the statistical significance of this comparative static by using the β_r and τ_r^a measures defined in Section 2.3. We estimate the regression specification in equation (4) separately for each value of p, and report the 9 estimated coefficients for β_r. Recall that $\beta_r > 0$ indicates subproportionality, and in fact the estimated β_r is positive for all p and statistically significant for $p \in \{0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$. However, the magnitudes are relatively modest—e.g., the largest estimated $\beta_r = 3.12$ when $p = 0.8$ implies that $RP(0.8,r)$ (and $m(0.8,r)$) increases by only \$2.49 as we move from $r = 0.1$ to $r = 0.9$. When we calculate τ_r^a, it is positive and statistically significant for all $p \geq 0.3$, roughly zero for $p = 0.2$, and negative and statistically significant for $p = 0.1$. Again, though, the magnitudes are all quite small, especially when compared to the magnitudes for τ_p^a.

As an alternative way to assess the extent of sub- versus superproportionality in our data, we take advantage of the fact that our design yields 1,890 potential subproportionality tests that compare $RP(p,r)$ across two values of r for a fixed p.[13] Of these, 15.7% indicate significant subproportionality, 3.4% indicate significant superproportionality, and 80.9% cannot be statistically differentiated from the EU null.[14]

On net, then, the aggregate data suggest relatively small comparative-static effects with respect to r, with small amounts of subproportionality at larger p, and with neutrality or perhaps small amounts of superproportionality at very small p. We summarize these aggregate findings in the following two results:

Result 1 (Risk aversion at high p, risk tolerance at low p for all r)**.** *For all r, mean normalized risk premium increases with p and crosses from risk tolerance at low p to risk aversion at high p, with a stable crossover around $p \approx 0.3$.*

Result 2 (Modest subproportionality for high p, neutral or mild superproportionality for low p)**.** *Mean behavior is modestly r-sensitive, and the direction varies by probability region: for $p \geq 0.3$, mean normalized risk premium tends to increase with r (subproportionality), while for $p < 0.3$ it is close to r-insensitive or mildly decreasing in r (superproportionality).*

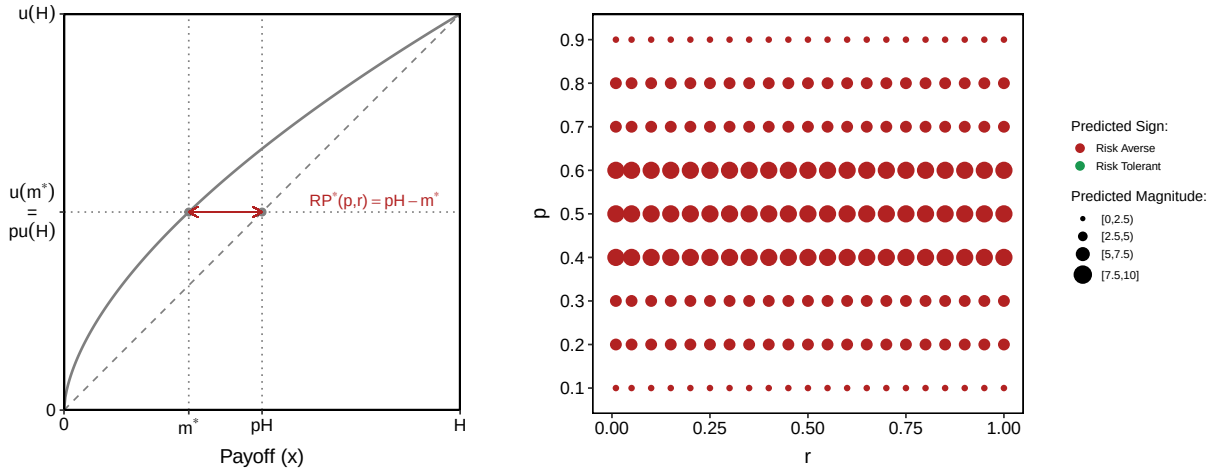
4 Implications for Models of Risk Preferences

In this section, we discuss the implications of Results 1 and 2 for various models of risk preferences. Note that, whereas above we use $m(p,r)$ and $RP(p,r)$ to denote empirical observations, here we use $m^*(p,r)$ and $RP^*(p,r)$ to denote theoretical predictions.

[13] Because for each p there are 21 values of r, there are $\binom{21}{2} = 210$ unique (r, r') pairs for each p. Given nine values for p, there are 1,890 unique (p, r, r') pairs.

[14] For the subproportionality test using its classic parameterization—that is, fixing $p = 0.8$ and comparing $r = 1$ to $r = 0.25$—we do find statistically significant subproportionality, with magnitude $RP(0.8,1) - RP(0.8,0.25) = 2.33$.

Figure 3: Predictions of Expected Utility (EU)



(a) Utility Function $u(x)$ (b) Predicted Risk Premia Over (p, r)

Notes: Figure depicts risk-premium map under EU assuming functional form $u(x) = x^{0.6}$. Panel (a) depicts the utility function, and it also depicts the normalized risk premium for one particular p highlighted in red. Given a (p, r) combination, m^* is defined by $ru(m^*) = pru(H)$ or $u(m^*) = pu(H)$, and then $RP^* = pH - m^*$. Note that normalized risk premium is independent of r under EU. In panel (b), for each (p, r), circle characterizes predicted $RP^*(p, r)$, where color of circle captures sign (green is positive, red is negative), and size of circle capture magnitude.

We organize this section around three main models: expected utility (EU), probability weighting (PW), and upside potential (UP). EU is the canonical benchmark in economics; PW is the component of prospect theory (PT) that was developed to accommodate the classic patterns in certainty equivalents and common ratio problems; and UP is a model recently proposed by McGranaghan et al. (2025) based on a different set of empirical results that contradict both EU and PW.

4.1 Expected Utility (EU)

It is useful to benchmark what we would expect to see in our data if, as commonly assumed in economics, people were risk-averse expected-utility maximizers—that is, if they behave according to EU with a globally concave utility function.

First, we would obviously expect to observe $RP^*(p, r) > 0$ for all p and r. Second, the magnitude of the normalized risk premium would be upside-down U-shaped in p—that is, for any fixed r, the normalized risk premium should be largest for intermediate p. Finally, the normalized risk premium should be independent of r. This last feature holds under EU with any utility function and is well known—it is why evidence on the CRE in particular and subproportionality more generally are seen as clear violations of EU. These predictions are depicted in Figure 3 and are clearly inconsistent with our data.

4.2 Probability Weighting (PW) and Prospect Theory (PT)

Evidence on (i) the transition from risk aversion at large p to risk tolerance at small p when $r = 1$ and (ii) the CRE has often been used to motivate PW models. Hence, it is natural to look at the

predictions of typical PW models throughout the parameter space covered by our data.

