

Measuring Economic Preferences in the Presence of Noise: The Connections Between Choices and Valuations

Ted O'Donoghue* Charles Sprenger[†] Po Hyun Sung[‡] Ben Wincelberg[§]

August 30, 2026

Abstract

Past research highlights failures of “procedural invariance” when measuring economic preferences using choices versus valuations. We reassess these failures by examining theoretical connections between choices and valuations when preferences are stable but measurements are noisy and individuals are heterogeneous. Even under strong assumptions governing noise and heterogeneity, stability does not generally imply identical measurements. We develop new tests of stable preferences in conjunction with various ancillary assumptions about heterogeneity and noise. We implement these tests using existing data to understand if, in the domain of risk preferences, choices and valuations truly differ and to provide quantitative assessments of any deviations. Limiting to the types of data used in the prior literature, we rarely reject the null of stable preferences. With richer data linking individual choices and valuations and structural assumptions, we find evidence of instability which differs qualitatively from the received wisdom that choices implicate greater risk aversion than valuations.

Keywords: Preference Reversal; Stochastic Choice; Choice Experiment.

JEL Codes: D01, D81

*Cornell University, Department of Economics. Email: edo1@cornell.edu

[†]Caltech, Division of Humanities and Social Sciences. Email: sprenger@caltech.edu

[‡]Caltech, Division of Humanities and Social Sciences. Email: psung@caltech.edu

[§]Caltech, Division of Humanities and Social Sciences. Email: bwincelb@caltech.edu

1 Introduction

Since the original economic experiment of [Thurstone \(1931\)](#), researchers have been interested in using behavioral measurement to assess individual preferences, evaluate adherence to underlying axioms, and develop predictions based on models fit to measurement data. To date, there exists no consensus on how such measurement should be conducted—that is, on what type of elicitation technique to use—with different researchers advocating for different methods.

Two prominent elicitation techniques are *choices* and *valuations*: On one hand, we could ask people to make binary choices between options; on the other hand, we could ask people to provide valuations of options using some index of value. Ideally, these two elicitation techniques would yield measurements that are consistent with each other, as they are both meant to measure the same underlying preferences. However, the literature suggests that they do not. [Lichtenstein and Slovic’s \(1971\)](#) famous “preference reversals” in risk preferences, popularized in economics by [Grether and Plott \(1979\)](#), suggest apparent inconsistencies between choices versus valuations: The proportion of subjects choosing a safer over a riskier alternative is frequently found to exceed the proportion of subjects whose valuations are higher for the safer alternative, suggesting greater risk aversion in choices. More recently, an experimental-economics literature (see, e.g., [Brown and Healy, 2018](#); [Freeman et al., 2019](#); [Freeman and Mayraz, 2019](#)) documents similar apparent inconsistencies when comparing choices versus multiple price lists (MPLs), which are a type of valuation.

Researchers have interpreted these failures of “procedural invariance” in (at least) two ways. The preference-reversals literature typically interprets the evidence as suggesting there do not exist stable economic primitives like preferences in the first place, with preferences constructed contextually in each environment ([Tversky and Thaler, 1990](#); [Slovic, 1995](#); [Lichtenstein and Slovic, 2006](#)). The choices-versus-MPLs literature interprets the evidence as suggesting that valuations, choices, or both fail to credibly measure the preferences of interest.¹

In this paper, we assess these claims. We develop a model of how people respond in choice tasks and valuation tasks. Inspired by the literature above, we frame this model in terms of elicitations of risk preferences, but we highlight in our concluding

¹For example, ([Freeman et al., 2019](#)) argue in favor of choices as the “gold standard” for measurement, and thus argue that, given the inconsistencies between measures, MPLs must be yielding biased measurement.

section how our theoretical insights apply to other domains. We use our model to investigate the conditions under which one would expect no apparent inconsistencies, and more generally to develop tests of the null of stable preferences in conjunction with various ancillary assumptions. We then implement these new tests using existing data to understand if, in the domain of risk preferences, choices and valuations truly differ and, if so, to provide quantitative assessments of any deviations.²

Our exercise begins with three basic observations. First, economic measurement of preferences—like all scientific measurement—is inherently noisy.³ Second, there is no a priori reason to expect that different elicitation techniques are subject to identical noise, and so we do not impose this assumption (and indeed our empirical analysis seems to contradict it). Third, most experiments study aggregate data across individuals, and thus we must account for the fact that individuals are fundamentally heterogeneous in their underlying preferences and possibly also in their decision noise.

With these preliminaries in hand, we provide three groups of theoretical results. First, we derive predictions for the empirical objects used in the prior literature: When subjects are asked to compare a safer option to a riskier option, we can study the proportion of subjects who choose the safer option (which we denote by $\widehat{\mathbb{P}}_c(\text{safe})$), or the proportion of subjects whose valuations indicate a preference for the safer option (which we denote by $\widehat{\mathbb{P}}_v(\text{safe})$). The prior literature tests the null hypothesis of stable preferences by examining whether $\widehat{\mathbb{P}}_c(\text{safe}) = \widehat{\mathbb{P}}_v(\text{safe})$, and the common finding is $\widehat{\mathbb{P}}_c(\text{safe}) > \widehat{\mathbb{P}}_v(\text{safe})$. We highlight that, even under the stable-preference null, our model need not predict that $\mathbb{P}_c(\text{safe}) = \mathbb{P}_v(\text{safe})$.⁴

²A more recent literature provides cognitive accounts for violations of procedural invariance under the maintained hypothesis of stable preferences—e.g., [Bouchouicha et al. \(2026\)](#) for choices presented independently versus in list form, and [Shubatt and Yang \(2026\)](#) for different types of valuation tasks. Within these frameworks, noisy coding of value (in the spirit of e.g., [Khaw et al. \(2021\)](#)) interacts with the differential demands of different procedures to generate apparent inconsistencies. [Bouchouicha et al. \(2026\)](#) and [Shubatt and Yang \(2026\)](#) both provide experimental evidence of apparent inconsistencies that are in line with their proposed cognitive accounts. The fundamental distinction with our own work is that we develop formal tests of stable preferences across unbiased measurements while assuming symmetric noise, while they develop potential theories for how cognitive noise can generate differentially biased measurements and thus lead to distinct measures under stable preferences.

³This observation is most often attributed to [Mosteller and Nogee \(1951\)](#), though even [Thurstone \(1931\)](#) carries some recognition of noise in measurement. In the domain of economics experiments, it is natural to interpret noise in measurement as decision noise, where people’s decisions reflect a noisy instantiation of their underlying preferences; we often use the terms measurement noise and decision noise interchangeably.

⁴Throughout, we use $\widehat{\mathbb{P}}_c(\text{safe})$ and $\widehat{\mathbb{P}}_v(\text{safe})$ to denote empirical objects, and $\mathbb{P}_c(\text{safe})$ and $\mathbb{P}_v(\text{safe})$

Hence, we derive the attainable set for $(\mathbb{P}_c(\text{safe}), \mathbb{P}_v(\text{safe}))$ —that is, combinations consistent with our model—under the assumption of stable preferences combined with different sets of assumptions about heterogeneity and noise. Under our most basic set of assumptions about noise at the level of individual subjects, the decision probabilities for each subject must satisfy the simple consistency condition of being on the same side of $\frac{1}{2}$ (i.e., $\mathbb{P}_{ci}(\text{safe})$ and $\mathbb{P}_{vi}(\text{safe})$ both greater than $\frac{1}{2}$ or both less than $\frac{1}{2}$). Aggregating across subjects with our weakest assumptions on heterogeneity in preferences and noise—which permit arbitrary combinations of individuals each satisfying the consistency condition above—generates an attainable set that encompasses much of the $(\mathbb{P}_c(\text{safe}), \mathbb{P}_v(\text{safe}))$ unit square. Stronger assumptions shrink the attainable set, but even under our strongest set of assumptions about heterogeneity and noise, the attainable set still encompasses half the unit square. These theoretical results motivate new tests of the null of stable preferences in which one compares an observed combination $(\widehat{\mathbb{P}}_c(\text{safe}), \widehat{\mathbb{P}}_v(\text{safe}))$ to null attainable regions instead of the traditional test of identical measurements.

Our second group of theoretical results demonstrates how richer data can permit additional tests. If, in addition to $(\widehat{\mathbb{P}}_c(\text{safe}), \widehat{\mathbb{P}}_v(\text{safe}))$, the data also provide information on the mean preference within the population, we can develop predictions for $\mathbb{P}_c(\text{safe})$ conditional on the mean preference. Similarly, if we have access to within-subject data on both choices and valuations, we can develop subject-level predictions for $\mathbb{P}_c(\text{safe})$ conditional on both the mean preference and their own stated valuation.

Our third group of theoretical results highlights what one can do if one has within-subject data on both choices and valuations and is willing to take a more structural approach by imposing functional assumptions on preference heterogeneity and noise. We focus on one example in which we assume that preference heterogeneity and noise are jointly normal and independent. We derive how this structure naturally leads to probit and maximum likelihood estimators that can be used both for testing the null of stable preferences and, when stability is rejected, for estimating the magnitude of deviations from stability.

Motivated by our theoretical development, we conduct empirical analyses that implement our new tests. We use two datasets constructed from existing data. First, we assemble a dataset of prior experiments from the preference-reversal and choices-versus-MPLs literatures (323 experiments in total), where we observe $\widehat{\mathbb{P}}_c(\text{safe})$ and $\widehat{\mathbb{P}}_v(\text{safe})$ to denote theoretical predictions for those objects.

for each experiment. Second, we pull data from two recent papers (McGranaghan et al., 2024, 2026) that collect choices and valuations for the same subjects (424 experiments in total).⁵

We document that, when approached from the perspective of the prior literature, both datasets are consistent with the received wisdom that people are more likely to choose a safer option than valuations would imply: across these datasets, approximately 70-80% of experiments indicate $\widehat{\mathbb{P}}_c(\text{safe}) > \widehat{\mathbb{P}}_v(\text{safe})$. However, when comparing $(\widehat{\mathbb{P}}_c(\text{safe}), \widehat{\mathbb{P}}_v(\text{safe}))$ to the attainable sets that we derive under our various sets of assumptions, we cannot reject the null of stable preferences for the majority of experiments. For instance, even under our strongest set of assumptions, only 45% of the prior literature’s experiments and 23% of the McGranaghan et al. (2024, 2026) experiments lie outside of the attainable region for stable preferences. Moreover, even fewer (20% and 12%) can be statistically differentiated from the attainable regions—that is, across the two datasets, 80-90% of experiments are consistent with stable preferences.

Because the McGranaghan et al. (2024, 2026) dataset contains within-subject data on both choices and valuations, we can use it to assess our conditional predictions. Here, we see more mixed evidence. When conditioning on an estimate for the mean preference within the population, $\widehat{\mathbb{P}}_c(\text{safe})$ looks largely consistent with stable preferences. In contrast, when conditioning on both the mean preference and a subject’s own stated valuation, $\widehat{\mathbb{P}}_c(\text{safe})$ often looks inconsistent with stable preferences.

Finally, we can also use the McGranaghan et al. (2024, 2026) dataset to pursue our more structural approach. The data contain 84 distinct groupings on which we can conduct our analysis (see Section 4 for details). Under the maintained assumption of joint normality, we are remarkably well-powered, and we reject the null of stable preferences in roughly 80% of the 84 groupings. That said, when we study the nature of these deviations, the message differs from the received wisdom that choices induce greater risk aversion than valuations. In fact, overall, there appears to be no systematic direction to violations: The estimates indicate that choices induce greater risk aversion in roughly half of groupings, whereas valuations induce greater risk aversion in the other half. Perhaps more interesting, different types of experiments reveal different directional deviations. For decisions between a certain amount and a binary lottery,

⁵McGranaghan et al. (2024, 2026) collect this data for different purposes from ours, and neither directly compares choices and valuations in the way that we do here.

the estimates indicate that choices tend to induce more risk aversion; for decisions between a certain amount and a trinary lottery, the estimates indicate that valuations tend to induce more risk aversion; and for decisions between two binary lotteries, the estimates indicate no systematic deviation.

Our analysis has three important implications for the central question of whether there is preference stability across measurement techniques. First, and most importantly, *prima facie* inconsistencies across measurement techniques do not necessarily imply a failure of stable preferences. Our analysis highlights how, to formally test for preference stability, the analyst must specify assumptions about heterogeneity and noise so as to identify the appropriate prediction to test. Our analysis further highlights that, for any test, rejection could mean a rejection of stable preferences but also could be a rejection of those ancillary assumptions.

Second, we provide guidance for experiments designed to assess preference stability. In particular, our analysis highlights the value of collecting within-subject data on both choices and valuations so as to conduct more powerful tests. One interpretation of our analysis is that tests based on whether observed combinations of $(\widehat{\mathbb{P}}_c(\text{safe}), \widehat{\mathbb{P}}_v(\text{safe}))$ fall within some acceptable range have low power to detect deviations from stability relative to our tests based on richer data. Indeed, under our strongest assumptions of full joint normality, we infrequently reject the null of stable preferences when using tests based on observed combinations of $(\widehat{\mathbb{P}}_c(\text{safe}), \widehat{\mathbb{P}}_v(\text{safe}))$, whereas with the same data we frequently reject the null of stable preferences when we use the connections in the data to study richer empirical objects.

Third, our empirical analysis that implements our tests suggest a very different message from the received wisdom that choices induce greater risk aversion than valuations. Our results add nuance to a literature that often draws broad conclusions about the robustness and directionality of preference reversals. Further work with purpose-built experiments would be valuable for assessing whether our nuanced findings are robust and for investigating their source.

This manuscript proceeds as follows: Section 2 reviews prior work evaluating procedural invariance and testing for preference stability; Section 3 presents our theoretical framework for evaluating the relationship between choices and valuations under various assumptions about heterogeneity and noise and develops our new tests; Section 4 reports on our empirical implementation of our new tests; and Section 5 provides a discussion drawing out implications for the empirical literature on procedural

invariance, and concludes.

2 Prior Evidence on Choices versus Valuations

While our theoretical insights can be applied more broadly (as we discuss in Section 5), we focus on apparent inconsistencies between two elicitation procedures: choices versus valuations. In this section, we clarify exactly what we mean by inconsistencies between these two procedures, and then provide an overview of prior evidence.

2.1 Choices and Valuations

Consider two types of tasks:

- **Choice task:** We ask a person to make a binary choice between two options.
- **Valuation task:** We ask a person to state a scalar value y_i that makes two options of equal value.

While choice tasks are straightforward, valuation tasks require additional description. In the domain of choice between lotteries, there are a variety of scalar values that we could elicit. Here are four examples:⁶

Valuation Task A: State a y_i such that $(y_i, 1)$ is of equal value to $(30, 0.6)$.

Valuation Task B: State a y_i such that $(20, 1)$ is of equal value to $(y_i, 0.2; 20, 0.3)$.

Valuation Task C: State a y_i such that $(y_i, 0.4)$ is of equal value to $(30, 0.2)$.

Valuation Task D: State a y_i such that $(20, y_i)$ is of equal value to $(30, 0.6)$.

The simplest form for a valuation task is when we ask a person to state a certainty equivalent for a lottery (as in Valuation Task A). But valuation tasks can come in many other forms depending on which value is elicited. For instance, even when comparing a certain amount to a lottery, rather than elicit the sure amount, we could elicit one of the amounts in the lottery (as in Valuation Task B). Alternatively, we might compare two lotteries, in which case we could elicit one of the amounts in one of

⁶To simplify notation, whenever we use lottery notation, we suppress zero outcomes. Hence, $(30, 0.6)$ represents a lottery to yields 30 with probability 0.6 and 0 with probability 0.4.

the lotteries (as in Valuation Task C) or one of the probabilities in one of the lotteries (as in Valuation Task D). In our language, all of these fall in the broad category of valuation tasks.

For any specific valuation task, there are multiple ways to elicit a scalar indifference value. For instance, we might simply ask a person to state a value (i.e., fill in the blank) or make a bid in the Becker-DeGroot-Marschak random second-price auction mechanism. Alternatively, we might use a multiple price list (MPL)—that is, a sequence of binary choices presented in list form where only the value of interest changes across rows.

2.2 Apparent Inconsistencies Between Choices and Valuations

Suppose we want to measure a person’s preference between two lotteries. One way to measure that preference is to use a choice task that asks the person to select between the two lotteries. A second way to measure that preference is through a valuation task (or tasks) such that the stated value (or values) convey information on the preference between the two lotteries.⁷

Our interest is studying apparent inconsistencies between these two approaches. Hence, for our purposes, an *experiment* fixes a pair of lotteries, and then one treatment asks subjects to make a binary choice between the two lotteries, while a second treatment (within- or between-subjects) elicits a valuation (or valuations) that suggests a preference between the two lotteries. In all experiments we consider, one lottery can clearly be labeled the safer lottery, and the empirical quantities (at least initially) are the proportion of subjects who choose the safer option in the choice task, which we denote by $\widehat{\mathbb{P}}_c(\text{safe})$, and the proportion of subjects whose valuation(s) suggests they prefer the safer option, which we denote by $\widehat{\mathbb{P}}_v(\text{safe})$. In line with the prior literature, we say there are apparent inconsistencies when these empirical quantities differ.

2.3 Prior Evidence

The earliest comparison of choices versus valuations (at least in the domain of risk) comes from the literature on “preference reversals” initially developed by [Lichtenstein](#)

⁷For instance, suppose the two lotteries of interest are $(14, 1)$ versus $(30, \frac{1}{2})$. If we use a valuation task in which we ask the person to state a y_i such that $(y_i, 1)$ is of equal value to $(30, \frac{1}{2})$, then a stated $y_i < 14$ suggests a preference for $(14, 1)$ over $(30, \frac{1}{2})$, whereas a stated $y_i > 14$ suggests a preference for $(30, \frac{1}{2})$ over $(14, 1)$.

and Slovic (1971) and later popularized in economics by Grether and Plott (1979). In a typical experiment, subjects are asked to compare two lotteries, one with a high chance of a moderate prize (called a P -bet) and one with a low chance of a high prize (called a $\$$ -bet); note that the P -bet is the safer option. First, subjects are presented with both bets and asked to make a choice. Second, subjects are asked to state a reservation value (i.e., a certainty equivalent) for each lottery independently. The researchers then compare (i) the proportion of subjects who choose the safer option but state a higher valuation for the riskier option to (ii) the proportion of subjects who choose the riskier option but state a higher valuation for the safer option. The typical finding is that the former substantially exceeds the latter. This finding is formally equivalent to $\widehat{\mathbb{P}}_c(\text{safe}) > \widehat{\mathbb{P}}_v(\text{safe})$, that is, that people appear more risk averse when making choices than they do when providing valuations.⁸

More recently, a small literature has investigated whether people behave consistently when making a binary choice versus when that binary choice is part of an MPL (Brown and Healy, 2018; Freeman et al., 2019; Freeman and Mayraz, 2019). This literature uses a between-subjects design with two treatments. In the choice treatment, subjects are asked to make a choice between two lotteries. In the MPL treatment, subjects are asked to complete an MPL in which one of the rows is the binary choice from the choice treatment. The researchers then compare the proportion of subjects who choose the safer option in the choice treatment versus the proportion of subjects who choose the safer option in the focal row in the MPL treatment. Again, the common empirical finding is that $\widehat{\mathbb{P}}_c(\text{safe}) > \widehat{\mathbb{P}}_v(\text{safe})$, that is, that people appear more risk averse when making choices than they do when providing valuations.

Most recently, McGranaghan et al. (2024, 2026) conduct experiments in which they collect choices and valuations from the same subjects. Their goal was not to compare choices versus valuations; rather, their focus was comparing pairs of valuations to study the phenomena of the common ratio effect, the common consequence effect, and preferences for probabilistic mixtures. They collected data on corresponding pairs of choices to illustrate an inference problem that arises when studying paired-choice tasks (in McGranaghan et al. (2024)) and to show that paired choices and paired valuations can yield similar conclusions in some instances (in McGranaghan

⁸To see this equivalence, let (a, b) denote the pattern choosing option $a \in \{\text{safe}, \text{risky}\}$ and stating a higher valuation for option $b \in \{\text{safe}, \text{risky}\}$. Then $\widehat{\mathbb{P}}_c(\text{safe}) = \widehat{\mathbb{P}}(\text{safe}, \text{safe}) + \widehat{\mathbb{P}}(\text{safe}, \text{risky})$ while $\widehat{\mathbb{P}}_v(\text{safe}) = \widehat{\mathbb{P}}(\text{safe}, \text{safe}) + \widehat{\mathbb{P}}(\text{risky}, \text{safe})$, and thus $\widehat{\mathbb{P}}_c(\text{safe}) > \widehat{\mathbb{P}}_v(\text{safe})$ if and only if $\widehat{\mathbb{P}}(\text{safe}, \text{risky}) > \widehat{\mathbb{P}}(\text{risky}, \text{safe})$.

et al. (2026)). However, their data provide a convenience dataset that we can use for the purposes of this paper. When we reanalyze their data to directly compare individual choices versus valuations, again it turns out that the vast majority (80%) of comparisons exhibit $\widehat{\mathbb{P}}_c(\text{safe}) > \widehat{\mathbb{P}}_v(\text{safe})$.

In Section 4, we provide more details on the literatures above and their data, and we then use that data for our empirical application. But first we turn to our main research question: What exactly can we infer when we observe apparent inconsistencies of the form described above?

3 Theoretical Analysis of Choices versus Valuations

Apparent inconsistencies between choices and valuations have been interpreted in two primary ways. On one hand, the preference-reversals literature uses apparent inconsistencies to suggest that there do not exist stable underlying preferences. On the other hand, the choices-versus-MPLs literature uses apparent inconsistencies to suggest that one or both elicitation procedures are generating biased measures of underlying preferences. Our goal is to assess these claims by developing a theoretical model that makes predictions for observed behavior under each elicitation procedure.

To simplify our exposition, our analysis in the text focuses on the simple case where we want person i to compare a lottery X to a sure payment of amount y . However, the main results and intuitions generalize to the case where a person makes lottery-versus-lottery comparisons, and they also generalize to the paradigm from the preference-reversals literature where a person chooses between two lotteries X and Z and also provides certainty equivalents for both (see Appendix C).

There are two ways in which we can ask a person to make a comparison between a lottery X and a sure payment of amount y :

- **Choice task:** We ask person i to make a binary choice between X and y .
- **Valuation task:** We ask person i to state the value y_i that is of equal value to them as the lottery X .

As described in Section 2, the prior literature compares two quantities: $\widehat{\mathbb{P}}_c(y)$, the empirical choice frequency of y over X , and $\widehat{\mathbb{P}}_v(y)$, the empirical proportion

of individuals who exhibit a valuation for X that is lower than the given y .⁹ In experimental data, $\widehat{\mathbb{P}}_c(y)$ and $\widehat{\mathbb{P}}_v(y)$ depend on a combination of underlying population values $\mathbb{P}_c(y)$ and $\mathbb{P}_v(y)$ and sampling variation due to having finite experimental samples. We address the latter in Section 4; in this section, we develop a theoretical model of the underlying population values $\mathbb{P}_c(y)$ and $\mathbb{P}_v(y)$.

The underlying population values $\mathbb{P}_c(y)$ and $\mathbb{P}_v(y)$ will reflect a combination of noise at the individual level and heterogeneity in preferences at the aggregate level. Hence, in building a stochastic model to make predictions for these quantities, we start with a model of stochastic behavior at the individual level (in Section 3.1), and we then incorporate heterogeneity in preferences to generate stochastic behavior at the aggregate level (in Section 3.2).

