Ever Since Allais*

Aluma Dembo, Shachar Kariv, Matthew Polisson, and John K.-H. Quah†

April 25, 2024

The Allais critique of expected utility theory (EUT) has led to the development of theories of choice under risk that relax the independence axiom but adhere to the fundamental/conventional axioms of ordering (completeness and transitivity) and monotonicity (with respect to first-order stochastic dominance). Unlike experimental work designed to test independence, our experiment is comprehensive—testing the entire set of axioms on which EUT is based. Our econometric analysis is also nonparametric and performed at the level of each individual subject. For the vast majority of subjects departures from independence are small relative to departures from ordering and/or monotonicity.

JEL Codes: D81, C91.

Keywords: revealed preference, rationality, ordering, completeness, transitivity, monotonicity, first-order stochastic dominance, independence, expected utility, non-expected utility, experiment.

*We are grateful to David Dillenberger, Federico Echenique, David Freeman, Georgios Gerasimou, Yoram Halevy, Joshua Lanier, Paola Manzini, Marco Mariotti, Yusufcan Masatlioglu, Peter Wakker, William Zame, and Lanny Zrill for discussions and comments, and to a number of seminar audiences for suggestions. The experiments reported in this paper were conducted in the Experimental Social Science Laboratory (Xlab) at UC Berkeley and the California Social Science Experimental Laboratory (CASSEL) at UCLA. This research has made use of the ALICE High Performance Computing Facility at the University of Leicester and BlueCrystal/BluePebble High Performance Computing at the University of Bristol. Financial support was provided by the National Science Foundation (Grant No. SES-0962543) and the British Academy/Leverhulme Trust (Grant No. SRG21\211614).

†Dembo: Reichman University (aluma.dembo@runi.ac.il); Kariv: University of California, Berkeley (kariv@berkeley.edu); Polisson: University of Leicester (matt.polisson@leicester.ac.uk); Quah: National University of Singapore (ecsqkhj@nus.edu.sg).

1 Introduction

Expected utility theory (EUT) lies at the very heart of economics so it is natural that experimentalists would want to test the empirical validity of the axioms on which EUT is based. Empirical violations of EUT bring into question the rationality of individual behavior and, more specifically, raise criticisms of the familiar von Neumann and Morgenstern (1947) *independence* axiom as the touchstone for rational decision making under risk. Such criticisms have motivated various theoretical alternatives to EUT, and the experimental investigation of these theories has resulted in new empirical regularities.

For the most part, generalizations of EUT weaken the independence axiom but embody ordering (completeness and transitivity) and monotonicity with respect to first-order stochastic dominance (FOSD).[1] To test EUT and its various generalizations, laboratory experiments typically use several pairwise alternatives, à la Allais, that these various theories rank differently, while making a presumption that subjects adhere to the (more) fundamental/conventional axioms of ordering and monotonicity.

Given that EUT is part of the core of economics—and not something that one can or should abandon lightly—we wish to provide a comprehensive assessment of all the axioms on which EUT is based, and not just the independence axiom. Our overall objective is to provide a better, positive account of choice behavior under risk by evaluating the performance of EUT (as well as non-EUT models) in a choice environment where all axioms underpinning these models can be evaluated.

To do this, we use an experiment where, through a graphical "point-and-click" design, subjects choose an allocation of contingent commodities from a budget set; the user-friendly interface makes it possible to present subjects with a large array of heterogeneous budget sets. Crucially, the rich information collected in this way allows us to perform (statistical) tests at the level of each individual subject. From these tests, we can ascertain each subject's degree of compliance with the different components of EUT. While budgetary experiments are not new, the vast majority of such experiments involve subjects choosing from two-dimensional budget lines (in particular, see Choi *et al.* (2007a,b)), whereas in our experiment, subjects

[1]Violations of FOSD are commonly regarded as mistakes/errors in decision making, and so monotonicity with respect to FOSD is a generally accepted principle in decision theory, as pointed out by Quiggin (1990), Wakker (1993), and Starmer (2000), among others.

choose from three-dimensional budget sets. The experiment involving three states and three associated securities has a number of important advantages over earlier experiments:

- A fundamental result attributable to Rose (1958) and extended by Banerjee and Murphy (2006) states that with only two goods, the *weak axiom of revealed preference* (WARP) and the *generalized axiom of revealed preference* (GARP) are observationally equivalent—any violation of GARP (cyclical inconsistency) must contain a violation of WARP (pairwise inconsistency). The important implication of this result is that with two goods we cannot separate incompleteness from intransitivity. With three (or more) goods, by contrast, choices can satisfy WARP (pairwise consistency) and violate GARP (cyclical inconsistency), indicating complete yet nontransitive preferences. This is a crucial point because the experimental design should first provide a discriminating test of the ordering axioms (completeness and transitivity) before jointly testing any additional properties, such as the independence axiom, which places strong restrictions on the precise form of preferences.

- Within the context of choices from three-dimensional budget sets, prominent non-EUT models give rise to distinct utility specifications, which yield empirically discriminating restrictions on observed behavior. However, these differences are no longer prominent within the context of choices from two-dimensional budget lines.[2] The greater empirical separation among non-EUT models in three-dimensional choice data allows for a more rigorous test of EUT (by testing it against a richer set of alternatives). This is consistent with our power analysis, which shows that data from three-dimensional budget sets provide a stronger test in terms of power—especially of EUT versus non-EUT models that respect FOSD—than data from two-dimensional budget lines.

Our empirical analysis is in the revealed preference tradition of Afriat (1967, 1973), Diewert (1973), and Varian (1982, 1983a, 1990). Afriat's (1967) theorem tells us that if a finite dataset generated by an individual's choices from linear budget sets (as in our experiment) obeys GARP, then the data can be rationalized by a continuous and increasing utility function. This result gives a practical way of checking whether a dataset is *rationalizable* in this

[2]For example, the rank-dependent utility (Quiggin, 1982, 1993) and disappointment aversion (Gul, 1991) models reduce to the same form with two equally likely states; we provide further details in the Appendix.

fundamental/basic sense. There are also extensions of Afriat's (1967) theorem that allow us to test whether a dataset can be rationalized by a utility function with stronger properties. In particular, we could test whether a dataset is *FOSD-rationalizable*, in the sense that it is consistent with the maximization of a continuous utility function that is increasing with respect to FOSD; and whether a dataset is *EUT-rationalizable*, in the sense that it is consistent with the maximization of a continuous utility function of the expected utility form.

When subjects make decisions, they often make mistakes; and these mistakes manifest themselves in inconsistent choices. Since GARP provides an exact test (either choices satisfy GARP or they do not) and choice data almost always contain at least some errors, we assess how nearly the data comply with GARP by calculating Afriat's (1973) *critical cost-efficiency index* (CCEI), which measures the extent by which each budget constraint must be reduced in order to remove all violations of GARP (thereby rendering the data rationalizable). The CCEI, denoted by e^*, is bounded between 0 and 1 and can be interpreted as saying that the decision maker is leaving (as much as) a fraction $1-e^*$ of the money on the table due to errors because the chosen allocation is utility-maximizing only when compared to allocations in the reduced budget set. The CCEI is thus the "least costly" adjustment required to remove all errors. In this sense, it measures the overall "noise" in individual behavior.

Using recent refinements of the CCEI (in particular, see Polisson, Quah, and Renou (2020)), we can also measure the extent to which budget sets need to be reduced in order for a dataset to be FOSD-rationalizable and EUT-rationalizable. Thus, for any dataset collected from an individual subject's choices, three CCEI-type scores can be calculated: e^* for (basic) rationalizability, e^{**} for FOSD-rationalizability (which can be no greater than e^* since FOSD-rationalizability is the more stringent requirement), and e^{***} for EUT-rationalizability (which can be no greater than e^{**} since EUT-rationalizability is the more stringent requirement).

The use of the same measure for all three models we consider has the very important advantage that we can decompose violations of EUT and compare the magnitudes of violations of the different axioms from which EUT can be derived. Perfect consistency with EUT implies that $1 = e^* = e^{**} = e^{***}$, whereas perfect consistency with any of the familiar non-EUT alternatives that respect FOSD but not EUT itself implies that $1 = e^* = e^{**} > e^{***}$. Our rich individual-level data also allow us to make statistical comparisons of the difference between *perfect* rationalizability and FOSD-rationalizability ($1-e^{**}$) and the difference between FOSD-rationalizability and EUT-rationalizability ($e^{**} - e^{***}$), using a purely nonparametric difference-in-differences econometric approach.