To isolate the implications of PW, we initially assume linear utility. Suppose a person evaluates lottery (x, q) as $\pi(q)x$, where π is a monotonic probability weighting function with $\pi(0) = 0$ and $\pi(1) = 1$.[15] Hence, the indifference valuation $m^*(p, r)$ is given by

$$\pi(r)m^*(p,r) = \pi(pr)H \qquad \Longleftrightarrow \qquad m^*(p,r) = \frac{\pi(pr)}{\pi(r)}H,$$

and the normalized risk premium is given by

$$RP^*(p,r) = pH - m^*(p,r) = pH - \frac{\pi(pr)}{\pi(r)}H = \frac{H}{\pi(r)}\left(p\pi(r) - \pi(pr)\right).$$

Figure 4 depicts the predictions of PW for our data when using the probability weighting function suggested by Tversky and Kahneman (1992). Panel A depicts the probability weighting function, and panel B presents the predictions for our data. For $r = 1$ (the far right column in panel B), we see the usual prediction of risk aversion for large p and risk tolerance for small p; indeed, the probability weighting function in panel A is typically motivated (and estimated) with such $r = 1$ certainty equivalent data. Panel B then illustrates the broader predictions of this probability weighting function through the rest of the parameter space. Starting from $r = 1$, we see clear subproportionality (as we move from right to left across a row, circles transition from larger red to smaller red to smaller green to larger green), except on the margin as r gets very small.[16] But the most striking feature of panel B is the prediction that risk tolerance should predominate through much of the parameter space except for a small region in the "northeast" portion.

The predictions depicted in panel B of Figure 4 do not accord well with our experimental findings. The model predicts much stronger subproportionality than we see in our data. But the most striking feature of our data—the robust transition for all r from risk aversion at high p to risk tolerance at low p—seems wholly inconsistent with standard formulations of PW.

The predictions in panel B of Figure 4 do not rely on the Tversky-Kahneman functional form, but rather are driven by the inverse-S shape that is needed to explain—within the probability-weighting framework—the robust empirical pattern of risk aversion at high p and risk tolerance at low p when $r = 1$. To see this, note that the sign of the normalized risk premium satisfies:

$$RP^*(p,r) > 0 \qquad \Longleftrightarrow \qquad \pi(pr) < p\pi(r)$$

and

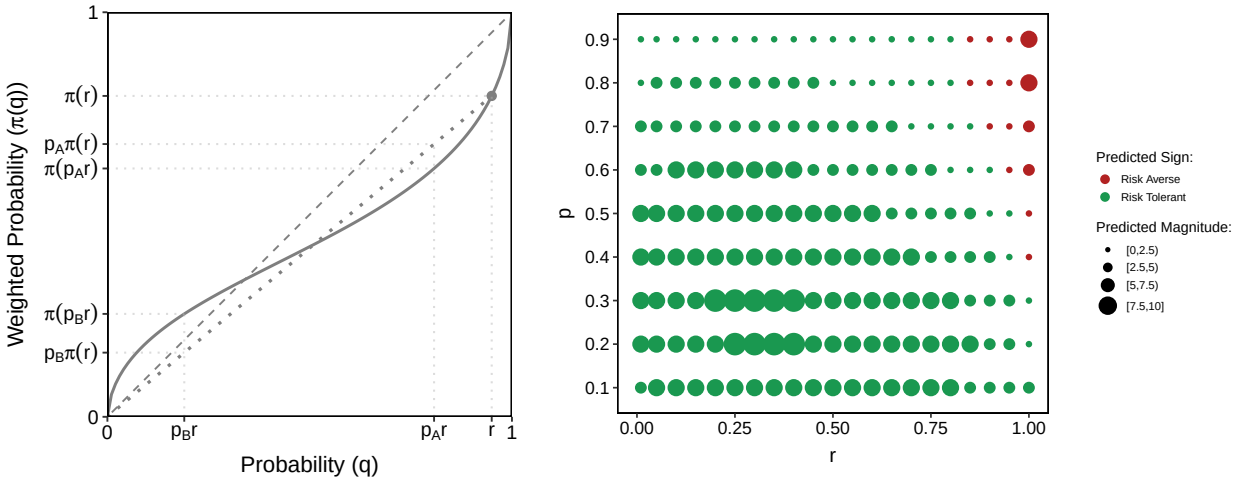
$$RP^*(p,r) < 0 \qquad \Longleftrightarrow \qquad \pi(pr) > p\pi(r).$$

When $r = 1$, these implications reduce to risk aversion when probabilities are underweighted,

[15] Given our data cover only binary lotteries with one zero outcome, the distinction between rank-dependent versus rank-independent probability weighting is irrelevant.

[16] Note that the Tversky and Kahneman (1992) functional form does not generate the global subproportionality suggested by Kahneman and Tversky (1979); however, one can show that it does generate a global CRE in the sense that, under its typically assumed parameter range of $\gamma \in (0.279, 1)$, $RP(p, 1) > RP(p, r)$ for all p and $r < 1$.

Figure 4: Predictions of Probability Weighting (PW)



(a) Probability Weighting Function $\pi(q)$ (b) Predicted Risk Premia Over (p, r)

Notes: Figure depicts risk-premium map under PW assuming functional form and parameter value suggested by Tversky and Kahneman (1992)—specifically, assuming probability weighting function $\pi(q) = q^\gamma/[q^\gamma + (1-q)^\gamma]^{1/\gamma}$ with $\gamma = 0.61$. Under PW, $RP^*(p, r) = \frac{H}{\pi(r)}(p\pi(r) - \pi(pr))$, and figure assumes $H = 30$. Panel (a) plots $\pi(q)$. In panel (b), for each (p, r), circle characterizes predicted $RP^*(p, r)$, where color of circle captures sign (green is positive, red is negative), and size of circle capture magnitude. In panel (a), dashed line is the line from the origin to $(1, \pi(1))$, which is the 45-degree line; when $r = 1$, model predicts risk aversion when $\pi(p)$ is below this line, and risk tolerance when $\pi(p)$ is above this line. Dotted line is the line from the origin to $(r, \pi(r))$; for that r, model predicts risk aversion for p where $\pi(pr)$ is below this line (e.g., p_A), and risk tolerance for p where $\pi(pr)$ is above this line (e.g., p_B).

$\pi(p) < p$, and risk tolerance when probabilities are overweighted, $\pi(p) > p$. Graphically, these conditions are represented by $\pi(p)$ being below (for risk aversion) or above (for risk tolerance) the ray from the origin to $(1, \pi(1))$, i.e., the 45 degree line. When $r < 1$, the intuition is identical but rather than comparing $\pi(pr)$ to the ray from the origin to $(1, \pi(1))$, one compares it to the ray from the origin to $(r, \pi(r))$. One such ray is illustrated in panel A along with two candidate values of p. For a larger candidate p_A, $\pi(p_A r)$ is below the ray, and thus the person is risk averse; for a smaller candidate p_B, $\pi(p_B r)$ is above the ray, and thus the person is risk tolerant. The value of r selected for illustration delivers regions of both risk aversion (for high p) and risk tolerance (for low p). Note, however that, relative to when $r = 1$, the set of p associated with risk aversion ($\pi(pr)$ below the ray) shrinks while the set of p associated with risk tolerance ($\pi(pr)$ above the ray) grows. Indeed, for any inverse S-shaped weighting function, for r sufficiently low, all values of p are associated with risk tolerance.[17]

PW is just one component of prospect theory (PT)(Kahneman and Tversky, 1979), and one might wonder whether the addition of a concave value function might generate predictions closer to what we see in our data—that is, if we assume that a person evaluates lottery (x, q) as $\pi(q)v(x)$

[17]For the functional form depicted in panel A, risk tolerance obtains for all p as soon as r drops below around 0.76. Alternative probability weighting specifications with similar shapes—such as those proposed by Lattimore et al. (1992) or Prelec (1998)—deliver similar predictions.

for some concave function v. In Appendix Figure D.1, we present predictions when using both the probability weighting in Figure 4 and the value function $v(x) = x^{0.88}$ that was also suggested by Tversky and Kahneman (1992). The essential qualitative predictions are maintained, although risk aversion occupies a larger portion of the "northeast" part of the parameter space.