3.1 Individual Behavior

We begin with a model of individual-level behavior, and in particular how underlying preferences combine with noise to generate stochastic behavior at the individual level. Throughout, we assume:

- For a choice task between X and y , individual i 's underlying preferences involve an indifference value y_{ci}^* such that $y_{ci}^* \sim_{ci} X$.¹⁰ For their actual choice, individual i chooses y over X when y is larger than $y_{ci} \equiv y_{ci}^* + \varepsilon_c$, where choice noise ε_c has CDF F_c . Individual i thus chooses y with probability $\mathbb{P}_{ci}(y) = \mathbb{P}(y_{ci}^* + \varepsilon_c \leq y) = F_c(y - y_{ci}^*)$.
- Likewise, for a valuation task, individual i 's underlying preferences involve an indifference value y_{vi}^* such that $y_{vi}^* \sim_{vi} X$. For their actual valuation, individual i reports valuation $y_{vi} = y_{vi}^* + \varepsilon_v$, where valuation noise ε_v has CDF F_v . Thus, the probability that individual i states a valuation for X that is lower than y is $\mathbb{P}_{vi}(y) = \mathbb{P}(y_{vi}^* + \varepsilon_v \leq y) = F_v(y - y_{vi}^*)$.
- We assume that ε_c and ε_v are mean zero.

⁹For a comparison between a lottery X and a sure amount y , the latter is clearly the safer option, and thus $\widehat{\mathbb{P}}_c(y)$ and $\widehat{\mathbb{P}}_v(y)$ correspond to $\widehat{\mathbb{P}}_c(\text{safe})$ and $\widehat{\mathbb{P}}_v(\text{safe})$ from Section 2.

¹⁰We use $\sim_{ci}$, $\prec_{ci}$, and $\succ_{ci}$ to denote underlying preferences for choice tasks, and $\sim_{vi}$, $\prec_{vi}$, and $\succ_{vi}$ to denote underlying preferences for valuation tasks. Under the null of stable preferences across the two tasks, we instead use $\sim_i$, $\prec_i$, and $\succ_i$ for both.

The key null hypothesis that we assess in this paper is whether each individual has a stable underlying preference that is used for both choice tasks and valuation tasks. We state this formally as Assumption 0:

Assumption 0 (stable preferences). *Each individual i has a stable indifference value; in other words, each individual i has $y_{ci}^* = y_{vi}^* \equiv y_i^*$.*

With stable preferences, the two key equations are:

$$\mathbb{P}_{ci}(y) = \mathbb{P}(\varepsilon_c \leq y - y_i^*) = F_c(y - y_i^*) \quad \text{and} \quad \mathbb{P}_{vi}(y) = \mathbb{P}(\varepsilon_v \leq y - y_i^*) = F_v(y - y_i^*). \quad (1)$$

This formulation assumes that decision noise enters at the level of the response variable—that is, for the response variable y , a person has an underlying value y_i^* that reflects their preferences, but their behavior is influenced by an additive distortion to y_i^* . In many analyses, it is assumed instead that decision noise enters at the level of utility—for instance, researchers often assume that decisions under risk are determined from expected utility preferences combined with an additive distortion to utility. In Appendix C, we demonstrate that most of the results below remain unchanged with that formulation.

We are interested in developing predictions for what combinations of individual-level decision probabilities $(\mathbb{P}_{ci}(y), \mathbb{P}_{vi}(y))$ are possible under stable preferences combined with various ancillary assumptions. Specifically, we focus on the following notion of attainability:

Definition 1. *Given a set of assumptions on the distribution over $(\varepsilon_c, \varepsilon_v)$, the attainable set $\mathcal{A}_i$ is the set of all $(a, b) \in [0, 1]^2$ such that $\mathbb{P}_{ci}(y) = F_c(y - y_i^*) = a$ and $\mathbb{P}_{vi}(y) = F_v(y - y_i^*) = b$ for some y_i^* and some distribution over $(\varepsilon_c, \varepsilon_v)$ that is consistent with the assumptions.*

In this framework, choices and valuations are both noisy measures of underlying preferences. The individual decision probabilities $\mathbb{P}_{ci}(y)$ and $\mathbb{P}_{vi}(y)$ depend on a combination of (i) how the offered y compares to the individual’s underlying preference y_i^* and (ii) how the person’s behavior is impacted by choice noise versus valuation noise. To build intuition, we consider a few examples.

First, note that if the choice noise and the valuation noise are identical (i.e., $F_c(x) = F_v(x)$ for all x), then

$$\mathbb{P}_{ci}(y) = F_c(y - y_i^*) = F_v(y - y_i^*) = \mathbb{P}_{vi}(y).$$

Hence:

Remark 1. *If noise is identical for choices and valuations, then $\mathcal{A}_i = \{(a, b) | a = b\}$.*

Thus, one way to justify the simple null of $\mathbb{P}_{ci}(y) = \mathbb{P}_{vi}(y)$ is to assume that noise is identical for choices and valuations. Without supporting evidence, however, this seems a very strong assumption, and indeed for our structural analysis in Section 4.5, if we were to assume stable preferences, we would conclude that choices are noisier than valuations. Once we permit differential noise for choices versus valuations, the attainable set can expand considerably. Indeed, Example 1 highlights that, if the only restriction on choice and valuation noise is that they both must be mean zero, any pattern is possible for $\mathbb{P}_{ci}(y)$ and $\mathbb{P}_{vi}(y)$.

Example 1: For any y and any $(a, b) \in (0, 1)^2$, one can find a y_i^* and mean-zero F_c and F_v such that $\mathbb{P}_{ci}(y) = F_c(y - y_i^*) = a$ and $\mathbb{P}_{vi}(y) = F_v(y - y_i^*) = b$. For instance, given a y , choose a $y_i^* < y$ and a $\delta > y - y_i^*$, and then assume

$$\varepsilon_c = \begin{cases} \delta & \text{with probability } 1 - a \\ \frac{-(1-a)\delta}{a} & \text{with probability } a \end{cases} \quad \varepsilon_v = \begin{cases} \delta & \text{with probability } 1 - b \\ \frac{-(1-b)\delta}{b} & \text{with probability } b. \end{cases}$$

By construction, a person with this y_i^* and these noise distributions would have $\mathbb{P}_{ci}(y) = a$ and $\mathbb{P}_{vi}(y) = b$.

Example 1 highlights that, if we want to have any hope of rejecting the null of stable preferences, we must impose restrictions on the nature of noise. One common restriction in applications is that noise is symmetric about its mean of zero.¹¹

Assumption 1 (symmetric noises). *The distributions of choice noise and valuation noise are both continuous and symmetric, i.e., F_c and F_v are both continuous with $F_c(x) = 1 - F_c(-x)$ and $F_v(x) = 1 - F_v(-x)$ for all x .*

Figure 1 illustrates what the attainable set $\mathcal{A}_i$ might look like under some known distributions for ε_c and ε_v that are consistent with symmetric noise. The curve in panel A depicts a generic $\mathcal{A}_i$ for fixed F_c and F_v that are symmetric and continuous with the same convex support. In general, once we fix an F_c , an F_v , and a y —so the

¹¹To simplify the exposition, whenever we assume a distribution is symmetric, we also assume it is continuous; however, except for a few minor details, continuity can be relaxed without affecting our results.

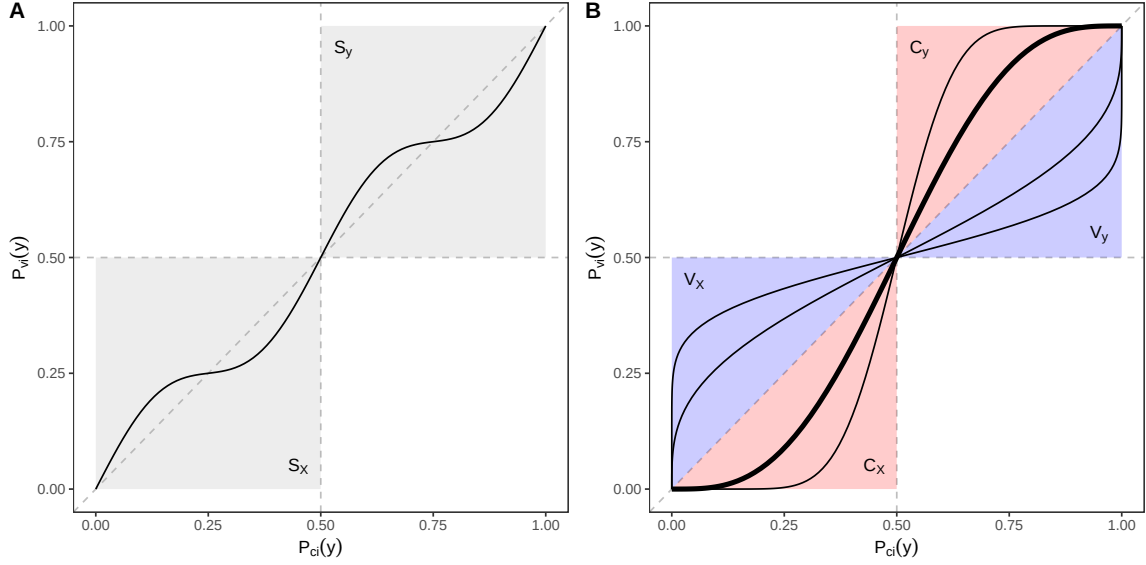


Figure 1: Individual Choice and Valuation Probabilities with Known Noise

Notes: Figure presents possible $(\mathbb{P}_{ci}(y), \mathbb{P}_{vi}(y))$ combinations given fixed choice and valuation noise distributions F_c and F_v . Panel A depicts possible combinations for generic F_c and F_v that are symmetric and continuous with the same convex support; any such curve must be contained in regions S_X and S_Y . Panel B depicts possible combinations when F_c and F_v are both mean-zero normal distributions with $k \equiv \frac{\sigma_c^2}{\sigma_v^2} \in \{4, 2, \frac{1}{2}, \frac{1}{4}\}$; bold line corresponds to $k = 2$. When $k > 1$ (choices noisier than valuations), curve must be contained in regions C_X and C_Y ; when $k < 1$ (valuations noisier than choices), curve must be contained in regions V_X and V_Y .

only unknown is where y_i^* falls in relation to that y —the set of possible $(\mathbb{P}_{ci}(y), \mathbb{P}_{vi}(y))$ will be a weakly increasing curve, where different points on the curve correspond to different y_i^* (i.e., a smaller y_i^* implies a larger $y - y_i^*$ and thus both $\mathbb{P}_{ci}(y)$ and $\mathbb{P}_{vi}(y)$ must be weakly larger). If F_c and F_v are symmetric, then this curve must pass through $(\frac{1}{2}, \frac{1}{2})$ and be symmetric around $(\frac{1}{2}, \frac{1}{2})$. Finally, if F_c and F_v have the same convex support, this curve must be strictly increasing from $(0, 0)$ to $(1, 1)$, as depicted by the curve in panel A.

Any weakly increasing curve that passes through $(\frac{1}{2}, \frac{1}{2})$ must be contained within the two shaded regions in panel A denoted S_y and S_X . Since these will appear in our results, we provide formal definitions:

$$S_y \equiv \left[\frac{1}{2}, 1\right]^2 \quad \text{and} \quad S_X \equiv \left[0, \frac{1}{2}\right]^2.$$

The labels reflect that, with symmetric noises, the outcome for individual i will be in

S_y when $y \succ_i X$ and in S_X when $X \succ_i y$.

Panel B of Figure 1 depicts the case where F_c and F_v are both known mean-zero normal distributions with variances σ_c^2 and σ_v^2 . The four curves reflect four different values of $k \equiv \sigma_c^2/\sigma_v^2$. When the ratio of the variances is larger than one, which means that choices are noisier than valuations, within S_y the curve must lie above the 45-degree line in the region labeled C_y , whereas within S_X the curve must lie below the 45-degree line in the region labeled C_X . Intuitively, when choices are noisier than valuations, $\mathbb{P}_{ci}(y)$ will be closer to $\frac{1}{2}$ (which reflects random choice), while $\mathbb{P}_{vi}(y)$ will be closer to 0 or 1 (which reflects implementing one's underlying preference). Analogously, when the ratio of the variances is smaller than one, which means that valuations are noisier than choices, within S_y the curve must lie below the 45-degree line in the region labeled V_y , whereas within S_X the curve must lie above the 45-degree line in the region labeled V_X . Finally, the further is the ratio of variances from one, the more the curve deviates from the 45-degree line.

Again, the curves in Figure 1 reflect the attainable set for known noise distributions. When analyzing experimental data, however, one won't know the specific noise distributions, and thus we want to identify the attainable set when we permit all possible noise distributions that satisfy a set of assumptions. From Figure 1, one can see that under symmetric normal-normal noise, if we have the flexibility to choose any σ_c^2 and σ_v^2 , we can produce a curve that goes through any point strictly in the interior of $S_y \cup S_X$. Lemma 1 establishes that under more general symmetric noise, the conclusion is much the same except we can generate outcomes along the edges as well.¹²

Lemma 1. *Under symmetric noises:*

1. $\mathcal{A}_i = S_y \cup S_X$;
2. If $y > y_i^*$, $(\mathbb{P}_{ci}(y), \mathbb{P}_{vi}(y)) \in S_y$, and if $y < y_i^*$, $(\mathbb{P}_{ci}(y), \mathbb{P}_{vi}(y)) \in S_X$; and
3. If $y = y_i^*$, then $(\mathbb{P}_{ci}(y), \mathbb{P}_{vi}(y)) = (\frac{1}{2}, \frac{1}{2})$.

The intuition for Lemma 1 is exactly as already expressed above. Whenever $y \geq y_i^*$ and thus $y \succeq_i X$ in terms of underlying preferences, symmetric noise implies $\mathbb{P}_{ci}(y) \geq \frac{1}{2}$ and $\mathbb{P}_{vi}(y) \geq \frac{1}{2}$ (because $\mathbb{P}_{ci}(y) = F_c(y - y_i^*) \geq F_c(0) = \frac{1}{2}$ and $\mathbb{P}_{vi}(y) =$

¹²The proof for Lemma 1 and all propositions are collected in Appendix B.

$F_v(y - y_i^*) \geq F_v(0) = \frac{1}{2}$). An analogous intuition holds whenever $y \leq y_i^*$ and thus $y \precsim_i X$ in terms of underlying preferences.¹³

The symmetric normal-normal case in Figure 1 also illustrates the impact of choices being noisier than valuations versus valuations being noisier than choices. We first formalize the regions in panel B of Figure 1:

$$C_y \equiv S_y \cap \{(a, b) \mid b \geq a\} \quad \text{and} \quad C_X \equiv S_X \cap \{(a, b) \mid b \leq a\},$$

and

$$V_y \equiv S_y \cap \{(a, b) \mid b \leq a\} \quad \text{and} \quad V_X \equiv S_X \cap \{(a, b) \mid b \geq a\}.$$

In the normal-normal case, when choices are noisier than valuations (i.e., $\sigma_c^2 > \sigma_v^2$), the outcome must lie in $C_y \cup C_X$, whereas when valuations are noisier than choices (i.e., $\sigma_v^2 > \sigma_c^2$), the outcome must lie in $V_y \cup V_X$. This conclusion generalizes to a more general definition of the noises being ordered (which we introduce later as Assumption 5). However, since one could easily imagine noises not being ordered or being ordered in different ways for different people, and since one cannot easily assess whether the noises are ordered without assuming additional structure, we do not yet make any assumptions on this dimension.

One might wonder whether additional restrictions on noise at the individual level could shrink $\mathcal{A}_i$ to something smaller than $S_y \cup S_X$. However, because the normal-normal case can attain any point in the interior of $S_y \cup S_X$, any such additional restriction would need to be something that excludes the normal-normal case. Since this seems unappealing from an applied perspective, we impose no further restrictions.

Lemma 1 characterizes the attainable set at the individual level under stable preferences and symmetric noises. If we had many repeated observations of choices and valuations at the individual level so that we could measure $\mathbb{P}_{ci}(y)$ and $\mathbb{P}_{vi}(y)$, we could assess the null of stable preferences combined with symmetric noises at the individual level. However, in experiments, we virtually always observe aggregate data from a heterogeneous population. Hence, before we can develop tests of the null of stable preferences, we need to build a model of aggregate behavior.

¹³Note that Assumption 1 is sufficient but not necessary for Lemma 1. All that is required is that F_c and F_v have the same median, in which case $y - y_i^*$ larger than that median must yield an outcome in S_y and $y - y_i^*$ smaller than that median must yield an outcome in S_X . Also note that part (3) of Lemma 1 is a detail that relies on continuity of F_c and F_v ; without continuity, $y = y_i^*$ implies $(\mathbb{P}_{ci}(y), \mathbb{P}_{vi}(y)) \in S_y$.

3.2 Aggregate Behavior

To specify a model of stochastic behavior at the aggregate level, we need to introduce heterogeneity in preferences and heterogeneity in decision noise. To do so, we use the following structure:

- Heterogeneity in preferences: Individual preferences are realizations of a random variable $y_i^* = \mu + \eta$, where $\mu \in \mathbb{R}$ is the population mean preference, and η is a mean-zero random variable with CDF H . Under Assumption 0, each individual gets a single draw of η (and thus a single y_i^*) that applies for both choices and valuations.
- Heterogeneity in decision noise: The additive noise shocks ε_c and ε_v can be correlated with η .

With this structure, the aggregate decision probabilities are:

$$\begin{aligned}\mathbb{P}_c(y) &= \mathbb{P}(y_i^* + \varepsilon_c \leq y) = \mathbb{P}(\mu + \eta + \varepsilon_c \leq y); \\ \mathbb{P}_v(y) &= \mathbb{P}(y_i^* + \varepsilon_v \leq y) = \mathbb{P}(\mu + \eta + \varepsilon_v \leq y).\end{aligned}$$

We are now interested in the attainable set $\mathcal{A}$ for the population proportions $(\mathbb{P}_c(y), \mathbb{P}_v(y))$, which we define analogously to the attainable set $\mathcal{A}_i$ for $(\mathbb{P}_{ci}(y), \mathbb{P}_{vi}(y))$.

Definition 2. *Given a set of assumptions on the distribution over $(\eta, \varepsilon_c, \varepsilon_v)$, the attainable set $\mathcal{A}$ is the set of all $(a, b) \in [0, 1]^2$ such that $\mathbb{P}_c(y) = a$ and $\mathbb{P}_v(y) = b$ for some μ and some distribution over $(\eta, \varepsilon_c, \varepsilon_v)$ that is consistent with the assumptions.*

We first note that if we make no assumptions beyond symmetry of each individual's noises, that is, symmetry of $\varepsilon_c|\eta$ and $\varepsilon_v|\eta$, then the attainable set at the population level is the convex hull of $S_y \cup S_X$.

Proposition 1. *Under stable preferences and (conditionally) symmetric noises, $\mathcal{A} = \text{conv}(S_y \cup S_X)$.*

Panel A in Figure 2 depicts the attainable set from Proposition 1. The intuition is straightforward. Lemma 1 establishes that for any point in $S_y \cup S_X$, an individual could exhibit those decision probabilities. If we then put no restrictions on the distribution of preferences and permit people with different y_i^* to have different noise distributions,

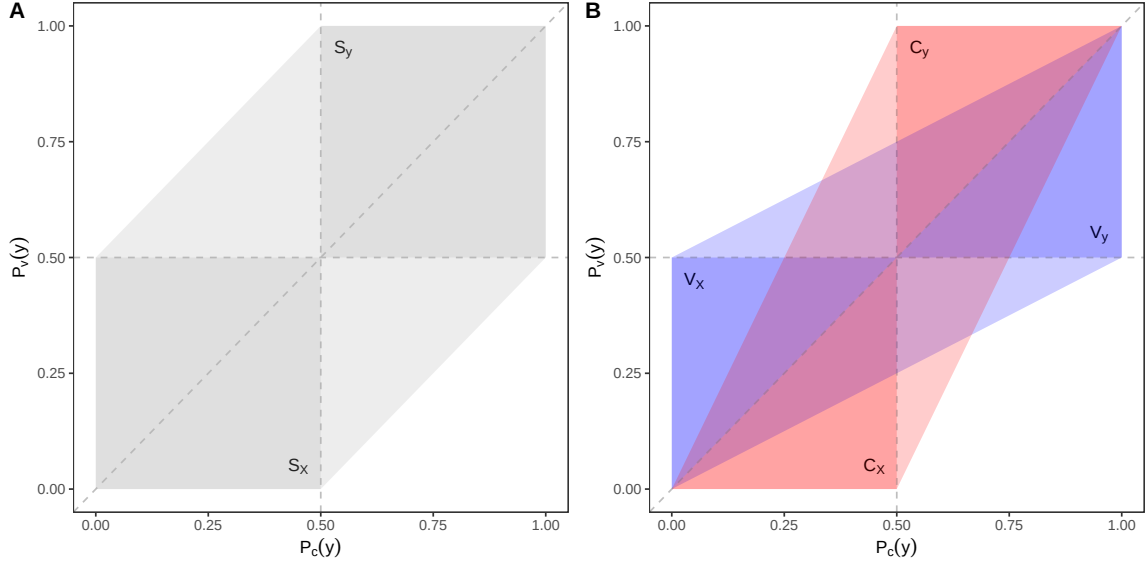


Figure 2: Attainable Sets from Propositions 1 and 2

Notes: Figure presents possible $(\mathbb{P}_c(y), \mathbb{P}_v(y))$ combinations given unknown choice and valuation noise that satisfy assumptions of Propositions 1 and 2. Panel A depicts possible combinations under unknown noise that is only symmetric (Assumption 1), with attainable set $\mathcal{A} = \text{conv}(S_y \cup S_x)$. Panel B depicts possible combinations under unknown noise that is both symmetric and independent of preferences (Assumptions 1 and 2), with attainable set $\mathcal{A} = \text{conv}(C_y \cup C_x) \cup \text{conv}(V_y \cup V_x)$.

we can create a population that consists of any distribution across points in $S_y \cup S_x$. It immediately follows that we can get average behavior anywhere in $\text{conv}(S_y \cup S_x)$.

Proposition 1 suggests a possible test of the null of stable preferences: If we want to assume very little—specifically, only that everyone has symmetric noises—then to reject the null of stable preferences, we would need to reject that an experiment lies in $\text{conv}(S_y \cup S_x)$. As we will see, nearly the entirety of the preference-reversals literature lies within this set. Thus, one might be interested in imposing additional assumptions to gain restrictions on the set of attainable outcomes.

A natural next step to restrict the attainable set is to assume that preferences and decision noise are independent from each other:

Assumption 2 (preferences independent of noise). $(\varepsilon_c, \varepsilon_v)$ is independent of η , and thus all individuals have the same CDF F_c for ε_c , and all individuals have the same CDF F_v for ε_v .

Under Assumption 2, for any fixed F_c and F_v that are symmetric and the same for everyone, the set of $(\mathbb{P}_{ci}(y), \mathbb{P}_{vi}(y))$ that could emerge across different y_i^* must be a

weakly increasing curve that is symmetric around $(\frac{1}{2}, \frac{1}{2})$ (as in Figure 1). Without any restrictions on the distribution of η , we could then attain any $(\mathbb{P}_c(y), \mathbb{P}_v(y))$ in the convex hull of that curve. Because we can choose any symmetric F_c and F_v that we want, it follows that a point is in the attainable set $\mathcal{A}$ if it is part of the convex hull of any weakly increasing curve that is symmetric around $(\frac{1}{2}, \frac{1}{2})$. Applying this insight yields Proposition 2.

Proposition 2. *Under stable preferences and symmetric noises, if preferences are independent of noise, then $\mathcal{A} = \text{conv}(C_y \cup C_X) \cup \text{conv}(V_y \cup V_X)$.*

To provide some intuition for Proposition 2, consider the restriction of the attainable set relative to Proposition 1. Specific values of $(\mathbb{P}_c(y), \mathbb{P}_v(y))$ are eliminated; these values correspond to combinations of correlated types (e.g., combining individuals exhibiting $y_i^* > y$ and valuations noisier than choices with those exhibiting $y_i^* < y$ and choices noisier than valuations). With preferences independent of noise, $(\mathbb{P}_c(y), \mathbb{P}_v(y))$ values attainable by relying on such correlation are no longer permissible.

Proposition 2 suggests another possible test of the null of stable preferences: If we are willing to assume that everyone has symmetric noises and that preferences and noise are independent, then to reject the null of stable preferences we would need to reject that an experiment lies in $\text{conv}(C_y \cup C_X) \cup \text{conv}(V_y \cup V_X)$. This region is illustrated in panel B of Figure 2. Notably, whether preferences are required to be independent of noise makes relatively little difference to the set of attainable outcomes, such that many studies (as we will see) will lie within the attainable regions for both Proposition 1 and 2.