Figure 1 depicts the distributions of the e^*, e^{**}, and e^{***} rationalizability scores. The horizontal axis presents score values; the vertical axis indicates the percent of subjects whose score is above each value. Only 16.1 percent of subjects are perfectly rationalizable ($e^* = 1$), but none are perfectly FOSD-rationalizable ($e^{**} = 1$) or EUT-rationalizable ($e^{***} = 1$). If we set a more permissive score value threshold of 0.95, we find that 63.1 percent of subjects are rationalizable ($e^* > 0.95$), 28.0 percent are FOSD-rationalizable ($e^{**} > 0.95$), and 16.1 percent are EUT-rationalizable ($e^{***} > 0.95$). In particular, this means that the subjects who simply fail FOSD-rationalizability are about *six* times more numerous as those who are FOSD-rationalizable but not EUT-rationalizable; at the other score values displayed in Figure 1, this ratio ranges between three and six. Moreover, the source of this aggregate phenomenon becomes clear when we compare these differences ($1-e^{**}$ and $e^{**}-e^{***}$) within each subject; for 84.5 percent of subjects this gap is positive and statistically significant.

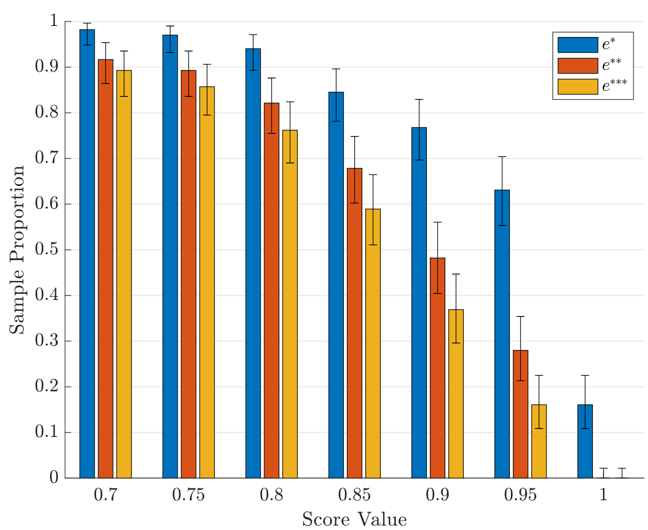


Figure 1: Distributions of Rationalizability Scores

The horizontal axis presents score values. The vertical axis indicates the percent of subjects whose score is above each value. The braces represent exact 95 percent confidence intervals on the proportions.

Violations of EUT thus appear to run much deeper than violations of merely independence, challenging many of the most prominent non-EUT alternatives which respect FOSD.

Furthermore, of the 83.9 percent of subjects who are not perfectly rationalizable, no subject has only GARP violations which are free of WARP violations; and the CCEI scores required to remove all violations of WARP and GARP are identical for all but 4 subjects (2.4 percent). Therefore, the scope for rationalizability even by a theory of choice based on a model of (complete and) nontransitive preference (as proposed by Bell (1982), Fishburn (1982), and Loomes and Sugden (1982)) also shows little promise in our context.

While there are alternative cost-efficiency measures of violations of rationalizability, none of the various indices are viable given our empirical exercise. We adopt the CCEI as a measure of rationalizability since it is straightforward to interpret and computationally feasible for moderately large datasets and for all classes of models we consider. Partly for those reasons, the CCEI is also the most commonly used measure in empirical revealed preference research. Nevertheless, we also develop an alternative distance-based approach which yields similar empirical conclusions, albeit in a somewhat limited empirical exercise.

The emphasis in our paper is to provide a *comprehensive* and *nonparametric* test of complete representations of preferences under risk rather than focusing on individual axioms. Our main result—that violations of EUT are relatively small after accounting for violations of ordering and monotonicity with respect to FOSD—is what Quiggin (1982) calls an "undesirable result" as ordering and monotonicity are more fundamental principles than the independence axiom, and are embodied in the most prominent non-EUT theories of choice under risk. As Starmer (2000) notes, economists have taken the view that the independence axiom needs to be weakened on the grounds of predictive validity and psychological realism, but have generally left ordering and monotonicity unchallenged.