Although our data, and especially Result 1, are inconsistent with the predictions of PT, we identified two post-hoc variants of PT that could explain Result 1. We present these for completeness and, in fact, we demonstrate in Section 5 that they do turn out to explain Result 1 quite well; however, we document in Section 6 that they are unlikely to explain more nuanced features of our data.

4.2.1 Proportional Probability Weighting (PPW)

When considering choices between binary lotteries, there is a natural way to modify the model of probability weighting to generate a transition from risk aversion for high p to risk tolerance at low p that is the same for all r: assume that people apply probability weighting to the *relative* probability between the two lotteries. In other words, when comparing lottery (H, q) to lottery (M, q') with $H > M$ and $q < q'$, one first calculates the relative probability q/q' and then applies probability weighting to the ratio. If so, then their valuation would be

$$m^*(p, r) = \pi\left(\frac{pr}{r}\right) H = \pi(p)H.$$

On the plus side, this post-hoc variant would by construction imply that a probability weighting function that generates a transition from risk aversion at high p to risk tolerance at low p for $r = 1$ would generate that same transition for all r, much as we see in Result 1. On the minus side, this model would imply that valuations and thus the normalized risk premium should be independent of r, and thus would imply that we should observe no sub- or superproportionality. Appendix Figure D.2 illustrates these implications by mapping the model-implied risk premium grid used in the experiment.

Proportional probability weighting is not often discussed in the literature, and perhaps one reason why is that it is not clear how to apply it outside the domain of choices between binary lotteries as it is no longer clear how to define proportional probabilities.

4.2.2 S-Shaped Value Function (SSV)

An alternative post-hoc way to explain Result 1 within prospect theory is to do so within the value function. In particular, with linear probability weighting, we can explain the transition from risk aversion to risk tolerance when $r = 1$ with an S-shaped value function that has $v(pH) < pv(H)$ for p small (which yields risk tolerance) and $v(pH) > pv(H)$ for p large (which yields risk aversion). Moreover, with linear probability weighting, the normalized risk premium is independent of r, and thus the same pattern would hold for all r. Appendix Figure D.3 illustrates these predictions. Of course, this model also cannot capture the classic CRE or our more nuanced findings on

subproportionality.

4.3 Upside Potential (UP)

McGranaghan et al. (2025) introduce a model of "upside potential" (UP) to explain the substantial preference for probabilistic mixtures that they see in their data on connected common ratio and common consequence problems. It turns out that their UP model, designed to rationalize other data, can naturally explain the patterns that we observe in our Results 1 and 2. Appendix D.2 provides a detailed treatment of this model and its predictions for our data. Here we provide an overview.

In the simplest form of the UP model, a person cares about the expected value of a lottery and also its upside potential, where the latter is the product of the total probability of winning something and an expected valuation of those winnings. Applied to the binary lotteries that are the focus of this paper, the UP model says that a person evaluates a lottery (x, q) as $qx + (q)(q\kappa(x))$, where κ is an upside function assumed to be monotonically increasing. The valuation $m^*(p, r)$ is then the M that satisfies the equation

$$rM + r^2\kappa(M) = prH + (pr)^2\kappa(H) \qquad \text{or}$$

$$[M - pH] + r[\kappa(M) - p^2\kappa(H)] = 0. \tag{6}$$

In Appendix D.2, we derive the following implication:

$$\kappa(pH) > p^2\kappa(H) \iff RP^*(p, r) > 0 \iff \frac{\partial RP^*(p, r)}{\partial r} > 0$$

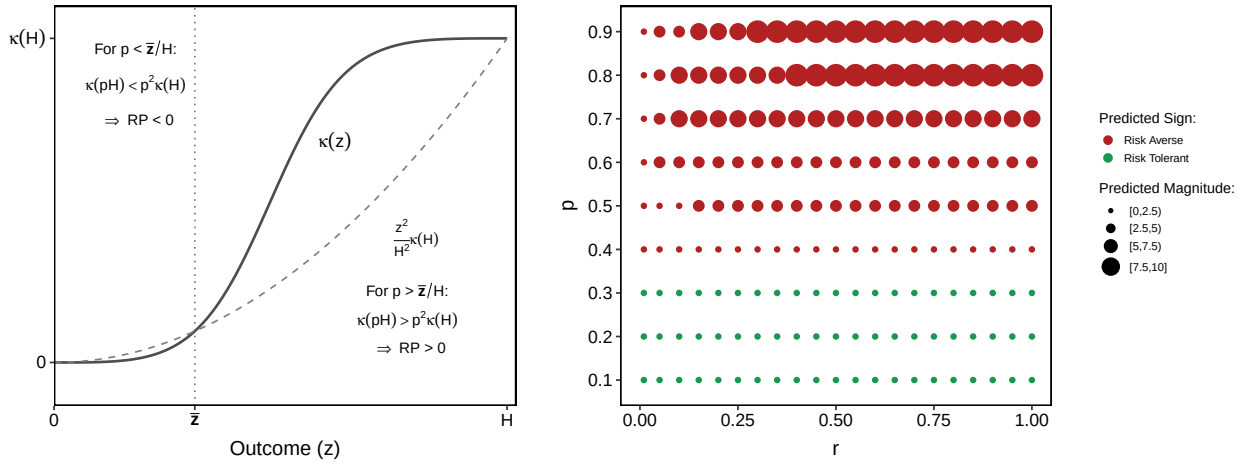
and

$$\kappa(pH) < p^2\kappa(H) \iff RP^*(p, r) < 0 \iff \frac{\partial RP^*(p, r)}{\partial r} < 0.$$

In other words, for any (p, r, H), a simple comparison of $\kappa(pH)$ to $p^2\kappa(H)$ determines both risk aversion versus risk tolerance *and* sub- versus superproportionality. Moreover, the model predicts a tight relationship between these two features: risk aversion goes hand in hand with subproportionality, while risk tolerance goes hand in hand with superproportionality.