Hence, further assumptions are required to more meaningfully restrict the set of attainable outcomes. Propositions 1 and 2 rely on the freedom to use any distribution of preferences. We next restrict preferences to be symmetric.

Assumption 3 (symmetric preferences). *The distribution of preferences is continuous and symmetric, i.e., H is continuous with $H(x) = 1 - H(-x)$ for all x .*

Under Assumptions 1, 2, and 3, it is natural to define the random variables y_c and y_v as

$$y_c \equiv y_i^* + \varepsilon_c = \mu + \eta + \varepsilon_c = \mu + \nu_c$$

and

$$y_v \equiv y_i^* + \varepsilon_v = \mu + \eta + \varepsilon_v = \mu + \nu_v$$

where $\nu_c = \eta + \varepsilon_c$ and $\nu_v = \eta + \varepsilon_v$. Note that ν_c and ν_v are both sums of independent symmetric mean-zero random variables and thus are themselves symmetric mean-zero random variables. Letting G_c and G_v denote the CDFs for ν_c and ν_v , we can write the aggregate decision probabilities as:

$$\mathbb{P}_c(y) = \mathbb{P}(y \geq \mu + \nu_c) = G_c(y - \mu) \quad \text{and} \quad \mathbb{P}_v(y) = \mathbb{P}(y \geq \mu + \nu_v) = G_v(y - \mu). \quad (2)$$

Comparing (1) and (2), we have:

Remark 2. *In terms of the attainable set, an aggregate-level model with symmetric noise and symmetric preferences with mean μ that are independent of noise is observationally equivalent to an individual-level model with $y_i^* = \mu$ and symmetric mean-zero noise.*

Remark 2 immediately leads to Proposition 3, which is the analogue of Lemma 1:

Proposition 3. *Under stable preferences and symmetric noises, if preferences are symmetric and independent of noise, then:*

1. $\mathcal{A} = S_y \cup S_X$;
2. If $y > \mu$, $(\mathbb{P}_c(y), \mathbb{P}_v(y)) \in S_y$, and if $y < \mu$, $(\mathbb{P}_c(y), \mathbb{P}_v(y)) \in S_X$; and
3. If $y = \mu$, then $(\mathbb{P}_c(y), \mathbb{P}_v(y)) = (\frac{1}{2}, \frac{1}{2})$.

Part 1 of Proposition 3 provides another test of the null of stable preferences when comparing aggregate shares in an experiment. If we are willing to make the ancillary assumptions that noise is symmetric and that preferences are symmetric and independent of noise, then we can test the null of stable preferences by assessing whether an experiment lies outside of $S_y \cup S_X$. Stated differently, Proposition 3(1) requires congruence between the signs of $\mathbb{P}_c(y) - \frac{1}{2}$ and $\mathbb{P}_v(y) - \frac{1}{2}$.¹⁴

While Proposition 3 permits $\mathbb{P}_c(y) = \mathbb{P}_v(y)$, it does not provide conditions under which we would predict $\mathbb{P}_c(y) = \mathbb{P}_v(y)$, which would then reflect conditions under which standard tests are valid. It turns out that, under two further assumptions, we would predict $\mathbb{P}_c(y) = \mathbb{P}_v(y)$ only in one special case.

¹⁴Analogous to part (3) of Lemma 1, part (3) or Proposition 3 is a detail that relies on continuity of F_c and F_v ; without continuity, $y = \mu$ implies $(\mathbb{P}_c(y), \mathbb{P}_v(y)) \in S_y$.

Assumption 4 (unimodal preferences). *H has a strictly unimodal density h on a convex support S . That is, for any $x_1 > x_2 \geq 0$, $h(x_1) \leq h(x_2)$ and $h(-x_1) \leq h(-x_2)$, where both inequalities are strict when $x_1 \in S$.*

Assumption 5 (ordered noise). *At least one of the following holds:*

1. *Choices are noisier than valuations: $F_v(z) < F_c(z)$ whenever $z < 0$ and $F_v(z) > 0$, and $F_v(z) > F_c(z)$ whenever $z > 0$ and $F_v(z) < 1$.*
2. *Valuations are noisier than choices: $F_c(z) < F_v(z)$ whenever $z < 0$ and $F_c(z) > 0$, and $F_c(z) > F_v(z)$ whenever $z > 0$ and $F_c(z) < 1$.*

The assumption of unimodal preferences is strong, but it is commonly used (e.g., if preferences are normally distributed). The assumption of ordered noise generalizes what we saw for the normal-normal case in Figure 1, where σ_c^2/σ_v^2 greater or less than one determined whether choices or valuations are noisier.

Proposition 4. *Under stable preferences and symmetric and ordered noises, if preferences are symmetric, unimodal, and independent of noise, then $\mathbb{P}_c(y) = \mathbb{P}_v(y) \in (0, 1)$ if and only if $y = \mu$, in which case $\mathbb{P}_c(y) = \mathbb{P}_v(y) = \frac{1}{2}$.*

Proposition 4 highlights a different problem with the typical approach of testing whether the aggregate shares are the same. Up to now, we have emphasized that a simple test of the null $\mathbb{P}_c(y) = \mathbb{P}_v(y)$ is not enough to test the null of stable preferences unless one is willing to make very strong ancillary assumptions (e.g., that the choice noise and the valuation noise are identical). In Proposition 4, we see a second problem: Under the types of regularity assumptions frequently made in economics—symmetric continuous noise and symmetric unimodal preferences—stable preferences combined with differential noise across elicitation techniques would in fact yield $\mathbb{P}_c(y) = \mathbb{P}_v(y)$ only if we happen to be conducting the experiment at $y = \mu$. Since the population mean preference is typically something one would not know when designing an experiment, a strong interpretation is that, in fact, under stable preferences, we should not expect to see $\mathbb{P}_c(y) = \mathbb{P}_v(y)$ in any experiments.

3.3 Leveraging Richer Data for Additional Tests

In line with the prior literatures on preference reversals and on comparing choices versus valuations, Section 3.2 focused on predictions for $\mathbb{P}_c(y)$ and $\mathbb{P}_v(y)$. However, the null of

stable preferences combined with various ancillary assumptions also makes qualitative predictions for other objects. In this section, we highlight two such predictions, where each takes advantage of richer data structures (if they are available).

3.3.1 Using Quantitative Information on Mean Preferences from Valuations

Part 2 of Proposition 3 suggests another test of stable preferences (combined with Assumptions 1–3): If $y > \mu$, then $\mathbb{P}_c(y) \geq \frac{1}{2}$, and if $y < \mu$, then $\mathbb{P}_c(y) \leq \frac{1}{2}$. Testing this prediction requires knowledge of μ . While we cannot observe μ , if a dataset contains the full distribution of y_{vi} in the valuation task, we can obtain an estimate of μ . We can then evaluate Proposition 3(2) using that empirical analog of μ . We describe one such implementation in Section 4.3.

3.3.2 Analyzing Conditional Choices

If one has linked data at the individual level—that is, for each individual we observe both their choice between y and X and their stated valuation y_{vi} for X —additional tests become possible.

Consider the following natural conjecture: Because it is true under symmetric noise that

$$\mathbb{P}_c(y|y = y_i^*) = \frac{1}{2}$$

and that $\mathbb{E}[y_{vi}] = y_i^*$, one might conjecture that¹⁵

$$\mathbb{P}_c(y|y = y_{vi}) = \frac{1}{2}.$$

It turns out that this property is unlikely to hold even in the types of symmetric, smooth, and independent environments often used in economics. In particular, suppose we assume symmetric noises and stable preferences that are symmetric, unimodal, and independent of noise. In addition, assume:

Assumption 6 (independent noise). ε_c and ε_v are independent.

¹⁵While the conditioning event is null, this is formally a regular conditional probability distribution, which is uniquely defined up to a null set. Also note that we can observe $\mathbb{P}_c(y|y = y_{vi})$ only if $y_{vi} - \mu$ is in the support of $\eta + \varepsilon_v$; hence, everything below makes this assumption.

Assumption 7 (positive densities). ε_c and ε_v have densities f_c and f_v , respectively, that are positive on an open interval containing zero.

The following proposition establishes that, under these conditions, this conjecture holds only if y is equal to the mean preference.

Proposition 5. *Under stable preferences and symmetric noises, if preferences are symmetric, unimodal, and independent of noise, and the noise is independent with positive densities, then $\mathbb{P}_c(y \mid y = y_{vi}) = \frac{1}{2}$ if and only if $y = \mu$. Moreover, if $y > \mu$ then $\mathbb{P}_c(y \mid y = y_{vi}) > \frac{1}{2}$, and if $y < \mu$ then $\mathbb{P}_c(y \mid y = y_{vi}) < \frac{1}{2}$.*

The intuition for Proposition 5 is easy to see when the noise is also unimodal. Consider an experiment with $y > \mu$. When we observe that a person reports a valuation $y_{vi} = y$, it need not be the case that $\mathbb{E}[y_i^* \mid y_{vi} = y] = y$; indeed, with symmetric and unimodal preferences and symmetric and unimodal noise, it must be that $\mathbb{E}[y_i^* \mid y_{vi} = y]$ is in between μ and y . Hence, for this person, y will be relatively attractive, and thus it leads to a conditional choice probability greater than $\frac{1}{2}$. An analogous argument holds for an experiment with $y < \mu$, leading to a conditional choice probability less than $\frac{1}{2}$.

On one hand, this result is a point of caution. It is tempting to think that one way to test for stable preferences is to test whether $\mathbb{P}(y \mid y_{vi} = y) = \frac{1}{2}$. Analogous to our discussion after Proposition 4, however, such a test is valid only if we are confident that we are conducting an experiment precisely at $y = \mu$.

On the other hand, Proposition 5 suggests another test of stable preferences (combined with Assumptions 1-4 and 6-7): If $y > \mu$ then $\mathbb{P}_c(y \mid y_{vi} = y) > \frac{1}{2}$, and if $y < \mu$ then $\mathbb{P}_c(y \mid y_{vi} = y) < \frac{1}{2}$. As for the prediction in Section 3.3.1, testing this prediction requires knowledge of μ . Again, if a linked dataset contains the full distribution of y_{vi} in the valuation task, we can obtain an estimate of μ and conduct the exercise with the empirical analog of μ . We describe one such implementation in Section 4.4.

3.4 Tests Based on Estimates of Structural Parameters

Up to now, our assumptions have involved qualitative features of the distributions of noise and preferences and have led to assessments of stability free from precise distributional assumptions. If one is willing to make structural assumptions about the

distributions, and if one has the appropriate variation in one's data, one can develop tests based on estimates of structural parameters. Moreover, with such estimates, one can also obtain quantitative estimates of the magnitudes of instability. In this section, we present an example of such a structural approach in which we assume all distributions are normal.

3.4.1 Underlying Structural Assumptions

Suppose that the population distribution of underlying preferences is given by

$$\begin{pmatrix} y_{ci}^* \\ y_{vi}^* \end{pmatrix} \sim N \left(\begin{pmatrix} \mu_c \\ \mu_v \end{pmatrix}, \begin{pmatrix} \gamma_c^2 & \gamma_{cv} \\ \gamma_{cv} & \gamma_v^2 \end{pmatrix} \right).$$

With this structure, the null of stable preferences would mean that $\mu_c = \mu_v$ and $\gamma_c^2 = \gamma_v^2 = \gamma_{cv}$ (which together imply that every individual has $y_{ci}^* = y_{vi}^*$). Because the approaches below do not provide direct estimates for γ_c^2 and γ_v^2 , we focus our analysis on testing the null of $\mu_c = \mu_v$, that is, of stable mean preferences.

We assume that noise is independent of preferences, and we further assume that ε_c and ε_v both have mean-zero normal distributions with variances σ_c^2 and σ_v^2 , respectively. Finally, we assume the two types of noise are independent of each other.¹⁶

Recall that a person chooses y over X when y is larger than $y_{ci} \equiv y_{ci}^* + \varepsilon_c$, and the person reports valuation for X of $y_{vi} \equiv y_{vi}^* + \varepsilon_v$. Under the assumptions above, the joint distribution of y_{ci} and y_{vi} is bivariate normal:

$$\begin{pmatrix} y_{ci} \\ y_{vi} \end{pmatrix} \sim N \left(\begin{pmatrix} \mu_c \\ \mu_v \end{pmatrix}, \begin{pmatrix} \sigma_c^2 + \gamma_c^2 & \gamma_{cv} \\ \gamma_{cv} & \sigma_v^2 + \gamma_v^2 \end{pmatrix} \right). \quad (3)$$

Equation (3) can be used to generate structural predictions from which one can base additional tests of stable preferences and quantification of deviations therefrom.

3.4.2 Probit Formulation

If one collects within-subjects responses for both choices and valuations, this framework generates a prediction for $\mathbb{P}_c(y|y_{vi})$. In particular, because y_{ci} and y_{vi} are joint normal,

¹⁶Note that this structure satisfies all of Assumptions 1-7, unless $\sigma_c^2 = \sigma_v^2$, in which case ordered noise only weakly holds.

the conditional distribution of y_{ci} given a specific valuation of y_{vi} is normal.¹⁷ From that conditional distribution one can derive that

$$\mathbb{P}_c(y|y_{vi}) = \Phi \left(\frac{\mu_v - \mu_c}{V} + \frac{\left(1 - \frac{\gamma_{cv}}{\sigma_v^2 + \gamma_v^2}\right)}{V}(y - \mu_v) + \frac{\frac{\gamma_{cv}}{\sigma_v^2 + \gamma_v^2}}{V}(y - y_{vi}) \right)$$

where $V^2 \equiv \sigma_c^2 + \gamma_c^2 - \frac{(\gamma_{cv})^2}{\sigma_v^2 + \gamma_v^2}$ and Φ is the CDF for a standard normal distribution. Replacing $y - \mu_v$ with its sample analog from valuations, $y - \overline{y_{vi}}$, yields a simple probit regression:

$$\mathbb{P}_c(y|y_{vi}) = \Phi(\alpha + \beta_1(y - \overline{y_{vi}}) + \beta_2(y - y_{vi})), \quad (4)$$

where $\alpha = \frac{\mu_v - \mu_c}{V}$, $\beta_1 = \frac{\left(1 - \frac{\gamma_{cv}}{\sigma_v^2 + \gamma_v^2}\right)}{V}$, $\beta_2 = \frac{\frac{\gamma_{cv}}{\sigma_v^2 + \gamma_v^2}}{V}$, and $\beta_1 + \beta_2 = \frac{1}{V}$. Hence, if one has “groupings” of experiments (i.e., multiple values of y for a fixed X such that there is variation in $y - \overline{y_{vi}}$), one can obtain estimates for α , β_1 , and β_2 .¹⁸

With these estimates in hand, one can test the null of stable mean preferences. Specifically, an estimate for $\mu_v - \mu_c$ can be derived from combining regression coefficients,

$$\widehat{\mu_v - \mu_c} = \frac{\hat{\alpha}}{\hat{\beta}_1 + \hat{\beta}_2},$$

with standard error derived from the delta method. We can then conduct a formal test of the null of $\mu_v - \mu_c = 0$.

The probit estimates can also be used to quantify the behavioral impact of deviations from stable preferences. The quantity $\mathbb{P}_c(\mu_v|y_{vi} = \mu_v)$ is the choice probability in the idealized experiment where we are able to select $y = \mu_v$ and in addition focus on participants who state a valuation equal to that μ_v . Under joint normality and stable preferences, this quantity would be $\frac{1}{2}$. Within our probit structure, $\Phi(\hat{\alpha})$ represents an estimate of $\mathbb{P}_c(\mu_v|y_{vi} = \mu_v)$.

3.4.3 MLE Formulation

The probit estimator of equation (4) relies on replacing $y - \mu_v$ with its sample analog $y - \overline{y_{vi}}$. This introduces the possible challenge of measurement error biasing corresponding estimates. Attenuation bias could naturally lead to measures of α

¹⁷Specifically, $y_{ci}|y_{vi} \sim N\left(\mu_c + \frac{\gamma_{cv}}{\sigma_v^2 + \gamma_v^2}(y_{vi} - \mu_v), V^2\right)$, where $V^2 \equiv \sigma_c^2 + \gamma_c^2 - \frac{(\gamma_{cv})^2}{\sigma_v^2 + \gamma_v^2}$.

¹⁸The key challenge is obtaining enough variation in $y - \overline{y_{vi}}$ to separately identify α and β_1 since $y - \overline{y_{vi}}$ will vary only across experiments through variation in y .

closer to zero, working against finding evidence of preference instability. A natural alternative is to estimate mean preferences alongside the other variables of interest by pursuing a maximum likelihood approach.

For such an approach, one again needs within-subjects responses for both choices and valuations. For an MLE approach, it is also natural to account for the interval nature of most valuations data. In particular, imagine a person provides a valuation in the region (y_L, y_H) (e.g., the interval that corresponds to the switching point in a MPL eliciting y_{vi}), and that when offered a choice between X and y , they choose y . The likelihood of such an observation is the probability of (y_{ci}, y_{vi}) falling within the rectangle $(-\infty, y) \times (y_L, y_H)$. Under the normal-normal-normal specification here, it is straightforward to calculate this probability given the joint normal distribution in equation (3). Maximizing the log likelihood of observations with respect to the parameters μ_c , μ_v , $(\sigma_c^2 + \gamma_c^2)$, $(\sigma_v^2 + \gamma_v^2)$, and γ_{cv} yields estimates of the corresponding parameters using conventional techniques. See Appendix E for details.

Because the MLE approach yields estimates for both $\hat{\mu}_c$ and $\hat{\mu}_v$, we can directly test the null of $\mu_v - \mu_c = 0$. In addition, under the MLE approach, we obtain an estimate of $\mathbb{P}_c(\mu_v | y_{vi} = \mu_v)$, which again provides a sense of the behavioral impact of deviations from stable preferences, using $\Phi\left(\frac{\hat{\mu}_v - \hat{\mu}_c}{\hat{V}}\right)$.¹⁹

4 Empirical Application

Our theoretical analysis in Section 3 provides multiple ways to test the null of stable preferences; we summarize these tests in Table 1. Note that each test is in fact a joint test of stable preferences combined with some combination of ancillary assumptions. In this section, we conduct an empirical implementation of these tests.

4.1 Dataset Construction

We begin by constructing two datasets. The first is constructed using data from the prior literature on preference reversals and the prior literature on choices-versus-MPLs. The second is constructed using data from two recent studies by McGranaghan et al. (2024, 2026) that are well-suited for the present paper.

Data from Prior Literature: To construct a dataset based on the preference-

¹⁹ $\hat{V}$ is the estimate for V , the conditional standard deviation for $y_{ci} | y_{vi}$; see footnote 17.

Table 1: Summary of Tests of Stable Preferences

Result	Assumptions	Prediction to Test
Panel A: Simple Qualitative Tests		
Remark 1	Stable preferences plus identical noise	$\mathbb{P}_c(y) = \mathbb{P}_v(y)$
Proposition 1	Stable preferences plus symmetric noise	$(\mathbb{P}_c(y), \mathbb{P}_v(y)) \in \text{conv}(S_y \cup S_X)$
Proposition 2	Stable preferences plus symmetric noise that is independent of preferences	$(\mathbb{P}_c(y), \mathbb{P}_v(y)) \in \text{conv}(C_y \cup C_X) \cup \text{conv}(V_y \cup V_X)$
Proposition 3(1)	Stable preferences plus symmetric preferences and symmetric noise that is independent of preferences	$(\mathbb{P}_c(y), \mathbb{P}_v(y)) \in S_y \cup S_X$
Panel B: Qualitative Tests Using Valuations to Estimate μ		
Proposition 3(2)	Stable preferences plus symmetric preferences and symmetric noise that is independent of preferences	If $y > \mu$, $\mathbb{P}_c(y) > \frac{1}{2}$; if $y < \mu$, $\mathbb{P}_c(y) < \frac{1}{2}$
Proposition 5	Stable preferences plus symmetric unimodal preferences and symmetric independent noise with positive densities that is independent of preferences	If $y > \mu$, $\mathbb{P}_c(y y_{vi} = y) > \frac{1}{2}$; if $y < \mu$, $\mathbb{P}_c(y y_{vi} = y) < \frac{1}{2}$
Panel C: Structural Tests Using Linked Choices and Valuations		
$\begin{pmatrix} y_{ci}^* \\ y_{vi}^* \end{pmatrix} \sim N \left(\begin{pmatrix} \mu_c \\ \mu_v \end{pmatrix}, \begin{pmatrix} \gamma_c^2 & \gamma_{cv} \\ \gamma_{cv} & \gamma_v^2 \end{pmatrix} \right),$ $\varepsilon_c \sim N(0, \sigma_c^2), \text{ and } \varepsilon_v \sim N(0, \sigma_v^2), \text{ all three independent}$		Structural (Probit or MLE) estimates for evaluation of $\mu_c = \mu_v$

reversal literature, we draw from a recent meta-analysis by [Lu \(2026\)](#). In Table 2 from Lu’s Online Supplementary Material, there are 462 experiments from 54 studies for which one could calculate the empirical quantities $(\widehat{\mathbb{P}}_c(y), \widehat{\mathbb{P}}_v(y))$. A research assistant attempted to collect the 54 original studies to verify (and, in very few instances, adjust) the numbers. The research assistant was able to confirm the numbers for 314 experiments from 42 studies, and that is the data that we use here. See Appendix [D.1](#) for details.²⁰

To construct data based on the choices-versus-MPLs literature, we pull the needed data ourselves from the three key studies in this literature ([Brown and Healy, 2018](#);

²⁰A number of the original studies, including the seminal work of [Grether and Plott \(1979\)](#), report only combined numbers across multiple conditions—i.e., across multiple “experiments”. Rather than exclude these, we chose to consider the combined numbers to be one experiment. See Appendix [D.1](#) for more details.

Freeman et al., 2019; Freeman and Mayraz, 2019). Specifically, we identify nine experiments in these studies; see Appendix D.2 for details.

On net, then, our prior-literature dataset consists of 323 experiments. For each experiment, we identify the safer and riskier lotteries, and we harmonize the data such that $\widehat{\mathbb{P}}_c(y)$ corresponds to the proportion of subjects choosing the safer alternative and $\widehat{\mathbb{P}}_v(y)$ corresponds to the proportion of subjects stating a higher valuation for the safer alternative (in the case of a preference-reversal experiment) or having a valuation that would indicate a preference for the safer alternative (in the case of a choices-versus-MPLS experiment). Hence, for each experiment, the data include $\widehat{\mathbb{P}}_c(y)$, $\widehat{\mathbb{P}}_v(y)$, and the number of participants for each. Across all 323 experiments, valuations are provided by an average of 101 subjects and choices are provided by an average of 96 subjects, yielding 32,695 total valuations and 30,949 total choices.

Data from McGranaghan et al. (2024, 2026): These studies collect choices and valuations from the same subjects. For each study, the goal was not to compare choices versus valuations; rather, their focus was comparing pairs of valuations to study the phenomena of the common ratio effect, the common consequence effect, and preferences for probabilistic mixtures. They collected data on corresponding pairs of choices to illustrate an inference problem that arises when studying paired choice tasks (in McGranaghan et al. (2024)) and to show that paired choices and paired valuations can yield similar conclusions in some instances (in McGranaghan et al. (2026)). However, their data provide a convenience dataset that we can use for the purposes of this paper.

In McGranaghan et al. (2024), there are 160 experiments—that is, 160 cases where a set of subjects make a binary choice and also provide a valuation linked to that binary choice. In McGranaghan et al. (2026), there are another 264 experiments. For each of these 424 experiments, we identify the safer and riskier lotteries, and we harmonize the data such that $\widehat{\mathbb{P}}_c(y)$ corresponds to the proportion of subjects choosing the safer alternative and $\widehat{\mathbb{P}}_v(y)$ corresponds to the proportion of subjects having a valuation that would indicate a preference for the safer alternative. Hence, for each experiment, the data include $\widehat{\mathbb{P}}_c(y)$, $\widehat{\mathbb{P}}_v(y)$, and the number of participants. Across all 424 experiments, there are 68,448 observations for both choices and valuations, and thus an average of 161 subjects per experiment (again, each experiment is within-subject).