The rest of the paper is organized as follows. The next section provides more background and motivation. Section 3 describes our tests of rationalizability, and Section 4 outlines the experimental design and procedures. Section 5 summarizes the experimental results. Section 6 explains how the paper is related to the literature, focusing on recent revealed preference papers on choice under risk. Section 7 outlines what we think theorists, experimentalists, and other economists should take away from the paper. In the interests of brevity, all technical details that are not essential for understanding the results are relegated to the Appendix.

Much of the experimental evidence of "anomalies" in choice behavior suggests that EUT may not the right model of choice under risk. To understand the role of each of the axioms on which EUT is based, suppose that there are three mutually non-indifferent outcomes $x_h \succ x_m \succ x_l$, such as monetary consequences where $x_h > x_m > x_l$. The probability triangle in Figure 2 depicts the set of all possible lotteries—each point in the triangle represents a lottery (π_h, π_m, π_l) over the outcomes (x_h, x_m, x_l), where $\pi_h = 0$ on the horizontal edge, $\pi_m = 0$ on the hypotenuse (because $\pi_h + \pi_l = 1$), and $\pi_l = 0$ on the vertical edge.[3]

Ordering (completeness and transitivity) plus continuity imply that there exists a map of (non-intersecting) indifference curves on the probability triangle. Monotonicity with respect to first-order stochastic dominance (FOSD) implies that preferences are increasing from right to left along horizontal lines, from bottom to top along vertical lines, and from bottom-right to top-left along lines parallel to the hypotenuse (Figure 2a). Assuming that ordering and monotonicity hold, the independence axiom then implies that preferences admit an expected utility representation, so that the indifference curves in the triangle are parallel straight lines (Figure 2b). Viewed within the context of the triangle, independence is a strong requirement, leaving only the slope of the indifference lines undetermined (with steeper lines implying higher risk aversion).

An example of the famous Allais (1953) paradox can be illustrated by a pair of binary choices—between lotteries $\mathbf{a}$ and $\mathbf{b}$ and between lotteries $\mathbf{a}'$ and $\mathbf{b}'$ (Figure 2c). The imaginary straight lines connecting lotteries $\mathbf{a}$ and $\mathbf{b}$ and lotteries $\mathbf{a}'$ and $\mathbf{b}'$ are parallel to each other and flatter than the indifference lines so $\mathbf{a} \succ \mathbf{b}$ and $\mathbf{a}' \succ \mathbf{b}'$. But experimental subjects often make choices revealing that $\mathbf{a} \succ \mathbf{b}$ and $\mathbf{b}' \succ \mathbf{a}'$ (or $\mathbf{b} \succ \mathbf{a}$ and $\mathbf{a}' \succ \mathbf{b}'$), which is commonly taken as evidence against independence. This persistent finding has led to a large literature with the objective of developing new models of choice under risk that weaken the independence axiom.[4]

[3]The probability triangle was introduced by Marschak (1950) and popularized by Machina (1982) as a way of representing the choice space over lotteries.

[4]For comprehensive reviews, see Camerer (1995) and Starmer (2000). More recently, based on data from 81 experiments across 29 studies, Blavatskyy, Ortmann, and Panchenko (2022) conclude that "the Allais Paradox is a fragile empirical finding," likely to be observed when subjects choose between lotteries near edges of the triangle (involving small probabilities) and/or in experiments with (high) hypothetical payoffs.