To see the intuition for this prediction, note that there are two terms in equation 6. The first reflects a comparison of the expected values of the safer versus the riskier option, and the second reflects a comparison of the upside potential of the safer versus the riskier option. Any deviations from risk neutrality must come from the UP term. If at the risk-neutral $M = pH$ the UP term favors the safer option (that is, if $\kappa(pH) > p^2\kappa(H)$), then the UP term will be generating risk aversion, which leads to $RP^*(p, r) > 0$. Moreover, for that case, if r becomes larger, the UP term in equation 6 becomes more important, which leads to more risk aversion and thus a larger $RP^*(p, r)$ (i.e., subproportionality). An analogous intuition applies if at the risk-neutral $M = pH$ the UP term favors the riskier option (that is, if $\kappa(pH) < p^2\kappa(H)$); in this case the UP term generates risk tolerance and thus $RP^*(p, r) < 0$, and a larger r increases the importance of the UP term and thus

Figure 5: Predictions of Upside Potential (UP)



(a) Upside-Potential Function $\kappa(z)$ (b) Predicted Risk Premia Over (p, r).

Notes: Figure depicts risk-premium map under UP assuming functional form and parameter values from full-sample estimation reported in Table 3 in Section 5—specifically, assuming $\kappa(x) = \gamma(x/H)^\delta/[\gamma(x/H)^\delta + (1-x/H)^\delta]$ with $\gamma = 1.310$ and $\delta = 3.145$. Under UP, $m^*(p,r)$ is M that solves equation 6, then $RP^*(p,r) = pH - m^*(p,r)$, and figure assumes $H = 30$. Panel (a) plots $\kappa(x)$, and in addition the dashed curve provides the quadratic benchmark used for sign comparisons. In panel (b), for each (p,r), circle characterizes predicted $RP^*(p,r)$, where color of circle captures sign (green is positive, red is negative), and size of circle capture magnitude.

leads to increased risk tolerance (i.e., superproportionality).

Whether $\kappa(pH)$ is larger or smaller than $p^2\kappa(H)$ depends on the shape of the κ function. Using their data, McGranaghan et al. (2025) estimate an S-shaped κ function. Interestingly, this shape can naturally generate $\kappa(pH) > p^2\kappa(H)$ for larger p while $\kappa(pH) < p^2\kappa(H)$ for smaller p. If so, then we would see risk aversion and subproportionality for larger p, and risk tolerance and superproportionality for smaller p.

Figure 5 illustrates this possible prediction of the model. In panel (a) we illustrate an S-shaped κ function for which $\kappa(pH) = p^2\kappa(H)$ exactly once, at $p = \bar{z}/H$ (this is in fact the curve we estimate off our data as described in Section 5). Panel (b) then illustrates how such a κ function generates risk aversion ($RP^*(p,r) > 0$) and subproportionality for large p and risk tolerance ($RP^*(p,r) < 0$) and superproportionality for small p (although the magnitude of the superproportionality is hard to see visually). Thus, the UP model can rationalize both Result 1 and Result 2.[18]

5 Estimating the Models and Comparing Performance

In Section 4, we saw that upside potential (UP) is the only model in our comparison set that can make qualitative predictions consistent with Results 1 and 2. However, it is also worth assessing the quantitative performance of the different models. To do so, we estimate each model on the aggregate means and assess their quantitative fit.

[18] McGranaghan et al. (2025) do note that their model may be able to rationalize findings from certainty equivalents of risk tolerance at low p and risk aversion at high p, though they have no certainty equivalents in their dataset.

Specifically, we conduct two exercises. First, we estimate each model using only the aggregate means when $r = 1$ (i.e., off 9 aggregate means). For these estimates, we compare models based on both their in-sample performance as well as their out-of-sample performance when we use the estimates to predict behavior throughout the remainder of the parameter space (i.e., for the 180 other aggregate means when $r < 1$). Second, we estimate each model using all 189 aggregate means, in which case we compare models using in-sample performance. Our two performance measures are the root mean-squared error (RMSE) and correlation between predicted and observed mean $RP(p, r)$.

Each model requires that we estimate a nonlinear function: PW, PT, and PPW require an inverse-S-shaped probability weighting function, while UP and SSV require an S-shaped transformation of outcomes. To put the models on equal footing, we impose a common functional form. In particular, the two-parameter specification of Lattimore et al. (1992) can capture both shapes (as well as other shapes). For any input $z \in [0, 1]$, define

$$f(z; \gamma, \delta) \equiv \frac{\gamma z^{\delta}}{\gamma z^{\delta} + (1 - z)^{\delta}}. \tag{7}$$

With this formulation, the parameter δ governs curvature (roughly, inverse-S-shape for $\delta < 1$ vs. S-shape for $\delta > 1$), while γ shifts the average level of the function. To ensure that f is increasing, δ and γ must be positive; in our estimation, we restrict them both to be at least 0.01.

To estimate each model's shape parameters (γ, δ), we use nonlinear least squares—see Appendix E for details. For PT, we additionally estimate the value-curvature parameter α from the power value function $v(x) = x^{\alpha}$.[19] Table 3 summarizes the resulting parameter estimates $(\hat{\gamma}, \hat{\delta}, \hat{\alpha})$. The top panel presents estimates using only the aggregate means when $r = 1$, along with the in-sample and out-of-sample performance measures. The bottom panel presents the corresponding estimates and performance measures using all 189 aggregate means.

Table 3 documents four key features of our estimation.[20] First, when estimating the PW model on the $r = 1$ data on which it is traditionally estimated, we estimate a typical inverse-S-shaped PW function and the estimated model performs well in sample. However, the performance of the estimated PW model deteriorates sharply when asked to extrapolate to the remainder of the parameter space. Indeed, the out-of-sample RMSE is roughly 9 times larger than the in-sample RMSE. This predictive inaccuracy derives from the estimated PW model incorrectly predicting substantial risk tolerance throughout the remainder of the (p, r) parameter space. When we instead estimate the PW model on the full (p, r) space, the model's shape changes dramatically, from an inverse-S-shape PW function to a convex PW function, and its predictive performance is poor.

Second, estimating the PT model—that is, adding a value function to the PW model—does

[19] In principle, estimating UP also requires estimating the scale of κ. Our analysis suggests the scale is quite large; however, we set the scale such that $\kappa(30) = 300$, as the estimates and RMSE's change very little if we permit a larger scale, see Appendix Table E.1.

[20] For a visual depiction of all the estimation results reported in this section, including the estimated functions and the in-sample and out-of-sample fit, see Appendix Figures E.1 - E.5.