These two studies have two additional advantages. First, the data contain each individual valuation, and thus we can calculate the average valuation (i.e., $\overline{y_{vi}}$) as

a proxy for mean preferences for different samples. Second, the experiments come in precisely the “groupings” needed to conduct the structural tests from Section 3.4. Specifically, a grouping consists of a set of experiments that all use the same valuation task for a specific lottery X , but which differ in y .²¹ In McGranaghan et al. (2024), there are 40 different groupings (40 different X ’s and corresponding valuation tasks), where each grouping contains four different experiments (4 different values for y). In McGranaghan et al. (2026), there are 44 different groupings, where each grouping contains six different experiments. On net, there are 84 groupings, with an average of 815 subjects in each grouping (i.e., completing each of the 84 valuation tasks across the two studies).

There are some differences across the 424 experiments and across the 84 groupings—see Appendix D.3 for more details. While most of those details are unimportant for our analysis in this paper, we do highlight one detail that turns out to be important. Across the two studies, subjects are asked to make three kinds of comparisons: (i) a comparison between a certain amount and a binary lottery (which they label an AB decision), (ii) a comparison between a certain amount and a trinary lottery (an AB' decision), and (iii) a comparison between two binary lotteries (a CD decision).²²

4.2 The Standard Test from the Prior Literature

Using our collected datasets, we first conduct the standard test used in the prior literature on preference reversals and the prior literature on choices-vs-MPLs—that is, we test the null of $\mathbb{P}_c(y) = \mathbb{P}_v(y)$.

In Figure 3, we plot for each experiment the observed values of $\widehat{\mathbb{P}}_c(y)$ and $\widehat{\mathbb{P}}_v(y)$ with reference to the 45-degree line of equality. In panel A, we present the data for the prior literature; in panel B, we present the data from McGranaghan et al. (2024, 2026). Within the prior literature, 232 of 323 experiments (72%) exhibit $\widehat{\mathbb{P}}_c(y) > \widehat{\mathbb{P}}_v(y)$, and within the McGranaghan et al. (2024, 2026) data 341 of 424 experiments (80%) exhibit $\widehat{\mathbb{P}}_c(y) > \widehat{\mathbb{P}}_v(y)$. Hence, both datasets are consistent with the received wisdom that people are more likely to choose a safer option than valuations would imply.

²¹This variation in y for a fixed valuation task yields variation in $y - \overline{y_{vi}}$, which is key to identification in our structural approaches (see footnote 18).

²²For certain-vs.-binary comparisons, there are 64 experiments (on average, 403 subjects per experiment) across 14 groupings (on average, 1844 subjects per grouping). For certain-vs.-trinary comparisons, there are 120 experiments (140 subjects per experiment) across 20 groupings (840 subjects per grouping). And for binary-vs.-binary comparisons, there are 240 experiments (107

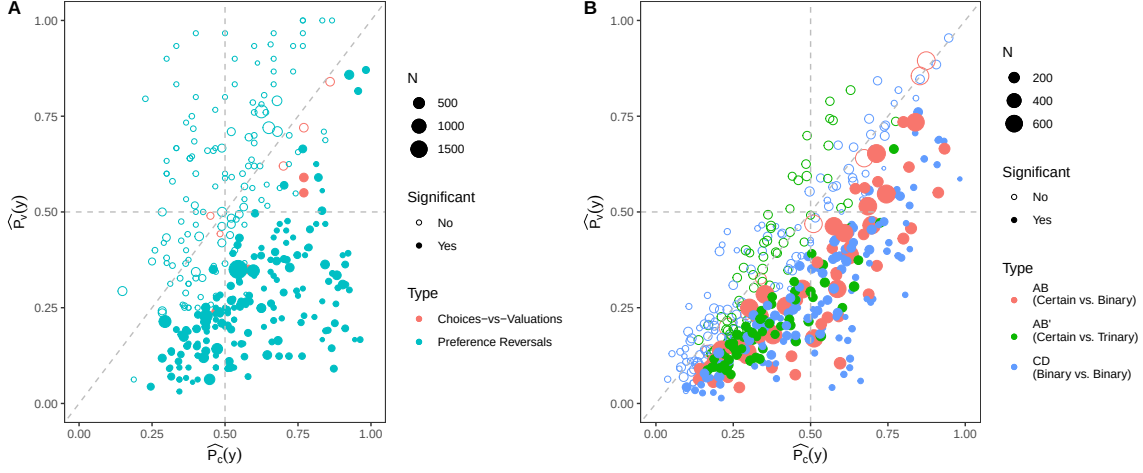


Figure 3: Applying the Standard Test of Stable Preferences

Notes: Figure presents observed aggregate choice and valuation proportions, $\widehat{\mathbb{P}}_c(y)$ and $\widehat{\mathbb{P}}_v(y)$, for 323 experiments from prior literature (panel A) and 424 experiments from [McGranaghan et al. \(2024, 2026\)](#) (panel B). Data harmonized such that greater probability corresponds to greater likelihood of preferring safer alternative. Each circle reflects one experiment, size corresponds to experimental sample size. Shading reflects application of standard test of stable preferences: solid circles correspond to experiments for which we reject the null of $\mathbb{P}_c(y) = \mathbb{P}_v(y)$ at the 5% level in a one-sided test.

Of course, merely observing $\widehat{\mathbb{P}}_c(y) \neq \widehat{\mathbb{P}}_v(y)$ is not sufficient to statistically reject the null of $\mathbb{P}_c(y) = \mathbb{P}_v(y)$ because even under the null we could observe different proportions due to sampling variation. To assess statistical significance, we use a standard test of proportions under the assumption of independence. Specifically, for each experiment, we calculate the z -score

$$z = \frac{\widehat{\mathbb{P}}_c(y) - \widehat{\mathbb{P}}_v(y)}{\sqrt{\frac{\widehat{\mathbb{P}}_c(y)(1-\widehat{\mathbb{P}}_c(y))}{N_c} + \frac{\widehat{\mathbb{P}}_v(y)(1-\widehat{\mathbb{P}}_v(y))}{N_v}}},$$

where N_c and N_v refer to the number of subjects completing choices and valuations, respectively. If we think a two-sided test is appropriate, we reject the null of $\mathbb{P}_c(y) = \mathbb{P}_v(y)$ if $|z| > 1.96$ (i.e., we impose a two-sided threshold of 0.05). Under a two-sided test, 67% of prior experiments and 55% of [McGranaghan et al. \(2024, 2026\)](#) experiments reject the null hypothesis of $\mathbb{P}_c(y) = \mathbb{P}_v(y)$. Given the received wisdom from the prior literature that people are more risk averse for choices than for valuations, we might instead reject the null of $\mathbb{P}_c(y) = \mathbb{P}_v(y)$ in favor of $\mathbb{P}_c(y) > \mathbb{P}_v(y)$ if $z > 1.645$

subjects per experiment) across 50 groupings (516 subjects per grouping).

(i.e., we impose a one-sided threshold of 0.05). Under this one-sided test, 56% of prior experiments and 56% of [McGranaghan et al. \(2024, 2026\)](#) experiments reject the null hypothesis of $\mathbb{P}_c(y) = \mathbb{P}_v(y)$. Figure 3 illustrates significance under the latter, one-sided test as shaded points.²³

Hence, if one uses the standard approach of testing the null of $\mathbb{P}_c(y) = \mathbb{P}_v(y)$, one indeed finds frequent rejections of the null in both datasets, and support for the received wisdom that people are more risk averse for choices than for valuations. As Table 1 highlights, however, rejecting the null of $\mathbb{P}_c(y) = \mathbb{P}_v(y)$ could reflect a failure of stable preferences, but it also could reflect a failure of identical noise. Because identical noise is a strong assumption to make without supporting evidence, we now turn to our new tests based on weaker assumptions regarding noise.

4.3 Statistical Tests Based on Weaker Assumptions

Propositions 1, 2, and 3(1) provide the basis for tests of stable preferences based on weaker assumptions, as summarized in Table 1. Each test relies on assessing whether observed values for $\widehat{\mathbb{P}}_c(y)$ and $\widehat{\mathbb{P}}_v(y)$ could come from some null region of attainable $(\mathbb{P}_c(y), \mathbb{P}_v(y))$ combinations derived in each proposition. In Figure 4, we again plot observed values of $\widehat{\mathbb{P}}_c(y)$ and $\widehat{\mathbb{P}}_v(y)$, but now with reference to the three attainable regions from Propositions 1-3.

Under our strongest set of assumptions—when noise is symmetric and independent of preferences and preferences are symmetric as well—Proposition 3 establishes that the attainable set is $S_X \cup S_y$. Within the prior literature, 45% of observations lie outside of this region, and within the [McGranaghan et al. \(2024, 2026\)](#) datasets, 23% of observations lie outside of this region.

To formally assess whether we can reject the null that an observation came from $S_X \cup S_y$, we deploy the test recommended by [Miller et al. \(2026\)](#) for sign congruence between two estimators.²⁴ Proposition 3(1) requires $\mathbb{P}_c(y) - 0.5$ and $\mathbb{P}_v(y) - 0.5$ to be sign-congruent such that $(\mathbb{P}_c(y) - 0.5)(\mathbb{P}_v(y) - 0.5) \geq 0$. The test is based on two

²³Within the paired data of [McGranaghan et al. \(2024, 2026\)](#), we can also evaluate the test of proportions accounting for the correlation between choices and valuations, adjusting the z -score. With this adjustment, 61% of [McGranaghan et al. \(2024, 2026\)](#) experiments exhibit $z > 1.645$, rejecting the null hypothesis of $\mathbb{P}_c(y) = \mathbb{P}_v(y)$ in a one-sided test.

²⁴This test was originally proposed in the statistics literature by [Russek-Cohen and Simon \(1993\)](#).

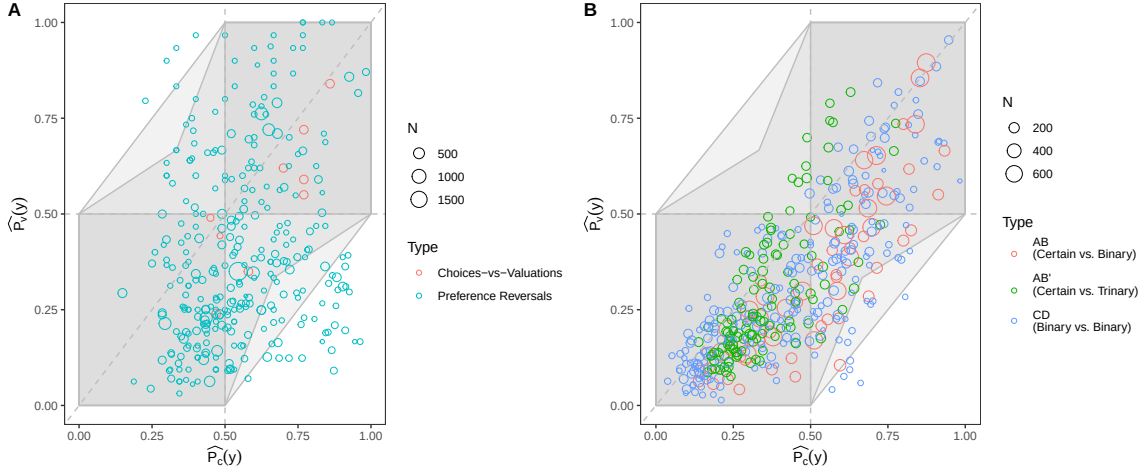


Figure 4: Applying Tests Based on Propositions 1, 2, and 3(1)

Notes: Figure presents observed aggregate choice and valuation proportions, $\widehat{\mathbb{P}}_c(y)$ and $\widehat{\mathbb{P}}_v(y)$, for 323 experiments from prior literature (panel A) and 424 experiments from [McGranaghan et al. \(2024, 2026\)](#) (panel B). Data harmonized such that greater probability corresponds to greater likelihood of preferring safer alternative. Each circle reflects one experiment, size corresponds to experimental sample size. Gray regions correspond to attainable sets (i.e., null hypotheses to test) under stable preferences and assumptions of Propositions 1, 2, and 3(1).

relevant z -scores:

$$z_c = \frac{\widehat{\mathbb{P}}_c(y) - 0.5}{\sqrt{\frac{\widehat{\mathbb{P}}_c(y)(1 - \widehat{\mathbb{P}}_c(y))}{N_c}}} \quad \text{and} \quad z_v = \frac{\widehat{\mathbb{P}}_v(y) - 0.5}{\sqrt{\frac{\widehat{\mathbb{P}}_v(y)(1 - \widehat{\mathbb{P}}_v(y))}{N_v}}}.$$

The test rejects the null hypothesis of sign congruence if both $(\widehat{\mathbb{P}}_c(y) - 0.5)(\widehat{\mathbb{P}}_v(y) - 0.5) < 0$ and $\min(|z_c|, |z_v|) > \bar{z}_\alpha$. The threshold $\bar{z}_\alpha$ is set to a critical value $\Phi^{-1}(1 - \alpha)$, and we select $\alpha = 0.05$ such that $\bar{z}_\alpha = 1.645$. For details see Appendix G.²⁵ Deploying this test, we reject the null that an observation came from $S_X \cup S_y$ for only 20% of experiments from the prior literature and only 12% of experiments from the [McGranaghan et al. \(2024, 2026\)](#) datasets. In other words, even under our strongest set of assumptions, roughly 80-90% of experiments across the two datasets are consistent

²⁵If $\mathbb{P}_c(y)$ and $\mathbb{P}_v(y)$ are known to be highly negatively correlated (i.e., a correlation smaller than around -0.8) an adjustment to $\bar{z}_\alpha$ is required, and when the correlation is -1 , the threshold value for $\alpha = 0.05$ is actually $\bar{z}_\alpha = 1.96$. [Russek-Cohen and Simon \(1993\)](#) Table 2 provides a tabulation of the relevant threshold for $\alpha = 0.05$ depending on the correlation. For testing Proposition 3(1), the issue of negative correlation is largely irrelevant as $\mathbb{P}_c(y)$ and $\mathbb{P}_v(y)$ should generally be positively correlated; indeed, within the [McGranaghan et al. \(2024, 2026\)](#) datasets 414 of 424 experiments have a positive correlation and the smallest correlation is -0.19 .

with stable preferences.

Moving to weaker sets of assumptions, even more experiments become consistent with stable preferences. If we only assume that noise is symmetric and independent of preferences, Proposition 2 establishes that the attainable set is $\text{conv}(C_y \cup C_X) \cup \text{conv}(V_y \cup V_X)$. Within the prior literature, only 16% of observations lie outside of this region, and within the McGranaghan et al. (2024, 2026) datasets, only 4% of observations lie outside of this region. Deploying a modified version of the sign-congruence test, we reject the null that an observation came from $\text{conv}(C_y \cup C_X) \cup \text{conv}(V_y \cup V_X)$ for only 9% of experiments from the prior literature and only 0.5% of experiments from the McGranaghan et al. (2024, 2026) datasets.²⁶

Finally, under our weakest set of assumptions, where we assume only that the noise is symmetric, Proposition 1 establishes that the attainable set is $\text{conv}(S_y \cup S_X)$. Only 11% of experiments from the prior literature and only 1% of experiments from the McGranaghan et al. (2024, 2026) datasets lie outside of this region; and we can reject the null that an observation came from $\text{conv}(S_y \cup S_X)$ for only 3% of experiments from the prior literature and for no experiments from the McGranaghan et al. (2024, 2026) datasets.²⁷

4.4 Using Information on Mean Preferences

Because the McGranaghan et al. (2024, 2026) dataset contains within-subject data on both choices and valuations, we can use it to assess our conditional predictions from Propositions 3(2) and 5. Formal tests of these predictions are challenging as we have only an estimate, rather than a perfect measure, of the conditioning events $y > \mu$ and $y < \mu$, and thus we pursue a more qualitative analysis. At the same time,

²⁶As suggested by Miller et al. (2026), we implement a base change corresponding to the linear restrictions of the boundaries for $\text{conv}(C_y \cup C_X) \cup \text{conv}(V_y \cup V_X)$, after which we can apply the sign-congruence test. Notably, the base change induces negative correlation that requires adjustment of the threshold value $\bar{z}_\alpha$, which we implement in our analysis. For details see Appendix G.

²⁷To develop a statistical test, we use the linear restrictions imposed by the lower and upper boundaries of $\text{conv}(S_y \cup S_X)$, $L : \mathbb{P}_v = \mathbb{P}_c - 0.5$ and $U : \mathbb{P}_v = \mathbb{P}_c + 0.5$. The associated z -scores for testing these linear restrictions are

$$z_L = \frac{\widehat{\mathbb{P}}_v(y) - \widehat{\mathbb{P}}_c(y) + 0.5}{\sqrt{\frac{\widehat{\mathbb{P}}_v(y)(1-\widehat{\mathbb{P}}_v(y))}{N_v} + \frac{\widehat{\mathbb{P}}_c(y)(1-\widehat{\mathbb{P}}_c(y))}{N_c}}} \quad \text{and} \quad z_U = \frac{\widehat{\mathbb{P}}_v(y) - \widehat{\mathbb{P}}_c(y) - 0.5}{\sqrt{\frac{\widehat{\mathbb{P}}_v(y)(1-\widehat{\mathbb{P}}_v(y))}{N_v} + \frac{\widehat{\mathbb{P}}_c(y)(1-\widehat{\mathbb{P}}_c(y))}{N_c}}}.$$

We reject the null that an observation came from $\text{conv}(S_y \cup S_X)$ if either (i) $\widehat{\mathbb{P}}_v(y) < \widehat{\mathbb{P}}_c(y)$ and $z_L < -1.645$ or (ii) $\widehat{\mathbb{P}}_v(y) > \widehat{\mathbb{P}}_c(y)$ and $z_U > 1.645$.

these predictions more naturally lend themselves to aggregating observations across experiments.

When noise is symmetric and independent of preferences and preferences are symmetric as well, an implication of Proposition 3(2) is that $\mathbb{P}_c(y) > \frac{1}{2}$ when $y > \mu$ and $\mathbb{P}_c(y) < \frac{1}{2}$ when $y < \mu$.²⁸ To assess this prediction, we need an estimate for μ ; we use the average valuation $\overline{y_{vi}}$ calculated for each of the 84 groupings.²⁹ We then analyze how $\widehat{\mathbb{P}}_c(y)$ varies with $y - \overline{y_{vi}}$, where a more positive (negative) $y - \overline{y_{vi}}$ is a stronger signal that $y > \mu$ ($y < \mu$).

Panel A of Figure 5 plots $\widehat{\mathbb{P}}_c(y)$ by $y - \overline{y_{vi}}$ in a bin-scatter. Consistent with the prediction, there is a tight connection between distance to average valuation and choice proportions. The correlation between binned values is 0.896, and 95% of binned combinations (38 of 40) lie within the prediction region for Proposition 3(2). In addition, overall average behavior is also consistent with Proposition 3(2): For the 47,748 observations with $y - \overline{y_{vi}} < 0$, the average choice proportion is 0.364, whereas for the 20,700 observations with $y - \overline{y_{vi}} > 0$, the average choice proportion is 0.645.

When we make additional assumptions that preferences are unimodal and that the choice and valuation noise are independent and both have positive densities, an implication of Proposition 5 is that $\mathbb{P}_c(y|y_{vi} = y) > \frac{1}{2}$ when $y > \mu$ and $\mathbb{P}_c(y|y_{vi} = y) < \frac{1}{2}$ when $y < \mu$. To assess this prediction, we use $y - \overline{y_{vi}}$ as above, but now we restrict attention to individuals where $y_{vi} \approx y$, specifically $|y - y_{vi}| \leq 1$ (i.e., whose valuation price list switch points are within 1 row of the choice experimental value of y). Of 68,448 individual observations, 3,280 satisfy this criterion.

Panel B of Figure 5 plots $\widehat{P}_c(y|y_{vi} \approx y)$ by $y - \overline{y_{vi}}$ in a bin-scatter. Here, the evidence looks inconsistent with stable preferences. While there is a positive correlation of 0.485 between binned values, only 55% of binned combinations (22 of 40) lie within the prediction region for Proposition 5. In terms of overall average behavior, for the 1532 observations with $y - \overline{y_{vi}} > 0$, the average choice proportion of 0.634 is consistent with Proposition 5, however, for the 1748 observations with $y - \overline{y_{vi}} < 0$, the average choice proportion 0.551 is inconsistent with Proposition 5.³⁰

²⁸Additionally, if the distributions of heterogeneity (i.e., of η) and noise (i.e., of ε_c and ε_v) are comparable across experiments, then the model would predict a positive relationship between $\widehat{\mathbb{P}}_c(y)$ and $y - \overline{y_{vi}}$.

²⁹Again, a grouping is a set of experiments that all use the same valuation task for a specific lottery X , but which differ in y . On average, there are 815 subjects per grouping.

³⁰While panel B of Figure 5 is inconsistent with the predictions of Proposition 5, and this inconsistency could be due to a failure of stable preferences, it could also be due to a failure of

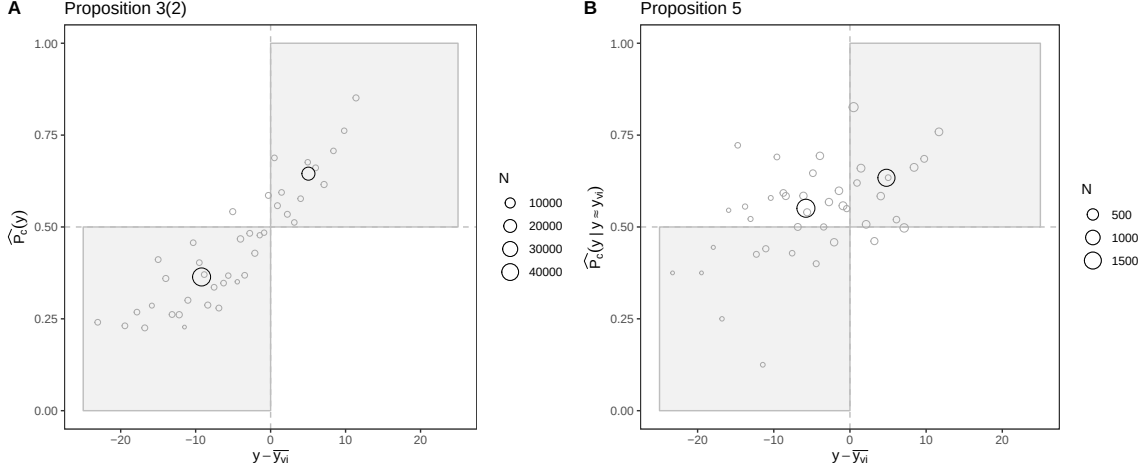


Figure 5: Evaluating Conditional Predictions from Propositions 3(2) and 5

Notes: Figure presents observed choice proportions $\widehat{\mathbb{P}}_c(y)$ (panel A) and conditional choice proportions $\widehat{\mathbb{P}}_c(y|y \approx y_{vi})$ (panel B) against distance to mean valuation ($y - \overline{y}_{vi}$), where light gray regions correspond to predicted behavior. Panel A uses all 68,448 observations from the [McGranaghan et al. \(2024, 2026\)](#) datasets. We first calculate $\overline{y}_{vi}$ for each of the 84 groupings and then calculate $y - \overline{y}_{vi}$ for each of the 68,448 observations. We bin x-axis into 40 approximately equally sized bins of $y - \overline{y}_{vi}$, and for each bin panel A plots the proportion $\widehat{\mathbb{P}}_c(y)$ choosing safer option against average value of $y - \overline{y}_{vi}$, where circle size corresponds to number of observations in a bin. Panel B presents the same exercise except, after calculating $y - \overline{y}_{vi}$ for all observations, we restrict data to the 3280 observations with $|y - y_{vi}| \leq 1$; thus, proportion choosing safer option in each bin is interpreted as $\widehat{\mathbb{P}}_c(y|y \approx y_{vi})$. In both panels, black circles correspond to overall averages when $y - \overline{y}_{vi} < 0$ versus when $y - \overline{y}_{vi} > 0$.