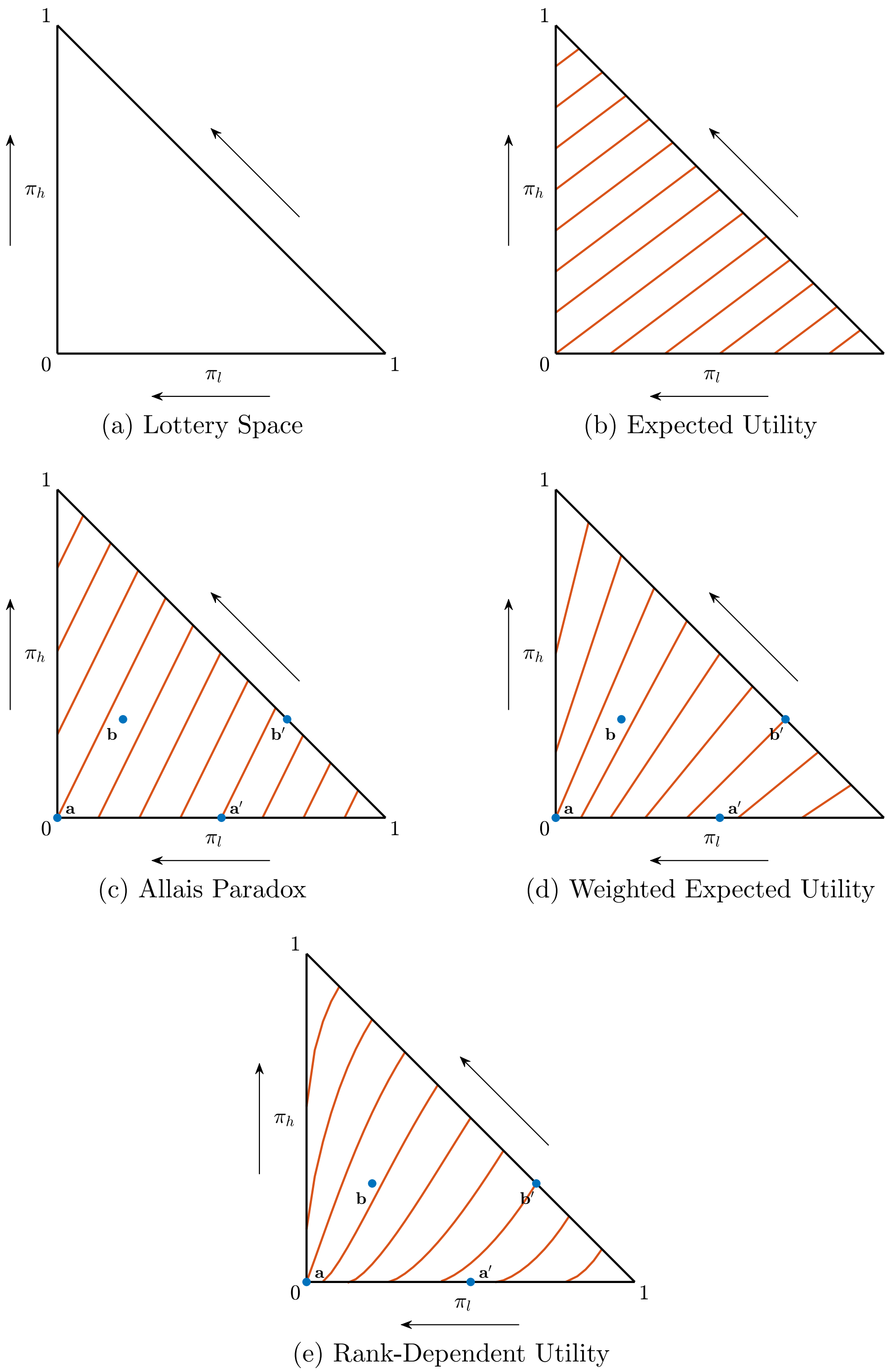


Figure 2: Probability Triangles

The probability triangle depicts the lottery space as a set of probability weights (π_h, π_m, π_l) over three fixed outcomes (x_h, x_m, x_l). (a) Ordering (completeness and transitivity) plus continuity guarantee non-intersecting indifference curves; monotonicity (with respect to FOSD) guarantees that preferences are increasing as shown (see arrows). (b) Adding independence gives rise to EUT, characterized by indifference curves that are parallel straight lines. (c) The Allais paradox arises because EUT requires $\mathbf{a} \succ \mathbf{b}$ and $\mathbf{a}' \succ \mathbf{b}'$, but experimental subjects often make choices revealing that $\mathbf{a} \succ \mathbf{b}$ but $\mathbf{b}' \succ \mathbf{a}'$. Alternatives to EUT like (d) weighted expected utility and (e) rank-dependent utility often avoid the Allais paradox by relaxing independence while adhering to ordering and monotonicity.