Table 3: Model Estimates and Predictive Performance

	Prospect Theory (PT)	Probability Weighting (PW)	Proportional Prob. Weight. (PPW)	S-Shaped Value (SSV)	Upside Potential (UP)
Panel A: Estimated Using Nine Observations of $m(p,r)$ with $r=1$					
Level ($\hat{\gamma}$)	0.050	0.641	0.641	2.404	1.035
	(0.154)	(0.031)	(0.031)	(0.255)	(0.215)
Curvature ($\hat{\delta}$)	0.983	0.506	0.506	1.975	2.699
	(0.722)	(0.041)	(0.041)	(0.161)	(0.362)
Value curv. ($\hat{\alpha}$)	3.153				
	(2.977)				
In-sample RMSE	0.746	0.835	0.835	0.835	1.149
In-sample Corr	0.978	0.972	0.972	0.972	0.948
Out-of-sample RMSE	8.606	7.638	1.291	1.291	1.766
Out-of-sample Corr	0.759	0.751	0.954	0.954	0.915
Panel B: Estimated Using All 189 Observations of $m(p,r)$					
Level ($\hat{\gamma}$)	0.283	0.357	0.671	2.375	1.310
	(0.110)	(0.045)	(0.009)	(0.076)	(0.101)
Curvature ($\hat{\delta}$)	1.143	1.025	0.462	2.166	3.145
	(0.198)	(0.038)	(0.011)	(0.052)	(0.152)
Value curv. ($\hat{\alpha}$)	1.156				
	(0.252)				
RMSE	3.794	3.799	1.207	1.207	1.669
Corr	0.320	0.309	0.952	0.952	0.911

Notes: Entries reflect estimated parameters (with standard errors in parentheses) and measures of model performance for each of the five models described in Section 4. Applying the function f from equation 7, for PT, PW, and PPW, functional form for probability weighting is $\pi(q) = f(q;\gamma,\delta)$; for PT, functional form for value function is $v(x) = x^{\alpha}$. For SSV, functional form for value function is $v(x) = f(x/H;\gamma,\delta)$. For UP, functional form for κ function is $\kappa(x) = 300 \cdot f(x/H;\gamma,\delta)$. Parameters estimated via non-linear least squares by minimizing the squared distance between model-implied $m^*(p,r)$ and observed mean response $m(p,r)$. In panel A, models estimated using only nine aggregate means when $r = 1$; in panel B, models estimated using all 189 aggregate means. Model performance measures are root mean-squared error (RMSE) and correlation between predicted and observed mean $RP(p,r)$ (Corr); panel A reports both in-sample performance across nine observations with $r = 1$ and out-of-sample performance across 180 observations with $r < 1$; panel B reports (in-sample) performance across all 189 observations.

not help. Given the extra parameter, we of course get an improvement in in-sample performance, but this improvement is tiny. Moreover, the estimated functions do not take the typical forms, as we estimate a convex value function and a convex PW function that has $\pi(q) < q$ for all q. Finally, when estimating the PT model on only the $r = 1$ data, the out-of-sample performance is substantially worse than the estimated PW model.

Third, when estimating the UP model on the $r = 1$ data, we indeed estimate an S-shaped κ function, and the estimated model performs relatively well in sample with only a slightly higher in-sample RMSE than the PW model. At the same time, it substantially outperforms the PW model out of sample. Moreover, when we estimate the UP model on the full (p, r) space, the estimated κ function is pretty stable, and it substantially outperforms the PW model.

Fourth, the post-hoc models (PPW and SSV) actually exhibit the best quantitative performance for the aggregate mean data. It turns out that, when using the functional form in equation 7, these two models are observationally equivalent, and thus their performance measures are identical (see Appendix E for details.) These models perform well both in-sample and out-of-sample, and are slightly superior to UP. The source of this superior performance seems to be that these models are able to better fit the PB-TK effect at $r = 1$ (as reflected by their lower in-sample RMSE), combined with UP predicting stronger comparative statics with regard to r than we actually see in the aggregate mean data.

6 Individual-Level Analysis

Although the post-hoc models of PPW and SSV outperform UP on the aggregate mean data, it is also worth investigating how well the models explain the heterogeneity that we see in individual-level data. In particular, in Section 4, we derived a key difference in predictions. The two post-hoc models imply that normalized risk premia should be independent of r, and thus any sub- or superproportionality observed in individuals must reflect sampling variation. In contrast, UP permits that people might exhibit subproportionality or superproportionality, and moreover that there should be a tight connection between whether they are risk averse versus risk tolerant and whether they are subproportional versus superproportional.

For our individual-level analysis, we focus on parametric regressions based on equations 3 and 4. Specifically, we estimate:

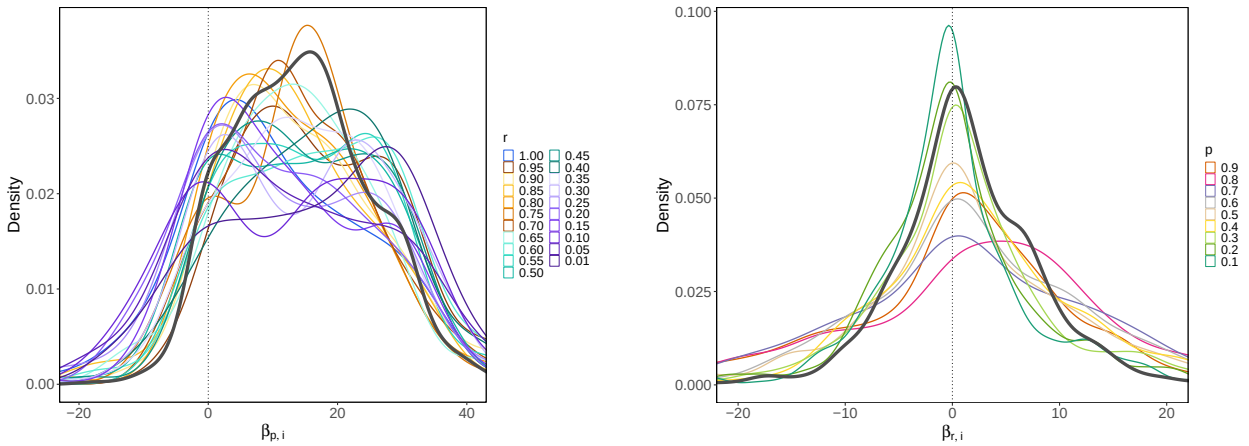
$$RP_i(p, \bar{r}) = \alpha_{p,i} - \beta_{p,i}(1 - p) + \epsilon_{p,i} \tag{8}$$

$$RP_i(\bar{p}, r) = \alpha_{r,i} - \beta_{r,i}(1 - r) + \epsilon_{r,i} \tag{9}$$

In our design, each subject provides valuations for (at least) five values of p for each of three values of r. Hence, for each subject, we can estimate $(\alpha_{p,i}, \beta_{p,i})$ for three values of r. Analogously, each subject provides valuations for (at least) five values of r for each of five values of p. Hence, for each subject, we can estimate $(\alpha_{r,i}, \beta_{r,i})$ for five values of p. Figure 6 provides an overview of the heterogeneity in our data by depicting the distributions of $\beta_{p,i}$ and $\beta_{r,i}$.[21]

[21]The caveat "at least" reflects that we sometimes have six or even seven observations for these individual-level

Figure 6: Distributions of Individual-Level Sensitivities to p and r



(a) Distribution of $\beta_{p,i}$

(b) Distribution of $\beta_{r,i}$

Notes: Panel (a) depicts kernel density estimates of distribution of $\beta_{p,i}$, where $\beta_{p,i}$ estimated from $RP_i(p, \bar{r}) = \alpha_{p,i} - \beta_{p,i}(1-p) + \varepsilon_{p,i}$. Colored curves correspond to the 21 values of $\bar{r}$; solid black curve overlays density of subject-level averages (each subject's mean $\beta_{p,i}$ across their three values of $\bar{r}$). Panel (b) analogously shows kernel density estimates of distribution of $\beta_{r,i}$, where $\beta_{r,i}$ estimated from $RP_i(\bar{p}, r) = \alpha_{r,i} - \beta_{r,i}(1-r) + \varepsilon_{r,i}$. Colored curves correspond to the 9 values of $\bar{p}$; solid black curve overlays the density of subject-level averages (each subject's mean $\beta_{r,i}$ across their five values of $\bar{p}$). Both panels truncated on horizontal axis to enhance readability; Appendix Figure F.1 contains untruncated versions.