4.5 Structured Empirical Analysis

When one has one has within-subject data on both choices and valuations (as in the [McGranaghan et al. \(2024, 2026\)](#) dataset), and one is willing to make structural assumptions about the distributions of preference heterogeneity and noise, then one can develop a structural test for the null of stable mean preferences and, when stability is rejected, one can estimate the magnitude of deviations from stability. In this section, we implement the two methods described in Section 3.4 that are built on an assumption that heterogeneity and noise are joint normally distributed.

For each of the 84 groupings in the [McGranaghan et al. \(2024, 2026\)](#) data, we use the probit specification in equation (4) from Section 3.4.2 to obtain estimates for α , β_1 , and β_2 . The left-hand top panel of Table 2 summarizes the results by presenting

ancillary assumptions. In fact, Appendix Figure A.1 assesses the symmetry assumptions by presenting a histogram of $y_{vi} - \overline{y}_{vi}$ across all 68,448 observations. While the mean value of $y_{vi} - \overline{y}_{vi}$ is 0 by construction, the median value is \$2.77 indicating a failure of symmetry.

the sample-weighted mean and median of each regression coefficient along with the sample-weighted 25th-75th percentiles. For each of the 84 groupings, we also conduct the MLE approach described in Section 3.4.3 (also see Appendix E). This approach yields, for each grouping, estimates of the structural parameters in equation (3): μ_v , μ_c , $\sqrt{\sigma_v^2 + \gamma_v^2}$, $\sqrt{\sigma_c^2 + \gamma_c^2}$, and γ_{cv} . The right-hand top panel of Table 2 summarizes the results in an analogous way.

The null of stable preferences requires $\mu_v - \mu_c = 0$, which we can test using either approach. For the probit estimation, the structure implies $\mu_v - \mu_c = \frac{\alpha}{\beta_1 + \beta_2}$ (see Section 3.4.2); hence, we can transform the estimated probit coefficients into an estimate for $\mu_v - \mu_c$. For the MLE, it is straightforward to transform the estimated μ_v and μ_c into an estimate for $\mu_v - \mu_c$. These estimates are summarized in the bottom panel of Table 2. Both approaches are well-powered, and we reject the null of $\mu_v - \mu_c = 0$ in 65 groupings (77%) under the probit approach and in 69 groupings (82%) under the MLE approach. Thus, with the additional structure of joint-normality, we frequently reject the null hypothesis of stability.

That said, when we study the nature of deviations from $\mu_v - \mu_c = 0$, the message differs from the received wisdom that choices induce greater risk aversion than valuations. Overall, there appears to be no systematic direction. The mean estimate $\widehat{\mu_v - \mu_c}$ is $-\$0.11$ under probit and $-\$2.20$ under MLE (and normalizing the estimates to expected-value units yields means of $\$0.47$ and $-\$0.21$). The estimates indicate that choices induce greater risk aversion in 54% of groupings under probit and 48% under MLE, whereas the estimates indicate that valuations induce greater risk aversion in 46% of groupings under probit and 52% under MLE.

In terms of the behavioral impact of these differences, it is useful to quantify $\mathbb{P}(\mu_v | y_{vi} = \mu_v)$, which is the choice probability in the idealized experiment where we are able to select $y = \mu_v$ and in addition focus on participants who state a valuation equal to that μ_v . Under joint normality and stable preferences, this quantity would be $\frac{1}{2}$. To obtain an estimate for this quantity, we can use $\Phi(\hat{\alpha})$ under probit, and we can use $\Phi\left(\frac{\widehat{\mu_v - \mu_c}}{\hat{V}}\right)$ under MLE. Estimates for these objects are also summarized in the bottom panel of Table 2.

Figure 6 provides a visual depiction of our estimates for both $\mu_v - \mu_c$ and $\mathbb{P}(\mu_v | y_{vi} = \mu_v)$ for both the probit approach (in panel A) and the MLE approach (in panel B). For each panel, a point reflects one of the 84 groupings, and solid points reflect the groupings for which we reject the null of $\mu_v - \mu_c = 0$. The two panels both highlight

Table 2: Summary of Structural Estimation

(1)		(2)	
Probit		MLE	
Constant ($\hat{\alpha}$)	0.13/0.14 [-0.26,0.47]	$\hat{\mu}_v$	32.04/35.4 [28.71,37.77]
$y - \overline{y_{vi}}$ ($\hat{\beta}_1$)	0.03/0.01 [0.00,0.03]	$\hat{\mu}_c$	29.44/30.75 [20.56,41.4]
$y - y_{vi}$ ($\hat{\beta}_2$)	0.03/0.03 [0.03,0.04]	$\sqrt{\sigma_v^2 + \gamma_v^2}$	12.94/13.79 [9.36,15.27]
		$\sqrt{\sigma_c^2 + \gamma_c^2}$	28.32/23.87 [17.26,37.25]
		$\sqrt{\gamma_{cv}}$	10.99/9.88 [8.67,15.7]
# Observations	68448	# Observations	68448
# Groupings	84	# Groupings	84
$\widehat{\mu_v - \mu_c} = \frac{\hat{\alpha}}{\hat{\beta}_1 + \hat{\beta}_2}$	-0.11/1.97 [-4.86,6.44]	$\widehat{\mu_v - \mu_c}$	-2.2/-1.58 [-9.15,5.52]
$(\widehat{\mu_v - \mu_c})^{\text{norm}}$	0.47/0.60 [-1.76,2.96]	$(\widehat{\mu_v - \mu_c})^{\text{norm}}$	-0.21/-0.21 [-2.23,2.12]
$\Phi(\hat{\alpha})$	0.54/0.56 [0.40,0.68]	$\Phi\left(\frac{\widehat{\mu_v - \mu_c}}{\hat{V}}\right)$	0.51/0.48 [0.38,0.62]
		$\frac{\widehat{\gamma_{cv}}}{\sqrt{\sigma_v^2 + \gamma_v^2} \sqrt{\sigma_c^2 + \gamma_c^2}}$	0.36/0.38 [0.31,0.42]

Notes: Upper panel presents summary of raw coefficient estimates for 84 separate probit regressions in column (1) and raw parameter estimates for 84 separate maximum likelihood estimations (MLE) in column (2). For each coefficient or parameter, table presents sample-weighted mean/median along with sample-weighted 25th and 75th percentiles in brackets. Lower panel uses raw estimates to provide estimates of other objects. Estimated preference difference $\widehat{\mu_v - \mu_c}$ calculated as $\frac{\hat{\alpha}}{\hat{\beta}_1 + \hat{\beta}_2}$ for probit and directly from parameter estimates with sign adjustments as appropriate for MLE (see Appendix E); for both, $\widehat{\mu_v - \mu_c} > 0$ indicates greater risk aversion in choices. $(\widehat{\mu_v - \mu_c})^{\text{norm}}$ expresses mean preference difference in expected value units: if elicited value occurs with probability q , $(\widehat{\mu_v - \mu_c})^{\text{norm}} = q(\widehat{\mu_v - \mu_c})$. Estimated $\mathbb{P}(\mu_v | y_{vi} = \mu_v)$ calculated as $\Phi(\hat{\alpha})$ in probit and $\Phi\left(\frac{\widehat{\mu_v - \mu_c}}{\hat{V}}\right)$ in MLE, where Φ is standard normal CDF and $\hat{V} = \sqrt{\sigma_c^2 + \gamma_c^2 - \frac{\widehat{\gamma_{cv}}^2}{\sigma_v^2 + \gamma_v^2}}$. In MLE, $\frac{\widehat{\gamma_{cv}}}{\sqrt{\sigma_v^2 + \gamma_v^2} \sqrt{\sigma_c^2 + \gamma_c^2}}$ provides lower bound on correlation between choice and valuation preferences.

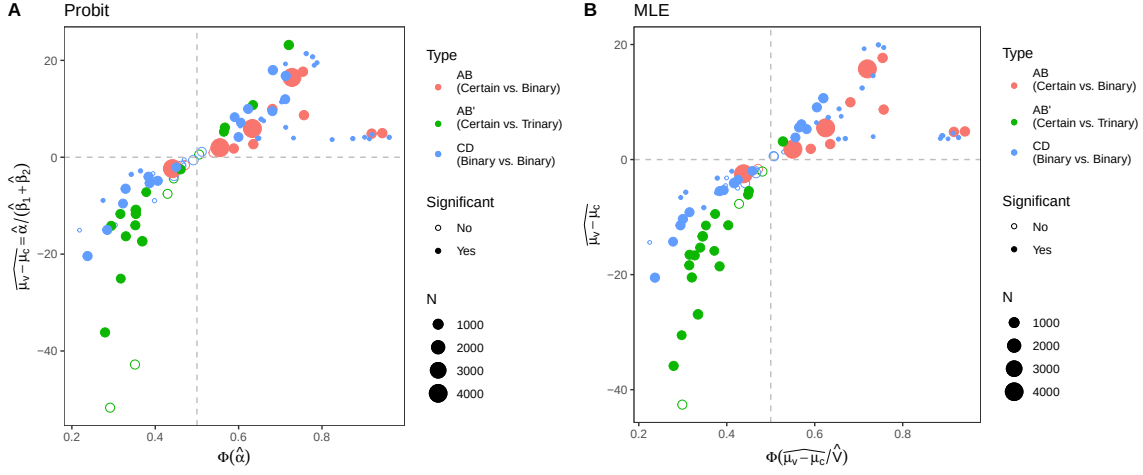


Figure 6: Visual Depiction of Probit and MLE Estimates

Notes: Figure presents estimates of $\widehat{\mu_v - \mu_c}$ (preference difference) against estimates of $\mathbb{P}(\mu_v | y_{vi} = \mu_v)$ (choice probability in idealized experiment) for each of 84 separate probit estimates in panel A and 84 separate MLE estimates in panel B. In each panel, each circle corresponds to one estimation; circle size corresponds to number of observations in grouping used for that estimation; and filled circles represent groupings for which the null hypothesis of stable mean preferences ($H_0 : \mu_v - \mu_c = 0$) is rejected at the 5% level in a two-sided test. Estimates from the two approaches are strongly correlated: estimates for $\mu_v - \mu_c$ are correlated 0.92; estimates for $\Phi(\cdot)$ are correlated 0.97.

how we frequently reject the null, and that the magnitude of the deviations can be large in terms of either the magnitude of $\widehat{\mu_v - \mu_c}$ or the deviation of $\mathbb{P}(\mu_v | y_{vi} = \mu_v)$ from $\frac{1}{2}$. At the same time, neither panel suggests a systematic overall trend toward more risk aversion in choices (i.e., $\widehat{\mu_v - \mu_c} > 0$ and $\mathbb{P}(\mu_v | y_{vi} = \mu_v) > \frac{1}{2}$) or more risk aversion in valuations (i.e., $\widehat{\mu_v - \mu_c} < 0$ and $\mathbb{P}(\mu_v | y_{vi} = \mu_v) < \frac{1}{2}$).

Beyond testing the null of $\mu_v - \mu_c = 0$, our structural approaches yields further insights. First, because they provide a measure of the magnitudes of deviations, they permit us to study differences across different types of problems. For instance, as described in Section 4.1, in the McGranaghan et al. (2024, 2026) data, subjects make three types of comparisons: (i) a comparison between a certain amount and a binary lottery (which they label an *AB* decision), (ii) a comparison between a certain amount and a trinary lottery (an *AB'* decision), and (iii) a comparison between two binary lotteries (a *CD* decision). Figure 6 uses different colors to denote the different types of comparisons (see also Appendix Table A.1 for a summary analogous to Table 2 for each type of comparison), and a clear pattern emerges: For certain vs. binary comparisons, the estimates indicate that choices tend to induce more risk aversion

(i.e., a trend toward $\widehat{\mu_v - \mu_c} > 0$ and $\mathbb{P}(\widehat{\mu_v | y_{vi}} = \mu_v) > \frac{1}{2}$); for certain vs. trinary comparisons, the estimates indicate that valuations tend to induce more risk aversion (i.e., a trend toward $\widehat{\mu_v - \mu_c} < 0$ and $\mathbb{P}(\widehat{\mu_v | y_{vi}} = \mu_v) < \frac{1}{2}$); and for binary vs. binary comparisons, the estimates indicate no systematic deviation.^{31,32}

Our MLE approach also provides evidence on the nature of heterogeneity and noise. For instance, our theoretical analysis highlights the importance of whether decision noise has a larger impact on choices versus valuations. Table 2 provides a summary of estimates for $\sqrt{\gamma_c^2 + \sigma_c^2}$ and $\sqrt{\gamma_v^2 + \sigma_v^2}$. In standard deviation units, choices are roughly twice as variable as valuations. Because under the null of stable preferences $\gamma_c^2 = \gamma_v^2$, these estimates suggest substantially greater decision noise for choices than for valuations.

Finally, given the structure that we have assumed, another feature of stable preferences is that y_{ci}^* should equal y_{vi}^* at the individual level, and thus $\text{corr}(y_{ci}^*, y_{vi}^*) = 1$. Using the MLE estimates, Table 2 provides a summary of estimates for the quantity $\frac{\gamma_{cv}}{\sqrt{\gamma_c^2 + \sigma_c^2} \sqrt{\gamma_v^2 + \sigma_v^2}}$, which reflects a lower bound on the correlation in preferences (since $\sigma_c^2 \geq 0$ and $\sigma_v^2 \geq 0$). Interestingly, these lower-bound correlations are relatively high, with an inter-quartile range of 0.31 to 0.42. These correspond favorably to correlations discussed in the psychology literature on stability of personality traits and correlations discussed in the economics literature on test-retest stability in preference measurements over time (for discussion on the relationship between stability in economic preference measurements and psychological personality traits, see [Meier and Sprenger, 2015](#)).

5 Discussion

Measurement of economic preferences requires confidence in our measurement or elicitation techniques. In this paper, we make both theoretical and empirical contributions to the study of preference stability in the domain of risk preferences across two elicitation techniques: choices versus valuations.

³¹[Freeman et al. \(2019\)](#) contrast certain vs. binary and binary vs. binary comparisons. They show significant deviations between $\widehat{\mathbb{P}}_c(y)$ and $\widehat{\mathbb{P}}_v(y)$ when considering (\$3, 1) vs. (\$4, .8) using the traditional test of equality; and they show insignificant deviations when considering (\$3, .5) vs. (\$4, .4). The statistical differences between one safe alternative and one risky alternative as opposed to two risky alternatives resonates with our own findings.

³²The [McGranaghan et al. \(2024, 2026\)](#) data also includes variation in the probabilities used in lotteries, and in Appendix F we document some systematic patterns in how $\widehat{\mu_v - \mu_c}$ varies with those probabilities.

Theoretically, we highlight that the null of stable preferences can be assessed only in conjunction with ancillary assumptions about preference heterogeneity and noise, and how it is important to make these ancillary assumptions explicit. We derive a series of predictions under varying ancillary assumptions, motivating new tests of the null of stable preferences in which one compares an observed combination $(\mathbb{P}_c(y), \mathbb{P}_v(y))$ to null attainable regions instead of the traditional test of identical measurements. We also develop more structured approaches to testing for stability if the analyst is willing to make functional assumptions for preference heterogeneity and noise. Critically, rejection in any of our tests may reflect a failure of stability or may reflect a failure of ancillary assumptions.

Empirically, we implement tests of the various predictions that we derive so as to obtain an updated sense of the extent to which the evidence is consistent with stable preferences. When we restrict ourselves to tests based on only $\mathbb{P}_c(y)$ and $\mathbb{P}_v(y)$, the message that emerges is that the vast majority of experiments from the prior literature and from the recent convenience dataset from [McGranaghan et al. \(2024, 2026\)](#) are consistent with stable preferences. When we instead use tests based on richer empirical objects, we start to see some evidence that is inconsistent with stable preferences (and/or our ancillary assumptions). However, the nature of these deviations is not in line with the received wisdom that choices induce more risk aversion than valuations. In fact, overall, we see no systematic directional pattern, although we do see some systematic variation with problem types that could be explored in future work.

We conclude by discussing some broader messages from our analysis. Our analysis highlights the value of focusing on richer empirical objects than traditional choice and valuation proportions. In our domain, if we limit attention to predictions for $\mathbb{P}_c(y)$ and $\mathbb{P}_v(y)$, then even under our strongest assumptions of full joint normality, the prediction is merely that $(\mathbb{P}_c(y), \mathbb{P}_v(y))$ is in $S_y \cup S_X$ (minus the extreme edges). However, under the same assumptions, if the data contain information on both choices and valuations for the same subjects, we can use this more detailed data to implement our structural approaches. For the data we study—and under the same assumptions of full joint normality—when using data on only $\mathbb{P}_c(y)$ and $\mathbb{P}_v(y)$, we find substantial consistency with stable preferences, while when using richer data, we find substantial inconsistency.

Important considerations emerge from our development for researchers linking models of decision making to data. First, it is critical to pay careful attention to

heterogeneity and to make one’s assumptions explicit. For instance, it is tempting to conjecture that if the average preference in the population is for y over X , then the aggregate population averages for $\mathbb{P}_c(y)$ and $\mathbb{P}_v(y)$ ought to lie in S_y . Our analysis highlights that this conjecture is not generally correct, and holds only under certain assumptions on the nature of heterogeneity and noise. Second, when assessing whether data is consistent with a theoretical prediction, sampling variation must be accounted for given (typically small) experimental samples. Our results show the importance of developing appropriate statistical tests, as we document sizable differences between the proportion of experiments that are outside our attainable regions and the proportion that are *significantly* outside our attainable regions.

We also provide some guidance to experimentalists deciding which elicitation techniques to use (even when they are not interested per se in the stability of preferences). In the data we study, it turns out that choice data and valuations data seem to be highly correlated, and thus it is not clear that either yields less biased measures of preferences. At the same time, our structural analysis suggests that, at least under the null of stable preferences, valuations yield more precise measurements than choices. While one might be tempted to conclude, then, that valuations are better to use, we add two caveats. First, we suspect both techniques can be subject to bias due to framing effects, and such effects are not part of our analysis in this paper. Second, and more important, our analysis—especially our structural analysis—highlights the value of collecting both (and more generally of using multiple elicitation techniques). Doing so can provide a richer picture of preferences, and can permit one to assess the existence of inconsistencies, and possibly to correct for the impact of elicitation procedures on behavior.

Finally, while we have framed our theoretical analysis in the domain of risk preferences, it can be applied more broadly. Whenever one is interested in measuring preferences between some generic options A and B, one could do so via binary choices or via valuations, and all of our results apply to any such generic comparison (see the general version of our model in Appendix C). Moreover, our key intuitions clearly extend to other assessments of procedural invariance. For example, a substantial debate continues on appropriate payment mechanisms in experimental protocols depending on whether subjects “isolate” each experimental choice (see, e.g., [Holt, 1986](#); [Starmer and Sugden, 1991](#); [Harrison and Swarthout, 2014](#); [Cox et al., 2015](#)). This debate revolves around evidence of inconsistencies between choice proportions depending

on whether only a single choice is made or the same choice is embedded in a larger experiment. Our analysis extends naturally to such comparisons and implies that a simple comparison of choice proportions is perhaps not appropriate. The techniques that we develop in this paper could thus prove useful in this and many other domains.

References

- M. Allais. Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école américaine. *Econometrica*, 21(4):503–546, 1953.
- R. Bouchouicha, R. Oprea, F. M. Vieder, and J. Wu. Cognitive frictions and canonical patterns in risk-taking. Technical report, Working Paper, 2026.
- A. L. Brown and P. J. Healy. Separated decisions. *European Economic Review*, 101: 20–34, 2018.
- J. C. Cox, V. Sadiraj, and U. Schmidt. Paradoxes and mechanisms for choice under risk. *Experimental Economics*, 18(2):215–250, 2015.
- D. J. Freeman and G. Mayraz. Why choice lists increase risk taking. *Experimental Economics*, 22:131–154, 2019.
- D. J. Freeman, Y. Halevy, and T. Kneeland. Eliciting risk preferences using choice lists. *Quantitative Economics*, 10(1):217–237, 2019.
- D. M. Grether and C. R. Plott. Economic theory of choice and the preference reversal phenomenon. *American Economic Review*, 69(4):623–638, 1979.
- G. W. Harrison and J. T. Swarthout. Experimental payment protocols and the bipolar behaviorist. *Theory and Decision*, 77(3):423–438, 2014.
- C. A. Holt. Preference reversals and the independence axiom. *American Economic Review*, 76(3):508–515, 1986.
- M. W. Khaw, Z. Li, and M. Woodford. Cognitive imprecision and small-stakes risk aversion. *Review of Economic Studies*, 88(4):1979–2013, 2021.
- S. Lichtenstein and P. Slovic. Reversals of preference between bids and choices in gambling decisions. *Journal of Experimental Psychology*, 89(1):46–55, 1971.

- S. Lichtenstein and P. Slovic, editors. *The Construction of Preference*. Cambridge University Press, New York, 2006.
- Y. Lu. Magnitude effects in preference reversals: A meta-analysis. *Review of Economic Design*, 30:169–202, 2026.
- C. McGranaghan, K. Nielsen, T. O’Donoghue, J. Somerville, and C. D. Sprenger. Distinguishing common ratio preferences from common ratio effects using paired valuation tasks. *American Economic Review*, 114(2):307–347, 2024.
- C. McGranaghan, K. Nielsen, T. O’Donoghue, J. Somerville, and C. D. Sprenger. Connecting common ratio and common consequence preferences. *Journal of Political Economy*, forthcoming, 2026.
- S. Meier and C. D. Sprenger. Temporal stability of time preferences. *Review of Economics and Statistics*, 97(2):273–286, 2015.
- D. L. Miller, F. Molinari, and J. Stoye. Testing sign congruence between two parameters. *Journal of Applied Econometrics*, 41(1):3–11, 2026.
- F. Mosteller and P. Nogue. An experimental measurement of utility. *Journal of Political Economy*, 59(5):371–404, 1951.
- E. Russek-Cohen and R. M. Simon. Qualitative interactions in multifactor studies. *Biometrics*, 49(2):467–477, 1993.
- C. Shubatt and J. Yang. Valuations under tradeoff complexity. Technical report, Working Paper, 2026.
- P. Slovic. The construction of preference. *American Psychologist*, 50(5):364–371, 1995.
- C. Starmer and R. Sugden. Does the random-lottery incentive system elicit true preferences? An experimental investigation. *American Economic Review*, 81(4):971–978, 1991.
- L. L. Thurstone. The indifference function. *Journal of Social Psychology*, 2(2):139–167, 1931.
- A. Tversky and D. Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323, 1992.