In weighted expected utility (Dekel, 1986; Chew, 1989), for example, all indifference curves are again straight lines but they typically "fan out"—that is, they become steeper (corresponding to higher risk aversion) when moving northwest in the triangle (Figure 2d). In loss/disappointment aversion (Gul, 1991), the indifference curves are also straight lines but "fan in" for lotteries better than x_m (top part of the triangle) and "fan out" for lotteries worse than x_m (bottom part of the triangle). In rank-dependent utility (Quiggin, 1982, 1993) and prospect theory (Kahneman and Tversky, 1979; Tversky and Kahneman, 1992), by contrast, the indifference curves are not straight lines and they can "fan out" or "fan in," especially near the triangle boundaries (Figure 2e). Each of the conventional alternatives to EUT gives rise to indifference curves with distinctive shapes in the triangle, but with the common feature that they avoid the Allais paradox.

In many experimental studies, the main criterion used to evaluate a given theory is the fraction of choices that it correctly predicts. A few studies have also estimated parametric utility functions for individual subjects. Generally speaking, these experiments involve collecting a small number of decisions from each subject, with the decisions involving specific choices that are narrowly tailored to test the independence axiom against its various generalizations. The accumulated experimental evidence against independence has prompted theorists to develop formal alternatives to EUT and, apart from a few notable exceptions, non-EUT models have relaxed the independence axiom while maintaining ordering and monotonicity with respect to FOSD. However, our basic contention is that we ought to have a wider view of the relative performance of EUT and therefore *all* the assumptions which underpin the model deserve close scrutiny.

In this paper, we test EUT and its generalizations in a canonical setting of decision making under risk—the portfolio choice problem of the consumer. Within this setting, we develop tests of rationalizability that are *comprehensive*, in the sense that we test whether a given model—taken as a whole—succeeds or fails in explaining the data, rather than focusing on specific individual axioms. Furthermore, by evaluating a set of progressively restrictive models using a common measure of performance, we can compare the relative impact of the different EUT axioms. Another important feature of our tests is that they are *nonparametric*, in the sense that we make no auxiliary functional form assumptions on the utility function.

3 Theory

In this section, we describe the theory on which the experimental design is based. In the experimental task we study, subjects make decisions under conditions of uncertainty about the objective parameters of the environment. There are three equally likely states of nature ($s = 1, 2, 3$) and an Arrow security for each state. An Arrow security for state s is defined to be a promise to deliver one token (the experimental currency) if state s occurs and nothing otherwise. Let $\mathbf{x} = (x_1, x_2, x_3)$ denote a portfolio of securities or bundle of contingent commodities, where $x_s \geqslant 0$ denotes the number of units of security s. Without essential loss of generality, assume the individual's endowment is normalized to 1. The budget set $\mathcal{B} = \{\mathbf{x} \in \mathbb{R}^3_+ : \mathbf{p} \cdot \mathbf{x} \leqslant 1\}$ is then all the bundles that are affordable given this endowment and given the price vector $\mathbf{p} = (p_1, p_2, p_3)$, where $p_s > 0$ denotes the price of security s.

The experimental design requires subjects to solve a sequence of 50 decision problems. Each decision problem is defined by a budget set, represented graphically on a computer screen. The subject uses the mouse to select a portfolio from the feasible set by "pointing and clicking." Subjects were informed that the states are equally likely, eliminating any ambiguity. For each subject, we have a set of 50 observations $\mathcal{D} := \{(\mathbf{p}^i, \mathbf{x}^i)\}_{i=1}^{50}$, where $\mathbf{p}^i = (p_1^i, p_2^i, p_3^i)$ denotes the i-th observation of the price vector and $\mathbf{x}^i = (x_1^i, x_2^i, x_3^i)$ denotes the corresponding allocation. The experiment thus provides a large amount of choice data consisting of many individual decisions over a wide range of three-dimensional budget sets.