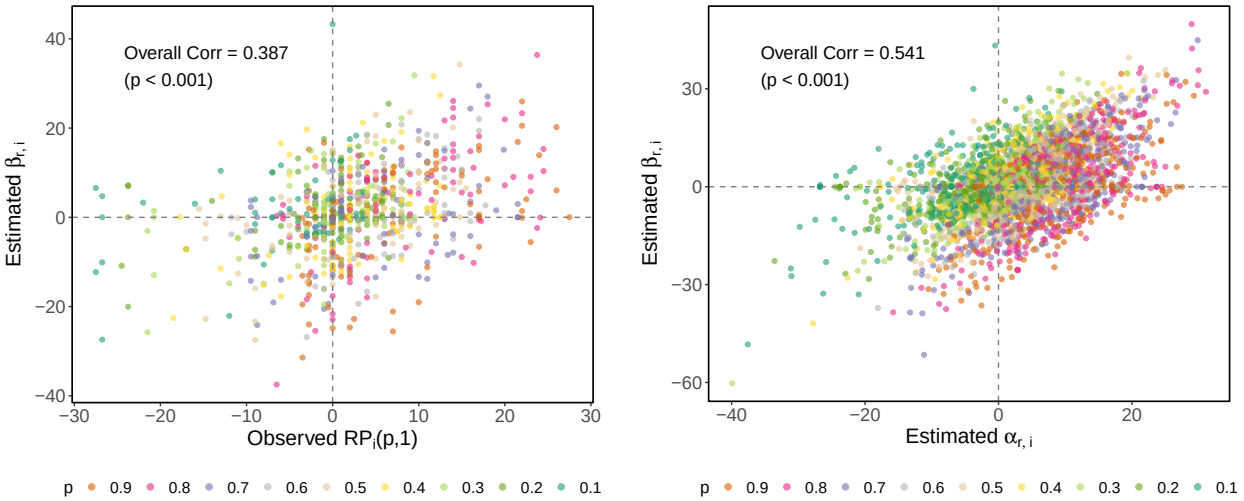
Panel (a) of Figure 6 provides the distribution of $\beta_{p,i}$. The colored curves depict kernel density estimates of $\beta_{p,i}$ separately for each of the 21 values of r. Each subject contributes to three of the colored curves; the solid black line in panel (a) depicts the kernel density estimate of subject-level average $\beta_{p,i}$, averaging across each subject's three $\beta_{p,i}$ values. Just as in the aggregate data, the feature of having normalized risk premia that increase with p appears to be a robust phenomena exhibited by most individuals. Across all individual-r combinations, 84.4% have a $\beta_{p,i}$ that is positive, and 59.6% have a $\beta_{p,i}$ that is larger than 10 (reflecting an \$8 increase in the normalized risk premium when moving from $p = 0.1$ to $p = 0.9$).

Panel (b) in Figure 6 provides the distribution of $\beta_{r,i}$. The colored curves depict kernel density estimates of $\beta_{r,i}$ separately for each of the 9 values of p. Here, each subject contributes to five of the colored curves; the solid black line in panel (b) depicts the kernel density estimate of subject-level average $\beta_{r,i}$, averaging across each subject's five $\beta_{r,i}$ values. Unlike $\beta_{p,i}$, the values for $\beta_{r,i}$ are relatively evenly distributed across positive and negative values: Across individual-p combinations, 55.1% are positive while 43.7% are negative. Moreover, at the individual-level, there are some large values—across individual-p combinations, 21.9% have $\beta_{r,i}$ more positive than 8 while 14.5% have $\beta_{r,i}$ more negative than -8. In other words, subjects sometimes exhibit both substantial subproportionality and substantial superproportionality.

This last finding brings us to the key question for this section: Are the positive and negative values for $\beta_{r,i}$ predicted by whether the subject is risk averse versus risk tolerant for that p (as

regressions (instead of five) due to the repeats. We can similarly calculate $\tau^a_{p,i}$ at the individual-r level and $\tau^a_{r,i}$ at the individual-p level; Appendix Figure F.2 is analogous to Figure 6 and looks much the same.

Figure 7: Correlation Between Risk Attitudes and Subproportionality



(a) Estimated $\beta_{r,i}$ vs. Observed $RP_i(p,1)$

(b) Estimated $\beta_{r,i}$ vs. Estimated $\alpha_{r,i}$

Notes: Panel (a) plots the estimated slope parameter ($\beta_{r,i}$) against the observed normalized risk premium at $r = 1$ ($RP_i(p,1)$) for the 838 participant-p combinations where we observe $RP_i(p,1)$. Panel (b) plots the estimated slope parameter ($\beta_{r,i}$) against the estimated intercept ($\alpha_{r,i}$) for all 4,000 participant-p combinations. For each participant-p combination, $(\alpha_{r,i}, \beta_{r,i})$ estimated from the regression $RP_i(p,r) = \alpha_{r,i} - \beta_{r,i}(1-r) + \varepsilon_{r,i}$. Points are color-coded by p. The reported Pearson correlation coefficient is the pooled correlation across all plotted observations. Note that points in panel (a) do not map identically to points in panel (b) because the x-axes differ.

predicted by the UP model)? To answer this question, we need a way to measure whether a person is risk averse or risk tolerant for a particular p. Given that valuations are elicited with noise combined with the fact that the UP model suggests the strongest risk aversion or risk tolerance when $r = 1$, we measure risk aversion versus risk tolerance at $r = 1$.

Hence, we initially limit the sample to individual-p combinations for which the individual provided a valuation when $r = 1$, which yields 838 individual-p combinations. For these observations, panel (a) of Figure 7 plots the estimated $\beta_{r,i}$ against the stated normalized risk premium at $r = 1$. Consistent with the UP model, when a participant exhibits risk aversion at $r = 1$ (i.e., $RP(p,1) > 0$), they tend to exhibit subproportionality (i.e., $\beta_{r,i} > 0$), and when a participant exhibits risk tolerance at $r = 1$ (i.e., $RP(p,1) < 0$), they tend to exhibit superproportionality (i.e., $\beta_{r,i} < 0$). The correlation between the normalized risk premium at $r = 1$ and $\beta_{r,i}$ is 0.39.

In order to use the full sample, we can instead measure risk aversion versus risk tolerance using the estimated $\alpha_{r,i}$, which in equation 9 represents the fitted normalized risk premium when $r = 1$. For all individual-p combinations, panel (b) plots the estimated $\beta_{r,i}$ against the estimated $\alpha_{r,i}$. The pattern is much the same, and the correlation between $\alpha_{r,i}$ and $\beta_{r,i}$ is 0.54. This connection between risk aversion/risk tolerance and sub/superproportionality is precisely what UP predicts and provides a sharp rejection of the post-hoc models of PPW and SSV.