A. Tversky and R. H. Thaler. Anomalies: preference reversals. *Journal of Economic Perspectives*, 4(2):201–211, 1990.

Measuring Economic Preferences in the Presence of Noise: The Connections Between Choices and Valuations

Ted O'Donoghue, Charles D. Sprenger, Po Hyun Sung, and Ben Wincelberg

August 30, 2026

Supplementary Appendix

A Additional Tables and Figures

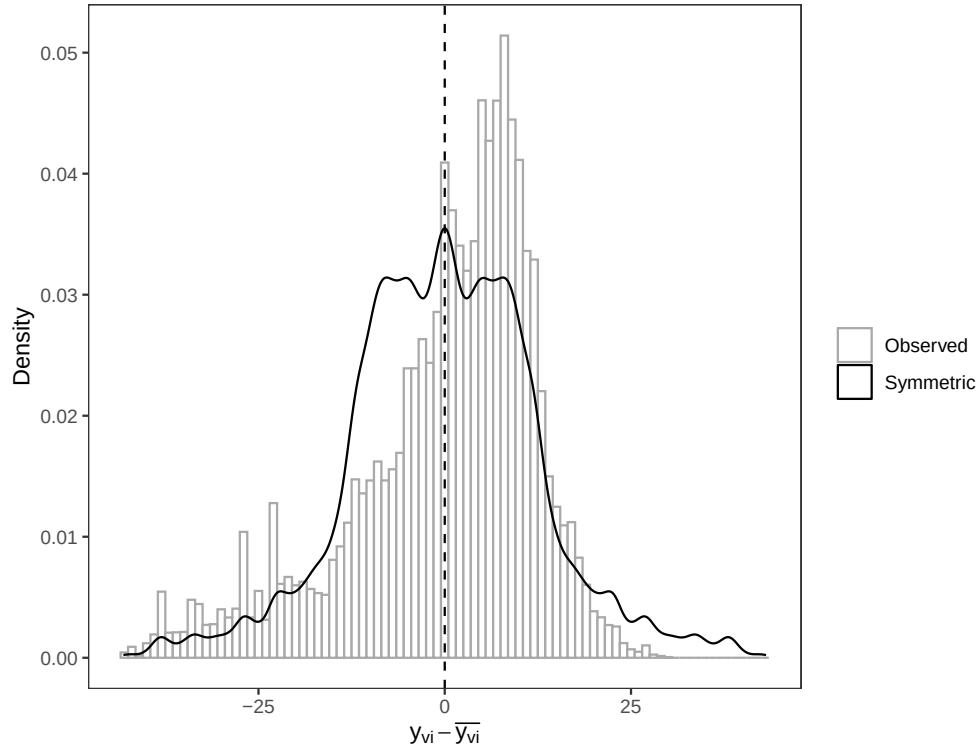


Figure A.1: Evaluating Symmetry of $y_{vi} - \bar{y}_{vi}$

Notes: Figure presents histogram of $y_{vi} - \bar{y}_{vi}$ for 68,448 observations in [McGranaghan et al. \(2024, 2026\)](#) (gray bars). Solid black line corresponds to symmetrized distribution constructed from a combined dataset of $y_{vi} - \bar{y}_{vi}$ and $\bar{y}_{vi} - y_{vi}$. Median of $y_{vi} - \bar{y}_{vi}$ is 2.77.

Table A.1: Summary of Structural Estimation Separated by Problem Type

Sample:	<i>AB</i>	<i>AB'</i>	<i>CD</i>	Sample:	<i>AB</i>	<i>AB'</i>	<i>CD</i>
	Certain vs. Binary (1)	Certain vs. Trinary (2)	Binary vs. Binary (3)		Certain vs. Binary (4)	Certain vs. Trinary (5)	Binary vs. Binary (6)
	Probit				MLE		
Constant ($\hat{\alpha}$)	0.35/0.34 [0.14,0.61]	-0.22/-0.38 [-0.44,-0.1]	0.12/0.03 [-0.29,0.47]	$\hat{\mu}_v$	33.28/37.67 [30.36,39.35]	32.97/33.98 [28.71,37.79]	30.21/33.9 [28.95,35.64]
$y - \overline{y}_{vi}$ ($\hat{\beta}_1$)	0.04/0.03 [0.01,0.04]	0/-0.01 [-0.01,0]	0.04/0.02 [0.01,0.03]	$\hat{\mu}_c$	37.49/41.69 [36.88,44.71]	16.83/20 [6.89,24.45]	29.62/30.65 [25.3,37.93]
$y - y_{vi}$ ($\hat{\beta}_2$)	0.04/0.04 [0.03,0.05]	0.03/0.03 [0.03,0.03]	0.03/0.03 [0.02,0.03]	$\sqrt{\sigma_v^2 + \gamma_v^2}$	12.14/13.53 [8.5,14.64]	15.4/15.29 [14.93,15.99]	12.14/12.84 [8.22,14.94]
				$\sqrt{\sigma_c^2 + \gamma_c^2}$	18.35/17.26 [15.72,22.98]	50.74/47.89 [41.93,62.44]	23.69/22.23 [19.6,29.47]
				$\sqrt{\gamma_{cv}}$	8.92/9.58 [7.88,12.02]	17.59/17.41 [16.26,19.17]	8.75/9.16 [6.39,11.62]
# Observations	25816	16816	25816	# Observations	25816	16816	25816
# Groupings	14	20	50	# Groupings	14	20	50
$\widehat{\mu_v - \mu_c} = \frac{\hat{\alpha}}{\hat{\beta}_1 + \hat{\beta}_2}$	5.51/4.93 [1.8,8.69]	-11.46/-11.7 [-17.34,-2.52]	1.68/1.12 [-4.52,8.28]	$\widehat{\mu_v - \mu_c}$	5.28/4.85 [1.8,8.69]	-16.14/-15.87 [-20.5,-9.46]	-0.59/-1.97 [-5.49,5.29]
$(\widehat{\mu_v - \mu_c})^{\text{norm}}$	2.14/2.83 [1,4.35]	-1.63/-2.25 [-2.95,-0.23]	0.17/0.03 [-0.83,1.66]	$(\widehat{\mu_v - \mu_c})^{\text{norm}}$	2.03/2.73 [1.12,4.35]	-3.27/-2.69 [-4.58,-1.28]	-0.45/-0.21 [-1.41,0.92]
$\Phi(\hat{\alpha})$	0.63/0.63 [0.56,0.73]	0.42/0.35 [0.33,0.46]	0.54/0.51 [0.39,0.68]	$\Phi\left(\frac{\widehat{\mu_v - \mu_c}}{\hat{V}}\right)$	0.62/0.62 [0.55,0.72]	0.37/0.35 [0.32,0.4]	0.5/0.47 [0.38,0.6]
				$\frac{\widehat{\gamma_{cv}}}{\sqrt{\sigma_v^2 + \gamma_v^2} \sqrt{\sigma_c^2 + \gamma_c^2}}$	0.4/0.4 [0.35,0.45]	0.41/0.41 [0.38,0.43]	0.3/0.3 [0.27,0.35]

Notes: Upper panel presents summary of raw coefficient estimates for 84 separate probit regressions in columns (1)-(3) and raw parameter estimates for 84 separate maximum likelihood estimations (MLE) in column (4)-(6). For each coefficient or parameter, table presents sample-weighted mean/median along with sample-weighted 25th and 75th percentiles in brackets. Lower panel uses raw estimates to provide estimates of other objects. Estimated preference difference $\widehat{\mu_v - \mu_c}$ calculated as $\frac{\hat{\alpha}}{\hat{\beta}_1 + \hat{\beta}_2}$ for probit and directly from parameter estimates with sign adjustments as appropriate for MLE (see Appendix E); for both, $\widehat{\mu_v - \mu_c} > 0$ indicates greater risk aversion in choices. $(\widehat{\mu_v - \mu_c})^{\text{norm}}$ expresses mean preference difference in expected value units: if elicited value occurs with probability q , $(\widehat{\mu_v - \mu_c})^{\text{norm}} = q(\widehat{\mu_v - \mu_c})$. Estimated $\mathbb{P}(\mu_v | y_{vi} = \mu_v)$ calculated as $\Phi(\hat{\alpha})$ in probit and $\Phi\left(\frac{\widehat{\mu_v - \mu_c}}{\hat{V}}\right)$ in MLE, where Φ is standard normal CDF and $\hat{V} = \sqrt{\sigma_c^2 + \gamma_c^2 - \frac{\widehat{\gamma_{cv}}^2}{\sigma_v^2 + \gamma_v^2}}$. In MLE, $\frac{\widehat{\gamma_{cv}}}{\sqrt{\sigma_v^2 + \gamma_v^2} \sqrt{\sigma_c^2 + \gamma_c^2}}$ provides lower bound on correlation between choice and valuation preferences.

B Proofs

B.1 Proof of Lemma 1

Proof of Lemma 1. Recall that $S_y = [\frac{1}{2}, 1]^2$ and $S_X = [0, \frac{1}{2}]^2$. To establish $\mathcal{A}_i = S_y \cup S_X$, we first prove that $\mathcal{A}_i \subseteq S_y \cup S_X$, and then prove that $S_y \cup S_X \subseteq \mathcal{A}_i$.

The former is straightforward. F_c and F_v both symmetric and mean zero implies: (i) if $y > y_i^*$ then $\mathbb{P}_{ci}(y) = F_c(y - y_i^*) \geq \frac{1}{2}$ and $\mathbb{P}_{vi}(y) = F_v(y - y_i^*) \geq \frac{1}{2}$, and thus $(\mathbb{P}_{ci}(y), \mathbb{P}_{vi}(y)) \in S_y$; (ii) if $y < y_i^*$ then $\mathbb{P}_{ci}(y) = F_c(y - y_i^*) \leq \frac{1}{2}$ and $\mathbb{P}_{vi}(y) = F_v(y - y_i^*) \leq \frac{1}{2}$, and thus $(\mathbb{P}_{ci}(y), \mathbb{P}_{vi}(y)) \in S_X$; (iii) if $y = y_i^*$ then $\mathbb{P}_{ci}(y) = F_c(0) = \frac{1}{2}$ and $\mathbb{P}_{vi}(y) = F_v(0) = \frac{1}{2}$, and thus $(\mathbb{P}_{ci}(y), \mathbb{P}_{vi}(y)) = (\frac{1}{2}, \frac{1}{2})$. Note that parts (2) and (3) follow from these statements, and it further follows that if $(a, b) \in \mathcal{A}_i$, then $(a, b) \in S_y \cup S_X$.³³

We prove $S_y \cup S_X \subseteq \mathcal{A}_i$ by construction; many are possible, we use a construction that follows the narrative in the text. First, suppose F_c is the CDF of $N(0, \sigma_c^2)$ and F_v is the CDF of $N(0, \sigma_v^2)$. Note that, for any $a \in (0, \frac{1}{2})$ and $z < 0$, there exists a σ_c^2 such that $F_c(z) = a$; and for any $a \in (\frac{1}{2}, 1)$ and $z > 0$, there exists a σ_c^2 such that $F_c(z) = a$. Analogous properties hold for F_v . Hence, for any (a, b) in the interior of S_X , if we select a $y_i^* > y$, there exists a σ_c^2 such that $F_c(y - y_i^*) = a$ and a σ_v^2 such that $F_v(y - y_i^*) = b$; thus any (a, b) in the interior of S_X is also in $\mathcal{A}_i$. Analogously, for any (a, b) in the interior of S_y , if we select a $y_i^* < y$, there exists a σ_c^2 such that $F_c(y - y_i^*) = a$ and a σ_v^2 such that $F_v(y - y_i^*) = b$; thus any (a, b) in the interior of S_y is also in $\mathcal{A}_i$.

It remains to prove that the edges of S_X and S_y are in $\mathcal{A}_i$. Suppose that for each $k \in \{c, v\}$, there exists $x_k'' > x_k' \geq 0$ such that the ε_k is uniformly distributed on $[-x_k'', -x_k'] \cup [x_k', x_k'']$. Note that ε_k is symmetric about zero with CDF

$$F_k(x) = \begin{cases} 0 & x \leq -x_k'' \\ \frac{1}{2(x_k'' - x_k')} (x - (-x_k'')) & -x_k'' \leq x \leq -x_k' \\ \frac{1}{2} & -x_k' \leq x \leq x_k' \\ \frac{1}{2} + \frac{1}{2(x_k'' - x_k')} (x - x_k') & x_k' \leq x \leq x_k'' \\ 1 & x \geq x_k'' \end{cases} \quad (\text{B.1})$$

³³Note that the same conclusion holds if we relax the symmetry assumption, requiring only that ε_c and ε_v have a median of zero.

Consider first the case where $x'_c \geq x''_v$. If we set y_i^* such that $y - y_i^* \in [-x''_v, x''_v]$, we can select all (a, b) on the edge with $a = \frac{1}{2}$ and $b \in [0, 1]$. If we set y_i^* such that $y - y_i^* \in [x'_c, x''_c]$, we can select all (a, b) on the edge with $a \in [\frac{1}{2}, 1]$ and $b = 1$. And if we set y_i^* such that $y - y_i^* \in [-x''_c, -x'_c]$, we can select all (a, b) on the edge with $a \in [0, \frac{1}{2}]$ and $b = 0$. If we instead consider the case where $x'_v \geq x''_c$, we can attain all (a, b) on the remaining three edges, completing the proof.³⁴

□

B.2 Proof of Proposition 1

Proof of Proposition 1. $\mathcal{A} \subseteq \text{conv}(S_y \cup S_X)$ follows directly from Lemma 1 because the aggregate probabilities are expectations of individual probabilities lying in $S_y \cup S_X$. We prove $\text{conv}(S_y \cup S_X) \subseteq \mathcal{A}$ by construction. Suppose a population consists of two groups: Group 1 has η_1 such that $y_1^* < y$; Group 2 has η_2 such that $y_2^* > y$; and the population shares are π_1 and π_2 such that $\pi_1 + \pi_2 = 1$. By Lemma 1, because Group 1 has $y - y_1^* > 0$, and because there are no restrictions on their noise beyond symmetry, their decision probabilities could be any $(a, b) \in S_y$. Analogously, because Group 2 has $y - y_1^* < 0$, and because there are also no restrictions on their noise beyond symmetry—including no requirement that their noise be the same as that for Group 1—their decision probabilities could be any $(a, b) \in S_X$. Finally, because there are no restrictions on the population proportions π_1 and π_2 , it follows that the aggregate decision probabilities can be any mixture of any point in S_y with any point in S_X , and thus $\text{conv}(S_y \cup S_X) \subseteq \mathcal{A}$.

□

B.3 Proof of Proposition 2

Proof of Proposition 2. Define $M \equiv \text{conv}(C_y \cup C_X) \cup \text{conv}(V_y \cup V_X)$, $\hat{M} = M - (\frac{1}{2}, \frac{1}{2})$, and $\hat{\mathcal{A}} = \mathcal{A} - (\frac{1}{2}, \frac{1}{2})$, so the goal is to prove that $\hat{\mathcal{A}} = \hat{M}$. Note that

$$\hat{M} = \left\{ (a, b) \in \left[-\frac{1}{2}, \frac{1}{2}\right]^2 : |b - 2a| \leq \frac{1}{2} \right\} \cup \left\{ (a, b) \in \left[-\frac{1}{2}, \frac{1}{2}\right]^2 : |2b - a| \leq \frac{1}{2} \right\}. \quad (\text{B.2})$$

³⁴To parallel the text, we used a normal-normal construction to prove that all points in the interior of $S_y \cup S_X$ are attainable; however, we could have also used the construction in (B.1) for that as well. Again, many different constructions are possible.

Since preferences are independent of noise, $(a, b) \in \hat{\mathcal{A}}$ if and only if there exist F_c and F_v that are symmetric such that $(a, b) \in \text{conv}(K)$, where

$$K \equiv \left\{ \left(F_c(x) - \frac{1}{2}, F_v(x) - \frac{1}{2} \right) \mid x \in \mathbb{R} \right\} \subseteq \left[-\frac{1}{2}, \frac{1}{2} \right]^2.$$

We first prove $\mathcal{A} \subseteq M$, or equivalently, $\hat{\mathcal{A}} \subseteq \hat{M}$. To this end, we fix F_c and F_v and show that $\text{conv}(K) \subseteq \hat{M}$. Since F_c and F_v are CDFs, K is weakly increasing, i.e., for any two points $(x_1, y_1), (x_2, y_2) \in K$, either $(x_1, y_1) \geq (x_2, y_2)$ or $(x_1, y_1) \leq (x_2, y_2)$. Moreover, since F_c and F_v are symmetric, $K = -K$, i.e., if $(x, y) \in K$, then $(-x, -y) \in K$. Thus, every point of K lies in one of the two diagonal orthants, since otherwise a point and its negative would violate monotonicity. Hence, if we define $K^+ \equiv K \cap \left[0, \frac{1}{2} \right]^2$, then $K = K^+ \cup (-K^+)$.

By Fenchel's refinement of Carathéodory's theorem for connected sets,³⁵ if $(a, b) \in \text{conv}(K)$, then (a, b) is a convex combination of two points in K . Suppose, toward a contradiction, that $(a, b) \notin \hat{M}$. Since $a \cdot b < 0$, (a, b) must be a convex combination of a point in K^+ and a point in K^- . Since $K = -K$, there exist $(x_1, y_1), (x_2, y_2) \in K^+$ such that

$$\begin{aligned} (a, b) &= \tau^+(x_1, y_1) + \tau^-(-x_2, -y_2) \\ &= (\tau^+x_1 - \tau^-x_2, \tau^+y_1 - \tau^-y_2) \end{aligned}$$

for some $\tau^+, \tau^- \in [0, 1]$ and $\tau^+ + \tau^- = 1$.

Then by (B.2),

$$|b - 2a| > \frac{1}{2} \quad \text{and} \quad |2b - a| > \frac{1}{2}. \quad (\text{B.3})$$

We consider first the case that $b - 2a > \frac{1}{2}$ and $2b - a > \frac{1}{2}$. Define

$$\begin{aligned} \Psi_C(x_1, y_1, x_2, y_2) &:= \tau^+(y_1 - 2x_1) - \tau^-(y_2 - 2x_2), \\ \Psi_V(x_1, y_1, x_2, y_2) &:= \tau^+(2y_1 - x_1) - \tau^-(2y_2 - x_2), \end{aligned}$$

so that

$$\Psi_C(x_1, y_1, x_2, y_2) = b - 2a, \quad \Psi_V(x_1, y_1, x_2, y_2) = 2b - a.$$

³⁵Note that K is connected since F_c and F_v are continuous. The theorem states that for a connected set $A \subseteq \mathbb{R}^n$, any point in $\text{conv}(A)$ is a convex combination of at most n points in A . Proposition 2 holds without continuity of F_c and F_v with a slightly longer proof.

Suppose first that $\tau^+ \geq \tau^-$, and note that either $(x_1, y_1) \leq (x_2, y_2)$ or $(x_1, y_1) \geq (x_2, y_2)$. If $(x_1, y_1) \leq (x_2, y_2)$ and thus $y_1 \leq y_2$, then

$$\Psi_C(x_1, y_1, x_2, y_2) \leq (\tau^+ - \tau^-)y_1 - 2\tau^+x_1 + 2\tau^-x_2 \leq \frac{\tau^+ - \tau^-}{2} + \tau^- = \frac{1}{2},$$

where the second inequality follows because $(x_1, y_1), (x_2, y_2) \in K^+$ implies boundary conditions $y_1 \leq \frac{1}{2}$, $x_1 \geq 0$, $x_2 \leq \frac{1}{2}$. Similarly, if $(x_1, y_1) \geq (x_2, y_2)$ and thus $x_2 \leq x_1$, then

$$\Psi_C(x_1, y_1, x_2, y_2) \leq \tau^+y_1 - \tau^-y_2 - 2(\tau^+ - \tau^-)x_2 \leq \frac{\tau^+}{2} \leq \frac{1}{2},$$

applying the boundary conditions $y_1 \leq \frac{1}{2}$, $y_2 \geq 0$, and $x_2 \geq 0$. Thus, if $\tau^+ \geq \tau^-$ then $\Psi_C(x_1, y_1, x_2, y_2) \leq \frac{1}{2}$, contradicting (B.3). An analogous argument implies that $\Psi_V(x_1, y_1, x_2, y_2) \leq \frac{1}{2}$ when $\tau^+ \leq \tau^-$, also contradicting (B.3). The case of $b - 2a < -\frac{1}{2}$ and $2b - a < -\frac{1}{2}$ follows a symmetric argument. The other two cases lie outside $\left[-\frac{1}{2}, \frac{1}{2}\right]^2$. Therefore $\text{conv}(K) \subseteq \hat{M}$.

We next show $M \subseteq \mathcal{A}$. We prove this by construction. Suppose F_c and F_v are as in (B.1) from the proof of Lemma 1, which again satisfies Assumption 1. Applying the logic there, when $x'_c \geq x''_v$, the resulting curve will include both the edge where $b = 0$ and $a \in [0, \frac{1}{2}]$ and the edge where $b = 1$ and $a \in [\frac{1}{2}, 1]$. Because we can choose any population distribution we want over this curve, we can attain any point in the convex hull of these two line segments, which is just $\text{conv}(C_y \cup C_X)$. Analogously, when $x'_v \geq x''_c$, the resulting curve will include both the edge where $a = 0$ and $b \in [0, \frac{1}{2}]$ and the edge where $a = 1$ and $b \in [\frac{1}{2}, 1]$, from which we can attain any point in $\text{conv}(V_y \cup V_X)$.

□

B.4 Proof of Proposition 3

Proof of Proposition 3. From the text, the aggregate choice proportions are:

$$\mathbb{P}_c(y) = G_c(y - \mu) \quad \text{and} \quad \mathbb{P}_v(y) = G_v(y - \mu)$$

where G_c is the CDF for random variable $\nu_c = \eta + \varepsilon_c$ and G_v is the CDF for random variable $\nu_v = \eta + \varepsilon_v$. Since η , ε_c , and ε_v are symmetric about zero and $\eta \perp (\varepsilon_c, \varepsilon_v)$, ν_c and ν_v are both symmetric about zero.

Given that ν_c and ν_v are both symmetric about zero, the proof that $\mathcal{A} \subseteq S_y \cup S_X$ is exactly analogous to the proof that $\mathcal{A}_i \subseteq S_y \cup S_X$ for Lemma 1. In addition, letting $\eta = 0$ degenerately, $S_y \cup S_X \subseteq \mathcal{A}$ follows from Lemma 1.

□

B.5 Proof of Proposition 4

Proof of Proposition 4. As in the proof of Proposition 3, the aggregate choice proportions are:

$$\mathbb{P}_c(y) = G_c(y - \mu) \quad \text{and} \quad \mathbb{P}_v(y) = G_v(y - \mu)$$

where G_c is the CDF for random variable $\nu_c = \eta + \varepsilon_c$ and G_v is the CDF for random variable $\nu_v = \eta + \varepsilon_v$. Since η , ε_c , and ε_v are symmetric about zero and $\eta \perp (\varepsilon_c, \varepsilon_v)$, ν_c and ν_v are both symmetric about zero.

When $y = \mu$, $\mathbb{P}_c(y) = G_c(0)$, $\mathbb{P}_v(y) = G_v(0)$, which equal $\frac{1}{2}$ since ν_c and ν_v are symmetric and continuous. Letting $t = y - \mu$, it remains to show $G_c(t) \neq G_v(t)$ for all $t \neq 0$.

Because η has a density by Assumption 4 and $\eta \perp (\varepsilon_c, \varepsilon_v)$, we have

$$G_v(t) - G_c(t) = \int_{-\infty}^{\infty} (F_v(t - \eta) - F_c(t - \eta))h(\eta)d\eta.$$

Setting $s = t - \eta$, we have

$$G_v(t) - G_c(t) = \int_{-\infty}^{\infty} (F_v(s) - F_c(s))h(t - s)ds.$$

Define $D(s) \equiv F_v(s) - F_c(s)$, and note that, because F_v and F_c are both symmetric, D is odd, i.e., $D(s) = -D(-s)$ for all s . We now have

$$G_v(t) - G_c(t) = \int_{-\infty}^{\infty} D(s)h(t - s)ds = \int_{-\infty}^0 D(s)h(t - s)ds + \int_0^{\infty} D(s)h(t - s)ds.$$

We can rewrite the first integral as

$$\int_{-\infty}^0 D(s)h(t - s)ds = \int_0^{\infty} D(-s)h(t + s)ds = \int_0^{\infty} -D(s)h(t + s)ds,$$

where the first equality follows from substituting $-s$ for s and the second equality

applies that D is odd. Hence, we have

$$G_v(t) - G_c(t) = \int_0^\infty D(s)(h(t-s) - h(t+s))ds.$$

Suppose $t > 0$. For any $s > 0$, symmetry and unimodality of h imply $h(t-s) - h(t+s) \geq 0$ because $|t-s| < t+s$. Moreover, the inequality is strict whenever $t-s$ lies in the interior of the support of h . Without loss of generality, we assume choices are noisier (Assumption 5). Then, for all $s > 0$, we have $D(s) \geq 0$ with strict inequality whenever $F_v(s) < 1$. Thus the integrand is non-negative. We now show that $D(s)(h(t-s) - h(t+s)) > 0$ on some interval of s . To this end, let $\bar{\eta} = \text{ess sup}(\eta)$ and let $\bar{\varepsilon}_v = \text{ess sup}(\varepsilon_v)$. Note that if $s \in (0, \bar{\varepsilon}_v)$, then $D(s) > 0$ and if $|t-s| < \bar{\eta}$, or equivalently, $s \in (t-\bar{\eta}, t+\bar{\eta})$, then $h(t-s) - h(t+s) > 0$. Since $G_v(t) = \mathbb{P}_v(y) \in (0, 1)$ by assumption and since η and ε_v are independent, $t < \bar{\varepsilon}_v + \bar{\eta}$. It follows that $(t-\bar{\eta}, t+\bar{\eta}) \cap (0, \bar{\varepsilon}_v)$ is a non-empty interval on which the integrand is positive. An analogous argument applies when $t < 0$, where $h(t-s) - h(t+s) \leq 0$. Thus, integrating over s , we have $G_v(t) - G_c(t) \neq 0$, or equivalently, $\mathbb{P}_c(y) \neq \mathbb{P}_v(y)$, whenever $y \neq \mu$ and $\mathbb{P}_c(y), \mathbb{P}_v(y) \in (0, 1)$. □

B.6 Proof of Proposition 5

Proof of Proposition 5. By independence of the choice and valuation noises, whenever the conditional probability is defined and since η and ε_v have densities by Assumption 4 and Assumption 7, we have

$$\mathbb{P}_c[y \mid y_{vi} = y] = \frac{\int F_c(y - \mu - \eta) f_v(y - \mu - \eta) h(\eta) d\eta}{\int f_v(y - \mu - \eta) h(\eta) d\eta} \quad (\text{B.4})$$

$$= \frac{1}{2} + \frac{\int O_c(y - \mu - \eta) f_v(y - \mu - \eta) h(\eta) d\eta}{\int f_v(y - \mu - \eta) h(\eta) d\eta}, \quad (\text{B.5})$$

where $O_c(x) = F_c(x) - \frac{1}{2}$, which is odd by the symmetric noises assumption.