Table 4: Classification of Subjects by $p_{\min}$-$p_{\max}$ Patterns

Risk Aversion at $r=1$ for $p_{\min}$ and $p_{\max}$	Sub- versus Superproportionality for $p_{\min}$ and $p_{\max}$				Total
	$\beta_{r,i}^{p_{\min}} \geq 0$ $\beta_{r,i}^{p_{\max}} \geq 0$	$\beta_{r,i}^{p_{\min}} \geq 0$ $\beta_{r,i}^{p_{\max}} < 0$	$\beta_{r,i}^{p_{\min}} < 0$ $\beta_{r,i}^{p_{\max}} \geq 0$	$\beta_{r,i}^{p_{\min}} < 0$ $\beta_{r,i}^{p_{\max}} < 0$	
$\alpha_{r,i}^{p_{\min}} \geq 0$ $\alpha_{r,i}^{p_{\max}} \geq 0$	127 (48.8% of Row) (15.9% of Total)	58 (22.3% of Row) (7.2% of Total)	39 (15.0% of Row) (4.9% of Total)	36 (13.8% of Row) (4.5% of Total)	260 (32.5%)
$\alpha_{r,i}^{p_{\min}} \geq 0$ $\alpha_{r,i}^{p_{\max}} < 0$	8 (19.5% of Row) (1.0% of Total)	19 (46.3% of Row) (2.4% of Total)	4 (9.8% of Row) (0.5% of Total)	10 (24.4% of Row) (1.2% of Total)	41 (5.1%)
$\alpha_{r,i}^{p_{\min}} < 0$ $\alpha_{r,i}^{p_{\max}} \geq 0$	84 (21.2% of Row) (10.5% of Total)	44 (11.1% of Row) (5.5% of Total)	180 (45.5% of Row) (22.5% of Total)	88 (22.2% of Row) (11.0% of Total)	396 (49.5%)
$\alpha_{r,i}^{p_{\min}} < 0$ $\alpha_{r,i}^{p_{\max}} < 0$	7 (6.8% of Row) (0.9% of Total)	24 (23.3% of Row) (3.0% of Total)	15 (14.6% of Row) (1.9% of Total)	57 (55.3% of Row) (7.1% of Total)	103 (12.9%)
Total	226 (28.2%)	145 (18.1%)	238 (29.8%)	191 (23.9%)	800 (100%)

Notes: For each subject, we select their largest and smallest values for p and label them $p_{\min}$ and $p_{\max}$. For each, $(\alpha_{r,i}, \beta_{r,i})$ estimated from the regression $RP_i(p,r) = \alpha_{r,i} - \beta_{r,i}(1-r) + \varepsilon_{r,i}$. Rows classify subjects by their risk aversion at $r=1$ for $p_{\min}$ and $p_{\max}$ (positive $\alpha_{r,i}$ is risk averse, negative $\alpha_{r,i}$ is risk tolerant). Columns classify subjects by their sub- versus superproportionality for $p_{\min}$ and $p_{\max}$ (positive $\beta_{r,i}$ is subproportional, negative $\beta_{r,i}$ is superproportional). Cells report the count, the within-row percentage (% of Row), and the overall percentage (% of Total). The last row/column reports marginal totals. Based on 800 subjects.

Result 3 (Risk attitudes predict sub- versus superproportionality)**.** *Consistent with the UP model, risk attitudes and r-responses are systematically linked: risk aversion is associated with subproportionality and risk tolerance is associated with superproportionality.*

We next investigate at the subject-level the tendency to show both the PB-TK effect and the connection between risk attitudes and sub- vs. superproportionality. Specifically, for each subject, we select the largest and smallest values for p that they saw and label them $p_{\max}$ and $p_{\min}$. We then estimate equation 9 for each, and label the estimates $(\alpha_{r,i}^{p_{\max}}, \beta_{r,i}^{p_{\max}})$ and $(\alpha_{r,i}^{p_{\min}}, \beta_{r,i}^{p_{\min}})$. Finally, we classify each of our 800 subjects by how the signs of their $\alpha_{r,i}$ and their $\beta_{r,i}$ change between their $p_{\min}$ and their $p_{\max}$.

Table 4 presents this classification. First note that, not surprisingly, the marginal distribution across the four rows clearly reflects the PB-TK effect. Some subjects are risk-averse for both their $p_{\min}$ and their $p_{\max}$ (32.5%), and some are risk-tolerant for both (12.9%), but among those who switch, the vast majority switch from risk tolerance at $p_{\min}$ to risk aversion for $p_{\max}$ (49.5% versus 5.1%). Second, within each row, we clearly see the connection between risk attitudes and sub- vs. superproportionality. Table 4 is organized such that the UP model predicts observations to be on the main diagonal, and indeed within each row, the cell on the main diagonal is the modal outcome. Moreover, we clearly reject the null hypothesis of independence between the rows and columns of

Table 5: Summary of Contributions

	Prospect Theory	Upside Potential
Panel A: Key Patterns in Prior Literature		
Key Pattern 1: PB-TK effect: risk aversion at high p, risk tolerance at low p when $r = 1$	Yes	Yes
Key Pattern 2: CRE: scaling down winning probabilities increases risk tolerance ($RP(p, r) < RP(p, 1)$) for p large and r small	Yes	Yes
Panel B: New Results from Our Data		
New Result 1: Risk aversion at high p, risk tolerance at low p for all r (and $RP(p, r)$ increasing in p for vast majority of individual-r combinations)	No	Yes
New Result 2: Modest subproportionality for high p, neutral or mild superproportionality for low p (and substantial heterogeneity across individual-p combinations)	No	Yes
New Result 3: Risk attitudes predict sub- versus superproportionality at the individual-p level	No	Yes

Notes: Table summarizes two key patterns from prior literature and three new results from our data. Last two columns characterize two models' ability to accommodate each pattern.

Table 4, Fisher's exact test $p < 0.01$.[22]

Before concluding our individual-level analysis, we revisit our model estimation at the individual level. For each subject, we estimate both the PW model and the UP model (i.e., estimations analogous to columns (2) and (5) in panel B of Table 3), using that subject's 30 observations for $m_i(p, r)$. Appendix Table F.2 compares individual RMSE values. For 2 subjects UP fails to converge. For the remainder, 403 of 798 are better fit by UP with an RMSE difference of at least 0.1, 295 are better fit by PW with an RMSE difference of at least 0.1, and 100 are approximately equally well fit with an RMSE difference of less than 0.1. Interestingly, the best-fitting UP models overwhelmingly point to an S-shaped κ, with 739 subjects estimated to have a shape that is initially convex and subsequently concave. In contrast, only 348 subjects have a best-fitting PW function that is inverse S-shaped (concave and then convex) in line with classical assumptions, while for 412 it is S-shaped (convex and then concave) and for 23 it is convex (with $\pi(q) < q$ for all q).

7 Discussion

In this paper, we attempt to gain a more complete understanding of people's preferences when comparing binary lotteries. The left-hand column of Table 5 summarizes our contribution.

The prior literature has highlighted two key patterns in comparisons of binary lotteries. First, there is the PB-TK Effect wherein certainty equivalents for binary lotteries reflect risk aversion for large probabilities and risk tolerance for small probabilities. Second, there is the CRE wherein, relative to risk attitudes when comparing a binary lottery to a sure amount, scaling down the

[22] Appendix Table F.1 reports the analogous transition matrix constructed over *all* within-subject pairs, where each subject contributes $\binom{5}{2} = 10$ pairwise comparisons. The message is essentially the same.

winning probabilities in each lottery by a common ratio increases risk tolerance. Importantly, evidence on these two key patterns covers a narrow range of the possible (p, r) parameter space. The PB-TK effect is a reliably observed comparative static with respect to p, but this comparative static has been studied only when $r = 1$. The CRE is a stylized fact about a comparative static with respect to r, and in principle it could hold for the entire parameter space, but the evidence for it comes from a small number of parameter combinations (typically for p large and r small; see McGranaghan et al. (2025)).

We instead study a broad set of binary-lottery comparisons that span the (p, r) parameter space and permit us to more fully understand comparative statics with respect to both p and r. Our findings are not what people might have expected.

Perhaps our most striking result is New Result 1: Not only do we see a transition from risk aversion at high p to risk tolerance at low p when $r = 1$, but we also see that same transition across all r that we consider. Indeed, we see this strong dependence of risk preferences on p at both the aggregate level and the individual level. New Result 2 is perhaps also surprising: At the aggregate level, we see only a mild impact of r on risk preferences—that is, our data clearly contradict the stylized fact of there being a substantial global CRE or global subproportionality. More interesting, this aggregate-level finding masks substantial heterogeneity across individual-p combinations, where we see both substantial subproportionality and substantial superproportionality, even for the same individual across different p. In trying to understand this heterogeneity, we came upon New Result 3: There seems to be a strong link between people's risk attitudes and whether they exhibit sub- versus superproportionality. Indeed, of the sixteen possible categories into which we parse participants in Table 4, participants in the modal category (22.5% of participants) exhibit both (i) risk aversion and subproportionality at the highest p that they saw and (ii) risk tolerance and superproportionality at the lowest p that they saw.

We have highlighted how our results might be surprising; the reason is that they are inconsistent with prospect theory, perhaps the most prominent alternative to EU. In particular, the shape of the probability weighting function in prospect theory is motivated by Key Patterns 1 and 2. However, this shape has implications for what we should expect to see through the rest of the (p, r) parameter space, and our New Results 1, 2, and 3 are wholly inconsistent with those predictions.[23] In light of this, we also provide some initial analysis on what might explain the full pattern of empirical results that we are seeing. Specifically, we discover that the model of upside potential (UP)—proposed by McGranaghan et al. (2025) to explain the substantial preference for probabilistic mixtures that they see in their data—turns out to be quite consistent with all three of our new results. The UP model has its limitations—see the discussion in McGranaghan et al. (2025). Nonetheless, the fact that it can capture the perhaps surprising new empirical results in both their paper and here can hopefully provide insight in the development of new models of risk preferences.

Although we view our empirical results as important, there is clearly more to explore. In

[23] In fact, many other alternatives to EU are motivated at least in part by an assumption of a global CRE, and thus our results are also inconsistent with those models. For an analysis of some other non-EU models' predictions for the CRE, see McGranaghan et al. (2025).

particular, we limited our attention to just one type of binary-lottery comparison in which each lottery involves some probability of winning a positive amount and otherwise one gets zero. While this type of comparison has been prominent in the prior literature, it would be valuable to explore other types of binary-lottery comparisons as well, such as comparing two binary lotteries with no shared outcomes, or comparing binary lotteries when some or all outcomes involve losing some amount. Furthermore, just as PT, UP, and other theories of risk preferences make predictions for the full parameter space of binary lotteries, they also make predictions for comparisons of lotteries with more than two possible outcomes, and thus we need to better understand preferences in that domain as well.

Finally, we highlight a broader message from our paper. In the decision theory and the behavioral economics literatures, theories are frequently built to explain evidence. However, in these exercises, the literature frequently takes an empirical finding from limited examples (or from a limited portion of the parameter space) and extrapolates it into an assumption of a global feature of preferences—analogous to taking evidence of a CRE in limited examples and extrapolating to an assumption of a global CRE or even global subproportionality. While such assumptions can be useful in the initial development of theories, our findings suggest caution, because at some point one must assess whether those extrapolations are valid. We have started that assessment in the domain of comparisons of binary lotteries, but we suspect there is scope for similar assessments in various other domains.

References

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école américaine. *Econometrica*, 21(4):503–546.

Bell, D. E. (1985). Disappointment in decision making under uncertainty. *Operations Research*, 33(1):1–27.

Bernheim, B. D. and Sprenger, C. (2020). On the empirical validity of cumulative prospect theory: Experimental evidence of rank-independent probability weighting. *Econometrica*, 88(4):1363–1409.

Blavatskyy, P., Panchenko, V., and Ortmann, A. (2023). How common is the common-ratio effect? *Experimental Economics*, 26:253–272.

Bruhin, A., Fehr-Duda, H., and Epper, T. (2010). Risk and rationality: Uncovering heterogeneity in probability distortion. *Econometrica*, 78(4):1375–1412.

Chen, D. L., Schonger, M., and Wickens, C. (2016). otree - an open-source platform for laboratory, online and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97.

Chew, S. H., Epstein, L. G., and Segal, U. (1991). Mixture symmetry and quadratic utility. *Econometrica*, 59(1):139–163.

Gonzalez, R. and Wu, G. (1999). On the shape of the probability weighting function. *Cognitive Psychology*, 38(1):129–166.

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.

Lattimore, P. K., Baker, J. R., and Witte, A. D. (1992). The influence of probability on risky choice: A parametric examination. *Journal of Economic Behavior and Organization*, 17(3):377–400.

MacCrimmon, K. R. and Larsson, S. (1979). Utility theory: Axioms versus 'paradoxes'. In *Expected utility hypotheses and the allais paradox: Contemporary discussions of the decisions under uncertainty with allais' rejoinder*, pages 333–409. Springer.

Machina, M. J. (1982). "Expected utility" analysis without the independence axiom. *Econometrica*, 50(2):277–323.

McGranaghan, C., Nielsen, K., O'Donoghue, T., Somerville, J., and Sprenger, C. D. (2025). Connecting common ratio and common consequence preferences. Working paper.

Prelec, D. (1998). The probability weighting function. *Econometrica*, 66(3):497–527.

Preston, M. G. and Baratta, P. (1948). An experimental study of the auction-value of an uncertain outcome. *The American Journal of Psychology*, 61(2):183–193.

Tversky, A. and Fox, C. R. (1995). Weighing risk and uncertainty. *Psychological Review*, 102(2):269.

Tversky, A. and Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323.