Thus at $y = \mu$, we have

$$\mathbb{P}_c[\mu \mid y_{vi} = \mu] = \frac{1}{2} + \frac{\int O_c(-\eta) f_v(-\eta) h(\eta) d\eta}{\int f_v(-\eta) h(\eta) d\eta} = \frac{1}{2},$$

since the numerator of the second term is zero as O_c is odd, f_v is even by symmetric

noises, and h is even by symmetric preferences.

We next show that $\mathbb{P}_c[y \mid y_{vi} = y] = \frac{1}{2}$ only at μ . Define $t = y - \mu$ and $z = t - \eta$. With this change of variables, we have

$$\mathbb{P}_c[y \mid y_{vi} = y] - \frac{1}{2} = \frac{\int O_c(z) f_v(z) h(t - z) dz}{\int f_v(z) h(t - z) dz}. \quad (\text{B.6})$$

By symmetry, O_c is odd and f_v is even. Pairing z and $-z$, the numerator becomes

$$\int_0^\infty O_c(z) f_v(z) [h(t - z) - h(t + z)] dz. \quad (\text{B.7})$$

Suppose first that $t > 0$. Symmetry and unimodality of h imply

$$h(t - z) \geq h(t + z) \quad \text{for every } z > 0,$$

since $|t - z| < t + z$. Moreover, the inequality is strict whenever either term is positive. Positive noise density around zero together with symmetry of noise imply that $O_c(z) > 0$ for all $z > 0$. Hence the integrand in (B.7) is nonnegative.

The denominator in (B.6) is positive (which follows from $\mathbb{P}_c[y \mid y_{vi} = y]$ being well-defined), so $f_v(z) h(t - z) > 0$ for a positive-measure set of z . If this occurs on a positive-measure set of $z > 0$, then the inequality $h(t - z) > h(t + z)$ is strict there. If instead it occurs on a positive-measure set of $z < 0$, symmetry of f_v implies that on this same set $f_v(-z) h(t - z) > 0$. Reflecting this set, we conclude that for a positive-measure set of $z > 0$, $f_v(z) h(t + z) > 0$, which again implies $h(t - z) > h(t + z)$. Thus the integrand in (B.7) is strictly positive on a set of positive measure. Therefore

$$\mathbb{P}_c[y \mid y_{vi} = y] > \frac{1}{2}.$$

A symmetric argument applies in the case that $t < 0$, demonstrating that

$$\mathbb{P}_c[y \mid y_{vi} = y] < \frac{1}{2}.$$

□

C General Version of Model

In this appendix, we develop a general version of our model that accommodates lottery-lottery comparisons, the preference-reversal paradigm, and measurement noise added to utility rather than the response variable.

Suppose we want a person to compare two lotteries, X and Y . When these lotteries can be ranked in terms of risk, we let Y denote the safer lottery, but they need not be rankable.³⁶ Individual i has a metric Ω_{ti} for comparing lotteries in task type $t \in \{c, v\}$ —that is, if $\Omega_{ti}(W) > \Omega_{ti}(W')$, then $W \succ_{ti} W'$. This metric could be a certainty equivalent, a more general form of valuation (e.g., ψ such that $(\psi, p) \sim_{ti} W$), an expected utility, or some other form of utility; we do assume that this metric respects first-order stochastic dominance (i.e., if W FOSD W' then $\Omega_{ti}(W) > \Omega_{ti}(W')$). Importantly, decision noise will enter at the level of this metric.

For the choice task between X and Y , we assume that individual i chooses Y when $\Omega_{ci}(Y) + \varepsilon_c(Y) \geq \Omega_{ci}(X) + \varepsilon_c(X)$ for some shocks $\varepsilon_c(X)$ and $\varepsilon_c(Y)$. Defining $\varepsilon_c \equiv \varepsilon_c(X) - \varepsilon_c(Y)$, we have $\mathbb{P}_{ci}(Y) = \mathbb{P}(\varepsilon_c \leq \Omega_{ci}(Y) - \Omega_{ci}(X))$.

To represent a generalized version of a valuation task, we introduce a parameterized lottery $Z(\psi)$, where $\psi \in \mathbb{R}$ and $\psi > \psi'$ implies $Z(\psi)$ FOSD $Z(\psi')$. One example is $Z(\psi) \equiv (\psi, p)$. Given a $Z(\psi)$, a valuation task for lottery W asks person i to state the value ψ_{Wi} such that $Z(\psi_{Wi})$ is of equal value to them as lottery W . For such tasks, we assume the person states the value ψ_{Wi} such that $\Omega_{vi}(Z(\psi_{Wi})) = \Omega_{vi}(W) + \varepsilon_v(W)$ (which is unique because Ω_i respects first-order stochastic dominance). We also define ψ_{Wi}^* such that $\Omega_{vi}(Z(\psi_{Wi}^*)) = \Omega_{vi}(W)$, which is the person's indifference value according to their underlying preferences (analogous to y_i^* in Section 3).

Across our data sources, there are three relevant types of valuation tasks:

- *X Valuation Task*: The parameterized lottery $Z(\psi)$ is such that $Z(\psi_Y) = Y$ for some ψ_Y , and we elicit ψ_{Xi} such that $Z(\psi_{Xi})$ is of equal value as X . For this type, $\mathbb{P}_{vi}(Y) = \mathbb{P}(\psi_{Xi} \leq \psi_Y) = \mathbb{P}(\Omega_{vi}(X) + \varepsilon_v(X) \leq \Omega_{vi}(Z(\psi_Y))) = \mathbb{P}(\varepsilon_v(X) \leq \Omega_{vi}(Y) - \Omega_{vi}(X))$.
- *Y Valuation Task*: The parameterized lottery $Z(\psi)$ is such that $Z(\psi_X) = X$ for some ψ_X , and we elicit ψ_{Yi} such that $Z(\psi_{Yi})$ is of equal value as Y . For this

³⁶In fact, none of what follows requires the structure of lotteries. We focus on lotteries rather than abstract alternatives because they are the subject of our data.

type, $\mathbb{P}_{vi}(Y) = \mathbb{P}(\psi_{Yi} \geq \psi_X) = \mathbb{P}(\Omega_{vi}(Y) + \varepsilon_v(Y) \geq \Omega_{vi}(Z(\psi_X))) = \mathbb{P}(-\varepsilon_v(Y) \leq \Omega_{vi}(Y) - \Omega_{vi}(X))$.

- *XY Valuation Task:* Given a parameterized lottery $Z(\psi)$, we elicit both the ψ_{Xi} such that $Z(\psi_{Xi})$ is of equal value as X and the ψ_{Yi} such that $Z(\psi_{Yi})$ is of equal value as Y . For this type, $\mathbb{P}_{vi}(Y) = \mathbb{P}(\psi_{Yi} \geq \psi_{Xi}) = \mathbb{P}(\Omega_{vi}(Y) + \varepsilon_v(Y) \geq \Omega_{vi}(X) + \varepsilon_v(X)) = \mathbb{P}(\varepsilon_v(X) - \varepsilon_v(Y) \leq \Omega_{vi}(Y) - \Omega_{vi}(X))$.

For all three types, $\mathbb{P}_{vi}(Y) = \mathbb{P}(\varepsilon_v \leq \Omega_{vi}(Y) - \Omega_{vi}(X))$ for some appropriately defined ε_v . Hence, when we apply the null of stable preferences—that is, that $\Omega_{ci}(W) = \Omega_{vi}(W) = \Omega_i(W)$ for all W —the two key equations are:

$$\mathbb{P}_{ci}(Y) = \mathbb{P}(\varepsilon_c \leq \Omega_i(Y) - \Omega_i(X)) \quad \text{and} \quad \mathbb{P}_{vi}(Y) = \mathbb{P}(\varepsilon_v \leq \Omega_i(Y) - \Omega_i(X)). \quad (\text{C.1})$$

Note that (C.1) takes the same form as (1) except that $\Omega_i(Y) - \Omega_i(X)$ replaces $y - y_i^*$. Assumptions about decision noise have the same meaning as in the text, and the two main individual results go through as follows.³⁷

Remark C1. If noise is identical for choices and valuations, then $\mathcal{A}_i = \{(a, b) | a = b\}$.

Lemma C1. Under symmetric noises:

1. $\mathcal{A}_i = S_y \cup S_X$; and
2. If $\Omega_i(Y) > \Omega_i(X)$, then $(\mathbb{P}_{ci}(Y), \mathbb{P}_{vi}(Y)) \in S_y$, and if $\Omega_i(Y) < \Omega_i(X)$, then $(\mathbb{P}_{ci}(Y), \mathbb{P}_{vi}(Y)) \in S_X$; and
3. If $\Omega_i(Y) = \Omega_i(X)$, then $(\mathbb{P}_{ci}(Y), \mathbb{P}_{vi}(Y)) = (\frac{1}{2}, \frac{1}{2})$.

To model heterogeneity in preferences, we use the following structure:

- Heterogeneity in preferences: $\Omega_i(Y) - \Omega_i(X) = \Delta - \eta$, where $\Delta \in \mathbb{R}$ is the population mean preference, and η is a mean-zero random variable with CDF H .³⁸
- Heterogeneity in decision noise: The additive noise shocks ε_c and ε_v can be correlated with η .

³⁷The proofs for all results for the general model are essentially the same as the proofs for the special case considered in the text, and thus are omitted.

³⁸We assume that η enters negatively so that, as in the main text, a larger η implies a stronger preference for the riskier lottery X .

This structure is identical to the structure in the text except that Δ replaces $y - \mu$. Assumptions about the distribution of preferences have the same meaning as in the text, and the main aggregate results go through as follows.

Proposition C1. Under stable preferences and (conditionally) symmetric noises, $\mathcal{A} = \text{conv}(S_y \cup S_X)$.

Proposition C2. Under stable preferences and symmetric noises, if preferences are independent of noise, then $\mathcal{A} = \text{conv}(C_y \cup C_X) \cup \text{conv}(V_y \cup V_X)$.

Proposition C3. Under stable preferences and symmetric noises, if preferences are symmetric and independent of noise, then:

1. $\mathcal{A} = S_y \cup S_X$; and
2. If $\Delta > 0$, $(\mathbb{P}_c(Y), \mathbb{P}_v(Y)) \in S_y$, and if $\Delta < 0$, $(\mathbb{P}_c(Y), \mathbb{P}_v(Y)) \in S_X$; and
3. If $\Delta = 0$, then $(\mathbb{P}_{ci}(Y), \mathbb{P}_{vi}(Y)) = (\frac{1}{2}, \frac{1}{2})$.

Proposition C4. Under stable preferences and symmetric and ordered noises, if preferences are symmetric, unimodal, and independent of noise, then $\mathbb{P}_c(y) = \mathbb{P}_v(y) \in (0, 1)$ if and only if $\Delta = 0$, in which case $\mathbb{P}_c(y) = \mathbb{P}_v(y) = \frac{1}{2}$.

For an analogue to Proposition 5, we need a way of expressing that a person's response(s) on a valuation task suggest that X and Y are of equal value to them. If we define $\psi_{Yi} \equiv \psi_Y$ for X valuation tasks and $\psi_{Xi} \equiv \psi_X$ for Y valuation tasks, then for all three types of valuation tasks, the needed condition is $\psi_{Yi} = \psi_{Xi}$.

Proposition C5. Under stable preferences and symmetric noises, if preferences are symmetric, unimodal, and independent of noise, and the noise is independent with positive densities, then $\mathbb{P}_c(Y \mid \psi_{Yi} = \psi_{Xi}) = \frac{1}{2}$ if and only if $\Delta = 0$. Moreover, if $\Delta > 0$ then $\mathbb{P}_c(Y \mid \psi_{Yi} = \psi_{Xi}) > \frac{1}{2}$, and if $\Delta < 0$ then $\mathbb{P}_c(Y \mid \psi_{Yi} = \psi_{Xi}) < \frac{1}{2}$.

Finally, for our structural assessments of stability, we can do analogous assessments based on the following population distribution of underlying preferences:

$$\begin{pmatrix} \Omega_{ci}(Y) - \Omega_{ci}(X) \\ \Omega_{vi}(Y) - \Omega_{vi}(X) \end{pmatrix} \sim N \left(\begin{pmatrix} \Delta_c \\ \Delta_v \end{pmatrix}, \begin{pmatrix} \gamma_c^2 & \gamma_{cv} \\ \gamma_{cv} & \gamma_v^2 \end{pmatrix} \right).$$

C.1 Translation to Specific Cases

Given the general version of the model, we can now discuss translations to specific cases, along with a few comments where relevant.

Lottery vs. certain amount, noise at level of certainty equivalents: This is the case from the main text where we assume $Y = (y, 1)$ and use X valuations. If the metric Ω_i is the certainty equivalent of a lottery, then clearly $\Omega_i(Y) = y$. If we further define y_i^* to be the certainty equivalent for X , then $\Omega_i(Y) - \Omega_i(X) = y - y_i^*$, and then the general results above translate directly to the results in the main text.

Lottery vs. lottery, noise at the level of valuations response variable: Two types of lottery-lottery comparisons appear in our data sources.

First, we might compare lotteries X and Y using X valuations. If the metric Ω_i is in terms of the valuation response variable ψ_{Xi} , then $\Omega_i(X) = \psi_{Xi}^*$ and $\Omega_i(Y) = \psi_Y$. If we then define $y = \psi_Y$ and $y_i^* = \psi_{Xi}^*$, then $\Omega_i(Y) - \Omega_i(X) = y - y_i^*$.

Second, we might compare lotteries X and Y using Y valuations. If the metric Ω_i is in terms of the valuation response variable ψ_{Yi} , then $\Omega_i(X) = \psi_X$ and $\Omega_i(Y) = \psi_{Yi}^*$. If we then define $y = -\psi_X$ and $y_i^* = -\psi_{Yi}^*$, then $\Omega_i(Y) - \Omega_i(X) = y - y_i^*$.

In either case, the general results above translate directly to the results in the main text. We note that this formulation is consistent with how we analyze lottery-lottery comparisons in Section 4.

Preference-reversal paradigm, noise at level of certainty equivalents: In the preference-reversals paradigm, we compare lotteries X and Y using XY valuations. Suppose the metric Ω_i is in terms of the certainty equivalent of a lottery, where we use $\text{CE}_i(W)$ to denote i 's certainty equivalent of lottery W . Then $\Omega_i(Y) - \Omega_i(X) = \text{CE}_i(Y) - \text{CE}_i(X)$. Hence, the general results hold with $\text{CE}_i(Y) - \text{CE}_i(X)$ in place of $\Omega_i(Y) - \Omega_i(X)$.

Lottery vs. lottery, noise at level of expected utility: Suppose the metric Ω_i is in terms of an expected-utility functional, where we use $\text{EU}_i(W)$ to denote i 's expected utility from lottery W . Because $\Omega_i(Y) - \Omega_i(X) = \text{EU}_i(Y) - \text{EU}_i(X)$, the general results hold with $\text{EU}_i(Y) - \text{EU}_i(X)$ in place of $\Omega_i(Y) - \Omega_i(X)$.³⁹

³⁹If individuals use some other form of utility functional (e.g., the cumulative prospect theory functional from [Tversky and Kahneman \(1992\)](#)), things would be analogous.

When decision noise enters at the level of expected utility, the assumption of symmetric preferences—that is, that the population distribution of $\text{EU}_i(Y) - \text{EU}_i(X)$ is symmetric—may be less appealing. In particular, a more natural assumption is symmetry on some underlying preference parameter, for instance, an assumption that $u(z; \eta) = z^{\Delta+\eta}$ with η symmetric and mean-zero. However, such an assumption would not generate symmetry in $\text{EU}_i(Y) - \text{EU}_i(X)$, and indeed the following example demonstrates that Proposition 3 does not hold.

Example: Suppose $X \equiv (100, 0.3)$ and $Y = (12, 1)$. Suppose $\Delta = 0.5$, in which case a person with the mean preference prefers Y to X . Suppose η is equally likely to be $-0.2, -0.1, 0.1$, and 0.2 . Then:

- For $\eta = -0.2$, $\text{EU}(Y; \eta) - \text{EU}(X; \eta) = 12^{0.3} - (0.3)100^{0.3} = 0.91$
- For $\eta = -0.1$, $\text{EU}(Y; \eta) - \text{EU}(X; \eta) = 12^{0.4} - (0.3)100^{0.4} = 0.81$.
- For $\eta = 0.1$, $\text{EU}(Y; \eta) - \text{EU}(X; \eta) = 12^{0.6} - (0.3)100^{0.6} = -0.31$.
- For $\eta = 0.2$, $\text{EU}(Y; \eta) - \text{EU}(X; \eta) = 12^{0.7} - (0.3)100^{0.7} = -1.84$.

Now suppose F_c is 50-50 $\bar{c}$ and $-\bar{c}$, and F_v is 50-50 $\bar{v}$ and $-\bar{v}$. If $\bar{c} \in (0.31, 0.81)$, then the choice probabilities for the four types are 1, 1, $\frac{1}{2}$, and 0, yielding an average of $\mathbb{P}_c(y) = 5/8$. If at the same time $\bar{v} \in (0.91, 1.84)$, then the valuation probabilities for the four types are $\frac{1}{2}$, $\frac{1}{2}$, $\frac{1}{2}$, and 0, yielding an average of $\mathbb{P}_v(y) = 3/8$. Hence, $(\mathbb{P}_c(y), \mathbb{P}_v(y)) = (5/8, 3/8)$, which is outside $S_y \cup S_X$.

D Details of Dataset Construction (for Section 4.1)

D.1 Data from the Preference-Reversal Literature

To construct a dataset based on preference-reversal literature, we draw from a recent meta-analysis by Lu (2026). In Table 2 from the Online Supplementary Material, there are 462 experiments from 54 studies for which one could calculate the empirical quantities $(\widehat{\mathbb{P}}_c(y), \widehat{\mathbb{P}}_v(y))$.⁴⁰

⁴⁰Note that Lu (2026)’s Table 2 also includes a number of additional entries for studies that report only the proportions of the two types of preference reversals. Because it is not possible to construct $(\widehat{\mathbb{P}}_c(y), \widehat{\mathbb{P}}_v(y))$ from only these proportions, we do not consider these studies.

We then hired a research assistant to reconstruct the data ourselves. Specifically, the research assistant collected the 54 original studies. There were three outcomes:

- For 33 studies, the research assistant was able to confirm the numbers from [Lu \(2026\)](#).
- For 9 studies, the research assistant found discrepancies between the numbers from [Lu \(2026\)](#) and the numbers in the original study; for our dataset, we use the corrected numbers.
- For 12 studies, the research assistant was unable to construct the needed numbers from the original study; for our dataset, we excluded these studies.⁴¹

In the end, our final dataset consists of 314 experiments from 42 studies, and that is the data that we use here. In line with the literature, for each experiment, we take the P -bet to be the safer option. We highlight three further details.

First, a number of the original studies report only the combined numbers across multiple conditions—i.e., across multiple “experiments”. For instance, in the seminal work of [Grether and Plott \(1979\)](#), each experiment covers six different pairs of bets (summarized in their Table 2), but they report decisions across all six pairs (in their Tables 5, 6, 8, and 9). Rather than exclude these, we chose to consider the combined numbers to be one experiment.

Second, for a number of the original studies, there is some ambiguity in how to define an experiment. So as not to give ourselves extra flexibility, we use the decisions from [Lu \(2026\)](#). As an example, Experiment 1 from [Grether and Plott \(1979\)](#) splits subjects into one group that does not have incentives and a second group that does have incentives. In principle, one could consider the two groups to be different experiments, or one could combine the behavior of the two groups into one experiment. [Lu \(2026\)](#) chose to do the former, and thus so do we.

Third, some studies permit participants to report indifference. For such studies, we include responses of indifference in the total count of observations when calculating $\widehat{\mathbb{P}}_c(y)$ or $\widehat{\mathbb{P}}_v(y)$. For instance, if in a choice task, 30 choose the safe option, 10 choose the risky option, and 10 express indifference, we would calculate $\widehat{\mathbb{P}}_c(y) = 0.60$.

⁴¹[Lu \(2026\)](#) comments that some numbers were obtained from personal correspondence with the original authors.

D.2 Data from the Choices-versus-MPLs Literature

To construct data based on the choices-versus-MPLs literature, we pull the needed data ourselves from the three key studies in this literature (Brown and Healy, 2018; Freeman et al., 2019; Freeman and Mayraz, 2019). Specifically:

- [Brown and Healy \(2018\)](#): In this study, we identified 1 experiment. Specifically, we use the results reported in their Table 3 for their Experiment 1, where we use O-14 as the choice treatment and L-RPS as the valuations treatment. For the choice experiment, $N = 61$ and $\widehat{\mathbb{P}}_c(y) = 0.44$; for the valuation experiment, $N = 60$ and $\widehat{\mathbb{P}}_v(y) = 0.48$. Their Experiment 2 does not include a choice treatment, and thus it does not yield an experiment for us.
- [Freeman et al. \(2019\)](#): In this study, we identified 2 experiments. The study includes some minor treatment differences in terms of number of questions asked and how payments are made. However, since the authors ignore these differences in presenting their headline results in their Figure 1, we focus on the two experiments represented in that figure. For the Q1 experiment, the choice treatment has $N = 81$ and $\widehat{\mathbb{P}}_c(y) = 0.77$, while the valuation treatment has $N = 355$ and $\widehat{\mathbb{P}}_v(y) = 0.55$. For the Q2 experiment, the choice treatment has $N = 83$ and $\widehat{\mathbb{P}}_c(y) = 0.70$, while the valuation treatment has $N = 354$ and $\widehat{\mathbb{P}}_v(y) = 0.62$.
- [Freeman and Mayraz \(2019\)](#): In this study, we identified 6 experiments, one associated with each of the six different single-choice treatments (SC-Q95, SC-Q90, SC-Q85, SC-Q80, SC-Q75, and SC-Allais). For the valuations treatment, we chose to use the combined R-List treatments. We then pull the proportions from their Table 3—e.g., for Q85, $\widehat{\mathbb{P}}_c(y) = 0.77$ and $\widehat{\mathbb{P}}_v(y) = 0.59$ (because the table reports the proportion choosing risky, these are one minus the numbers reported in the table). The paper does not report precise sample sizes for each of the thirteen treatments, but each has an average of 120. Hence, roughly, each choice experiment has $N = 120$, one valuation experiment (R-List Allais) has $N = 120$, and the other five valuation experiments each have $N = 360$ (as each combines three treatments).

For all three studies, we follow the authors' labels for which lottery is riskier.

D.3 Data from [McGranaghan et al. \(2024, 2026\)](#)

D.3.1 Overview of Studies

The primary focus of [McGranaghan et al. \(2024\)](#) and [McGranaghan et al. \(2026\)](#) is to better understand how people react to variants of the common-ratio (CR) and common-consequence (CC) manipulations proposed by [Allais \(1953\)](#). [McGranaghan et al. \(2024\)](#) highlight an inference problem when using paired-choice tasks to study CR problems, and propose using paired-valuation tasks instead to overcome this problem. [McGranaghan et al. \(2026\)](#) highlight how there has been limited exploration of the parameter space for both CR and CC problems, and they conduct such exploration using both paired-valuation tasks and paired-choice tasks. While neither study focuses directly on comparing decisions for choices versus valuations, it turns out their data is ideally situated for doing so.

These studies ask subjects to make three types of comparisons for risky lotteries:

AB Comparison: Safer $A \equiv (M, 1)$ *vs.* Riskier $B \equiv (H, p)$

AB' Comparison: Safer $A \equiv (M, 1)$ *vs.* Riskier $B' \equiv (H, pr; M, 1 - r)$

CD Comparison: Safer $C \equiv (M, r)$ *vs.* Riskier $D \equiv (H, pr)$

These studies use two types of valuation tasks that differ in terms of the scalar value that is elicited (both types are implemented using a multiple-price list). In an *m*-valuation, they fix (H, p, r) and ask subjects to state a value m_{vi} that makes two options of equal value. In an *h*-valuation, they fix (M, p, r) and ask subjects to state a value h_{vi} that makes two options of equal value. For instance, for a *CD* comparison with $p = 0.5$ and $r = 0.4$, the two tasks might be:

m-Valuation: State an m_{vi} such that $(m_{vi}, 0.4)$ is of equal value to $(30, 0.2)$.

h-Valuation: State an h_{vi} such that $(15, 0.4)$ is of equal value to $(h_{vi}, 0.2)$.

As we describe below, across the two studies, *h*-valuations are used for all three types of comparisons, while *m*-valuations are used only for *AB* and *CD* comparisons. It follows that *m*-valuations always elicit a value within the safer option, and thus are analogous to the *X* valuation tasks defined in [Appendix C](#), while *h*-valuations always elicit a value within the riskier option, and thus are analogous to the *Y* valuation tasks defined in [Appendix C](#).

These studies also present the same subjects with linked choice tasks. Specifically, if a subject faces an m -valuation for a particular (H, p, r) , then they also face one choice task that uses that same (H, p, r) along with a specific m_c that is randomly selected from a set of values. Analogously, if a subject faces an h -valuation for a particular (M, p, r) , then they also face one choice task that uses that same (M, p, r) along with a specific h_c that is randomly selected from a set of values. The sets of values were chosen based on pilot data to (ideally) span both sides of the mean preference.⁴²

The primary analysis in [McGranaghan et al. \(2024\)](#) studies paired-valuation tasks. Specifically, they study CR problems by comparing an AB and a CD valuation for the same (p, r) , and they do so using both m -valuations and h -valuations. The primary analysis in [McGranaghan et al. \(2026\)](#) also studies paired-valuation tasks. Specifically, they study CR problems by comparing an AB and a CD valuation for the same (p, r) , CC problems by comparing an AB' and a CD valuation for the same (p, r) , and also preferences for probabilistic mixtures by comparing an AB and a AB' valuation for the same (p, r) . They do so using only h -valuations. Each study use the data on choice tasks to support their conclusions.

D.3.2 Construction of Data for Our Purposes

Again, neither study focuses directly on comparing decisions for choices versus valuations, but it turns out their data is ideally situated for doing so.

[McGranaghan et al. \(2024\)](#) collects data on AB and CD decisions for five different values of $p \in \{0.1, 0.2, 0.5, 0.8, 0.9\}$ and three different values of $r \in \{0.2, 0.4, 0.6\}$ for both m -valuations and h -valuations. Since r is irrelevant for AB decisions, they collect a total of 10 AB valuations (five m -valuations and five h -valuations). Since r matters for CD decisions, they collect a total of 30 CD valuations (15 m -valuations and 15 h -valuations). For each of these 40 valuations, they collect data for linked choices for four different values of M or H . Hence, this study yields data on 40 groupings with 4 experiments within each group, and thus a total of 160 experiments. For the AB decisions, there is an average of 225 subjects per experiment and thus 900 subjects per grouping. For the CD decisions, there is an average of 75 subjects per experiment

⁴²Analogous to the choices-versus-MPLs literature, a choice task corresponds to one row of the multiple-price list for the linked valuation task. But unlike the choices-versus-MPLs literature, the comparison can be done within-subjects.

and thus 300 subjects per grouping.

McGranaghan et al. (2026) collects data on AB , AB' , and CD decisions for four different values of $p \in \{0.3, 0.5, 0.8, 0.9\}$ and five different values of $r \in \{0.1, 0.2, 0.3, 0.5, 0.8\}$ using only h -valuations. Again, since r is irrelevant for AB decisions, they collect a total of 4 AB valuations. Since r matters for AB' and CD decisions, they collect a total of 20 AB' valuations and 20 CD valuations. For each of these 44 valuations, they collect data for linked choices for six different values of H . Hence, this study yields data on 44 groupings with 6 experiments within each group, and thus a total of 264 experiments. For the AB decisions, there is an average of 701 subjects per experiment and thus 4204 subjects per grouping. For the AB' and CD decisions, there is an average of 140 subjects per experiment and thus 841 subjects per grouping.

Combining the data from both studies, there is a total of 424 experiments across 84 groupings. In total, there are 68,448 observations for both choices and valuations, and thus an average of 161 subjects per experiment 815 subjects per grouping.

D.3.3 Harmonizing m -Tasks and h -Tasks

In an m -valuation, a larger m_{vi} reflects a stronger preference for the riskier option, whereas in an h -valuation, a larger h_{vi} reflects a stronger preference for the safer option. Analogously, for a choice task linked to an m -valuation, a larger selected value for M makes the safer option more attractive, whereas for a choice task linked to an h -valuation, a larger selected value for H makes the riskier option more attractive. Given this asymmetry, we must code the data in a way that harmonizes m -tasks and h -tasks.

When calculating $\widehat{\mathbb{P}}_c(y)$ and $\widehat{\mathbb{P}}_v(y)$ for a particular experiment, this harmonization is straightforward. Throughout, we have defined these as the proportion that select the safer alternative, and it is clear that the safer alternative is A in an AB task, A in an AB' task, and C in a CD task. Hence, for each experiment, we merely define $\widehat{\mathbb{P}}_c(y)$ to be the proportion of subjects choosing the safer alternative and $\widehat{\mathbb{P}}_v(y)$ to be to the proportion of subjects having a valuation that would indicate a preference for the safer alternative.

When using quantitative information on valuations, as needed for the tests from Sections 3.3 and 3.4, we must be a little more careful. In our theory, a larger y_{vi} means the person has a stronger preference for the risky option, and a larger y implies the safer option is more attractive. m -tasks match this logic, and thus we can use data

from m -tasks at face value. In other words, we can merely set $y_{vi} = m_{vi}$ and $y = m_c$, because the equations from our theory apply directly when $y - \overline{y_{vi}} = m_c - \overline{m_{vi}}$ and $y - y_{vi} = m_c - m_{vi}$. And for the MLE, as in our theory, a person chooses the safer option when m_c is larger than y_{ci} .

For h -tasks, however, the logic is inverted, and thus we must correct for this. To do so, we set $y - \overline{y_{vi}} = \overline{h_{vi}} - h_c$ and $y - y_{vi} = h_{vi} - h_c$, and then apply the equations from our theory. And for the MLE, a person chooses the safer option when h_c is smaller than y_{ci} .

E Details of Maximum Likelihood Estimation

We conduct our MLE approach on each of the 84 groupings defined in Section 4.1. However, we must be careful to harmonize the estimation for m -valuations, where a larger valuation indicates less risk aversion, and h -valuations, where a larger valuation indicates more risk aversion (see Appendix D.3 for details on these).

For either type of valuation task, we let w denote the object that is elicited (so $w \in \{m, h\}$). We then assume, analogous to equation 3, that a person's choice and valuation are determined from w_{ci} and w_{vi} , where

$$\begin{pmatrix} w_{ci} \\ w_{vi} \end{pmatrix} \sim N \left(\begin{pmatrix} \mu_c \\ \mu_v \end{pmatrix}, \begin{pmatrix} \sigma_c^2 + \gamma_c^2 & \gamma_{cv} \\ \gamma_{cv} & \sigma_v^2 + \gamma_v^2 \end{pmatrix} \right). \quad (\text{E.1})$$

Within a grouping, each individual faces the valuation task plus one of the (four or six) choice tasks. Hence, individual i 's data comes in the form $(w_c, a_{ci}, w_{vi}^L, w_{vi}^H)$, where

- w_c is the value randomly assigned to i for their choice task.
- $a_{ci} \in \{0, 1\}$ is i 's chosen option in that choice task, where $a_{ci} = 1$ denotes choosing the safer option and $a_{ci} = 0$ denotes choosing the riskier option.
- w_{vi}^L and w_{vi}^H reflect the values of i 's switching rows in the valuation task—that is, in the price list, they choose the fixed option for rows with $w \leq w_{vi}^L$, and they choose the varying option for rows with $w \geq w_{vi}^H$. Note that if the person always chooses the varying option, we code $w_{vi}^L = -\infty$; and if the person always chooses the fixed option, we code $w_{vi}^H = \infty$.

Given this structure, it is straightforward to derive an individual's likelihood function, although we must do it separately for m versus h valuations. For m valuations, we use:⁴³

$$L_i(m_c, 0, m_{vi}^L, m_{vi}^H) = N \left(\begin{pmatrix} m_c, \infty \\ m_{vi}^L, m_{vi}^H \end{pmatrix}; \begin{pmatrix} \mu_c \\ \mu_v \end{pmatrix}, \begin{pmatrix} \sigma_c^2 + \gamma_c^2 & \gamma_{cv} \\ \gamma_{cv} & \sigma_v^2 + \gamma_v^2 \end{pmatrix} \right), \text{ and}$$

$$L_i(m_c, 1, m_{vi}^L, m_{vi}^H) = N \left(\begin{pmatrix} -\infty, m_c \\ m_{vi}^L, m_{vi}^H \end{pmatrix}; \begin{pmatrix} \mu_c \\ \mu_v \end{pmatrix}, \begin{pmatrix} \sigma_c^2 + \gamma_c^2 & \gamma_{cv} \\ \gamma_{cv} & \sigma_v^2 + \gamma_v^2 \end{pmatrix} \right).$$

For h valuations, we instead use:

$$L_i(h_c, 0, h_{vi}^L, h_{vi}^H) = N \left(\begin{pmatrix} -\infty, h_c \\ h_{vi}^L, h_{vi}^H \end{pmatrix}; \begin{pmatrix} \mu_c \\ \mu_v \end{pmatrix}, \begin{pmatrix} \sigma_c^2 + \gamma_c^2 & \gamma_{cv} \\ \gamma_{cv} & \sigma_v^2 + \gamma_v^2 \end{pmatrix} \right), \text{ and}$$

$$L_i(h_c, 1, h_{vi}^L, h_{vi}^H) = N \left(\begin{pmatrix} h_c, \infty \\ h_{vi}^L, h_{vi}^H \end{pmatrix}; \begin{pmatrix} \mu_c \\ \mu_v \end{pmatrix}, \begin{pmatrix} \sigma_c^2 + \gamma_c^2 & \gamma_{cv} \\ \gamma_{cv} & \sigma_v^2 + \gamma_v^2 \end{pmatrix} \right).$$

Combining all individuals in the grouping, the aggregate log likelihood is:

$$L = \sum_i \log L_i(w_c, a_{ci}, w_{vi}^L, w_{vi}^H).$$

Maximizing the aggregate log likelihood with respect to the parameters μ_c , μ_v , $(\sigma_c^2 + \gamma_c^2)$, $(\sigma_v^2 + \gamma_v^2)$, and γ_{cv} is a straightforward exercise yielding estimates of the corresponding parameters using conventional techniques. In our implementation, for each of the 84 groupings, we use the conjoint gradient algorithm with a maximum of 1000 iterations. If convergence is not achieved within 1000 iterations, we take the final values as our estimates.⁴⁴

Note that the raw estimates in Table 2 and Appendix Table A.1 do not distinguish between m -valuations and h -valuations, and more generally are not really comparable across groups. In interpreting these estimates, we are more interested in their difference, as that reflects which elicitation technique is more likely to induce risk aversion. For that, however, we must account for the fact that larger values have opposite meanings

⁴³The function N represents the CDF for a bivariate normal distribution.

⁴⁴Of our 84 MLE implementations, only 8 fail to achieve convergence within 1000 iterations.

for m -valuations versus h -valuations. Given raw estimates $\hat{\mu}_v$ and $\hat{\mu}_c$, we define

$$\widehat{\mu_v - \mu_c} = \begin{cases} \hat{\mu}_v - \hat{\mu}_c & \text{if it is an } m\text{-valuation} \\ \hat{\mu}_c - \hat{\mu}_v & \text{if it is an } h\text{-valuation} \end{cases}$$

Note that $\widehat{\mu_v - \mu_c}$ is defined such that a positive value means that choices induce more risk aversion than valuations for either type of valuation task. When describing the implications of the MLE estimates in the lower panel of Table 2 and Appendix Table A.1 and in panel B of Figure 6, we report results in terms of $\widehat{\mu_v - \mu_c}$.

F Problem Parameters and Estimates of Instability

The 84 probit and maximum likelihood estimates presented in Table 2 and Table A.1 reveal that different broad conditions lead to different quantitative estimates of mean preference instability, $\widehat{\mu_v - \mu_c}$. For AB problems, the tendency is towards $\widehat{\mu_v - \mu_c} > 0$ and thus choices yield greater risk aversion; for AB' problems, the tendency is towards $\widehat{\mu_v - \mu_c} < 0$ and thus valuations yield greater risk aversion; and for CD problems, the tendency is towards $\widehat{\mu_v - \mu_c} \approx 0$ (particularly when expressed in expected value units).

Importantly, there is also substantial heterogeneity within problem types, with each of AB , AB' , and CD problems indicating some instances of $\widehat{\mu_v - \mu_c} > 0$ and some instances of $\widehat{\mu_v - \mu_c} < 0$. Understanding the drivers of this heterogeneity may lead to further insights on when preference stability is more or less likely to hold.

To analyze the sources of this within-problem heterogeneity, in Table F.1 we regress probit estimates and MLE estimates of $\widehat{\mu_v - \mu_c}$ on the probability parameters p and r controlling for whether the corresponding grouping covers an AB , AB' , or CD problem, and whether the grouping involves m -valuation tasks or h -valuation tasks. Columns (1) and (3) examine p effects for all 84 groupings. Because r is not defined for AB problems, columns (2) and (4) examine r effects for the 70 groupings which cover AB' and CD problems. We find that both p and r are highly and differentially correlated with estimates of instability. Choices yield greater risk aversion than valuations at low values of p , and less risk aversion than valuations at high values of p . In contrast, choices yield less risk aversion than valuations at low values of r , and greater risk aversion than valuations at high values of r .

Table F.1: Problem Parameters and Estimates of Instability

Probit Estimates			MLE Estimates		
Dependent Variable :	$\widehat{\mu_v - \mu_c}$		Dependent Variable :	$\widehat{\mu_v - \mu_c}$	
	(1)	(2)		(3)	(4)
p	-21.72		p	-20.82	
	(3.89)			(2.61)	
r		26.57	r		15.79
		(6.24)			(4.93)
$\mathbf{1}(AB')$	-16.71		$\mathbf{1}(AB')$	-20.77	
	(3.64)			(2.44)	
$\mathbf{1}(CD)$	-3.03	14.61	$\mathbf{1}(CD)$	-4.48	17.12
	(3.04)	(3.16)		(2.04)	(2.50)
$\mathbf{1}(m\text{-valuation})$	-4.91	-3.63	$\mathbf{1}(m\text{-valuation})$	-3.16	-1.35
	(2.73)	(3.48)		(1.83)	(2.75)
Constant	18.91	-21.48	Constant	17.70	-22.08
	(3.59)	(3.46)		(2.41)	(2.74)
# Observations	84	70	# Observations	84	70
R ²	0.45	0.38	R ²	0.69	0.48

Notes: Table presents OLS regressions of either probit or MLE estimated value of $\widehat{\mu_v - \mu_c}$ on problem parameters p and r controlling additionally for whether the grouping covers an AB , AB' , or CD decision and whether the grouping's valuations elicit a middle prize (m -valuation) or a high prize. Columns (1) and (3) focus on all 84 groupings and examines effect of p ; columns (2) and (4) focus on the 70 groupings for AB' and CD decisions and examines effect of r . Standard errors in parentheses.

Findings of preference instability may be quite sensitive not only to the type of lotteries used (e.g., certain vs. binary compared to certain vs. trinary compared to binary vs binary), but also to the probabilities implemented. Understanding the deeper reasons for these predictable effects on instability will be a critical path for future studies.

G Testing Sign Congruence (Miller et al., 2026)

In Section 4.3, we deploy the test recommended by Miller et al. (2026) for sign congruence between two estimators.⁴⁵ In this section, we provide more detail.

Miller et al. (2026) develop a methodology for testing the sign congruence between two theoretical quantities, μ_1 and μ_2 , that is, for testing the null $H_0 : \mu_1 \cdot \mu_2 \geq 0$. The test requires estimators $\hat{\mu}_1$ and $\hat{\mu}_2$ that satisfy the following assumption:

Assumption G1. The estimators $\hat{\mu}_1$ and $\hat{\mu}_2$ are asymptotically normal

$$\sqrt{n} \begin{pmatrix} \hat{\mu}_1 - \mu_1 \\ \hat{\mu}_2 - \mu_2 \end{pmatrix} \rightarrow_d N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right),$$

with convergent variance-covariance estimates,

$$\frac{\hat{\sigma}_1}{\sigma_1} \rightarrow_p 1, \quad \frac{\hat{\sigma}_2}{\sigma_2} \rightarrow_p 1, \quad \frac{\hat{\rho}}{\rho} \rightarrow_p 1.$$

Their recommended test rejects H_0 if

$$\hat{\mu}_1 \cdot \hat{\mu}_2 < 0 \quad \text{and} \quad \min \left\{ \frac{|\hat{\mu}_1|}{\hat{\sigma}_1/\sqrt{n}}, \frac{|\hat{\mu}_2|}{\hat{\sigma}_2/\sqrt{n}} \right\} \geq \bar{z}_\alpha.$$

The critical value is $\bar{z}_\alpha = \Phi^{-1}(1 - \alpha)$ when it is known that $\rho \geq 0$; otherwise, the critical value depends on ρ . For $\alpha = 0.05$, the critical threshold is $\bar{z}_\alpha = 1.645$ when $\rho \geq 0$; it remains approximately $\bar{z}_\alpha = 1.645$ until ρ is around -0.8 ; and then it reaches $\bar{z}_\alpha = 1.96$ when $\rho = -1$. Russek-Cohen and Simon (1993) provide a tabulation of the relevant thresholds for $\alpha = 0.05$ depending on the correlation. Where necessary (and feasible) we use this tabulation to target the threshold for $\alpha = 0.05$.

Under the assumptions of Proposition 3, the attainable set satisfies $(\mathbb{P}_c(y) - 0.5) \cdot (\mathbb{P}_v(y) - 0.5) \geq 0$. We can thus directly apply the Miller et al. (2026) test if we maintain Assumption G1 for the relevant empirical quantities—that is, for the sample proportions $\widehat{\mathbb{P}}_c(y)$ and $\widehat{\mathbb{P}}_v(y)$ and for their sample standard errors $\sqrt{\frac{\widehat{\mathbb{P}}_c(y)(1-\widehat{\mathbb{P}}_c(y))}{N_c}}$ and $\sqrt{\frac{\widehat{\mathbb{P}}_v(y)(1-\widehat{\mathbb{P}}_v(y))}{N_v}}$. Because the data from the prior literature do not contain information on the sample correlation, we apply the test assuming $\rho = 0$ (and thus use $\bar{z}_\alpha = 1.645$)

⁴⁵This test was originally proposed in the statistics literature by Russek-Cohen and Simon (1993).

for both of our datasets.⁴⁶

Under the assumptions of Proposition 2, the attainable set is $\text{conv}(C_y \cup C_X) \cup \text{conv}(V_y \cup V_X)$. We can test this prediction using an adjustment to the basic test. Miller et al. (2026) note that more general hypotheses akin to sign congruence can be tested by considering a change of base. We apply a base change to $\mathbb{P}_c(y)$ and $\mathbb{P}_v(y)$ using the cone spanned by the vectors $\{(4/3, 2/3), (2/3, 4/3)\}$. Define:

$$A \equiv \begin{pmatrix} 4/3 & 2/3 \\ 2/3 & 4/3 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix} \equiv A \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} \mu_1 - \frac{1}{2}\mu_2 \\ -\frac{1}{2}\mu_1 + \mu_2 \end{pmatrix}.$$

If $\mu_1 = \mathbb{P}_c(y) - 2/3$ and $\mu_2 = \mathbb{P}_v(y) - 1/3$ then, $\nu_1 = \mathbb{P}_c(y) - \frac{1}{2} - \frac{1}{2}\mathbb{P}_v(y)$ and $\nu_2 = -\frac{1}{2}\mathbb{P}_c(y) + \mathbb{P}_v(y)$. Thus, $\nu_1 = 0$ is equivalent to $\mathbb{P}_v(y) = 2\mathbb{P}_c(y) - 1$, the lower boundary of $\text{conv}(C_y \cup C_X)$. Similarly, $\nu_2 = 0$ is equivalent to $\mathbb{P}_v(y) = \frac{1}{2}\mathbb{P}_c(y)$, the lower boundary of $\text{conv}(V_y \cup V_X)$. With such definitions for μ_1 and μ_2 , an observation lies below both lines if $\nu_1 > 0$ and $\nu_2 < 0$. Defining $\mu'_1 = \mathbb{P}_c(y) - 1/3$ and $\mu'_2 = \mathbb{P}_v(y) - 2/3$ with the same A matrix delivers $\nu'_1 = 0$ and $\nu'_2 = 0$ being equivalent to the upper boundaries of $\text{conv}(C_y \cup C_X)$ and $\text{conv}(V_y \cup V_X)$. An observation lies above both lines if $\nu'_1 < 0$ and $\nu'_2 > 0$.

Miller et al. (2026) show that asymptotic normality of (ν_1, ν_2) with an arbitrary base change follows from asymptotic normality of (μ_1, μ_2) and thus the same recommended test applies. However, in our case the specific change of base induces negative correlation between ν_1 and ν_2 . Because, again, the data from the prior literature do not contain information on the sample correlation, we proceed under an assumption that $\mathbb{P}_c(y)$ and $\mathbb{P}_v(y)$ are uncorrelated. Under this assumption,

$$\text{corr}(\nu_1, \nu_2) = \text{corr}(\nu'_1, \nu'_2) = \frac{-(\sigma_1^2 + \sigma_2^2)/2}{\sqrt{\sigma_1^2 + 1/4\sigma_2^2}\sqrt{1/4\sigma_1^2 + \sigma_2^2}} < 0.$$

This negative correlation can be sufficiently large in magnitude to require adjustment of the critical value associated with $\alpha = 0.05$. When we apply this test to assess whether an observation comes from $\text{conv}(C_y \cup C_X) \cup \text{conv}(V_y \cup V_X)$, we calculate the sample analog for this correlation and use it to identify the corresponding threshold for $\bar{z}_\alpha$ from Russek-Cohen and Simon (1993) to target $\alpha = 0.05$.

⁴⁶This simplification appears largely justified: In the McGranaghan et al. (2024, 2026) data, 414 of the 424 experiments exhibit positive correlation, and no correlation is smaller than -0.19 .