3.1 Rationalizability

The most fundamental/basic question to ask of each individual-level dataset obtained in our experiment is whether it is consistent with utility maximization. We refer to a utility function defined over contingent commodities, $U : \mathbb{R}^3_+ \to \mathbb{R}$, as *well-behaved* if it is continuous and increasing. An individual-level dataset $\mathcal{D}$ is said to be *rationalizable* if there is a well-behaved utility function U that rationalizes $\mathcal{D}$, in the sense that $U(\mathbf{x}^i) \geqslant U(\mathbf{x})$ for all

$$\mathbf{x} \in \mathcal{B}^i = \{\mathbf{x} \in \mathbb{R}^3_+ : \mathbf{p}^i \cdot \mathbf{x} \leqslant 1\}.$$

We want to know whether each dataset $\mathcal{D}$ is rationalizable in the sense just defined—that is, whether it could have been generated by an individual maximizing a well-behaved utility function. The canonical test for this involves checking the *generalized axiom of revealed preference* (GARP). A well-known result, due to Afriat (1967), tells us that a dataset $\mathcal{D}$ is rationalizable if and only if $\mathcal{D}$ satisfies GARP.[5] GARP requires that if allocation $\mathbf{x}^i$ is revealed preferred to $\mathbf{x}^j$, then $\mathbf{x}^i$ must cost at least as much as $\mathbf{x}^j$ at the prices prevailing when $\mathbf{x}^j$ is chosen, $\mathbf{p}^j \cdot \mathbf{x}^i \geqslant 1$. Put precisely, allocation $\mathbf{x}^i$ is said to be *directly revealed preferred* to $\mathbf{x}^j$, denoted $\mathbf{x}^i \, R^D \, \mathbf{x}^j$, if $\mathbf{p}^i \cdot \mathbf{x}^j \leqslant 1$ (equivalently, $\mathbf{x}^j \in \mathcal{B}^i$) and *directly revealed strictly preferred* if the inequality is strict. The *revealed preference* relation, denoted R, is the transitive closure of the direct revealed preference relation.[6] GARP requires that if $\mathbf{x}^i \, R \, \mathbf{x}^j$ then $\mathbf{x}^j$ is not directly revealed strictly preferred to $\mathbf{x}^i$. To verify GARP, it is thus necessary to have an efficient way to compute the transitive closure R of the direct revealed preference relation R^D and check that, for every pair of allocations $\mathbf{x}^i$ and $\mathbf{x}^j$ satisfying $\mathbf{x}^i \, R \, \mathbf{x}^j$, we do not have $\mathbf{x}^j$ being directly revealed strictly preferred to $\mathbf{x}^i$.

GARP fails whenever $\mathcal{D}$ contains a revealed preference cycle: a sequence of allocations $\{\mathbf{x}^k\}_{k=1}^K$ with $\mathbf{x}^1 = \mathbf{x}^i$ and $\mathbf{x}^K = \mathbf{x}^i$, such that $\mathbf{x}^k \, R^D \, \mathbf{x}^{k+1}$ for every $k = 1, \ldots, K-1$, where at least one direct revealed preference relation is strict; we refer to such a revealed preference cycle as a GARP violation. The *weak axiom of revealed preference* (WARP) can then be seen as a weakening of GARP which only requires that if $\mathbf{x}^i \, R^D \, \mathbf{x}^j$ then $\mathbf{x}^j$ is not directly revealed strictly preferred to $\mathbf{x}^i$. WARP fails whenever $\mathcal{D}$ contains a pairwise revealed preference cycle $\mathbf{x}^i \, R^D \, \mathbf{x}^j$ and $\mathbf{x}^j \, R^D \, \mathbf{x}^i$, where at least one direct revealed preference relation is strict; we refer to such cycle as a WARP violation.

Choices from two-dimensional budget lines—as collected in previous experiments—cannot satisfy WARP and violate GARP since every GARP violation must contain a WARP violation (see Rose (1958) and Banerjee and Murphy (2006)). This is particularly important because choice data which violate not only GARP but also WARP cannot be rationalized even by a complete but *nontransitive* preference ordering.[7] This is not the case if choices are from three-dimensional (or more) budget sets, where it is indeed possible for choice data

[5]It is straightforward to show that choice data generated by the maximization of a locally nonsatiated utility function must obey GARP. Conversely, and somewhat more difficult to show, if an individual's choice data obey GARP, then they can be rationalized by a well-behaved and concave utility function.

[6]That is, $\mathbf{x}^i \, R \, \mathbf{x}^j$ if there exists a sequence of allocations $\{\mathbf{x}^k\}_{k=1}^K$ with $\mathbf{x}^1 = \mathbf{x}^i$ and $\mathbf{x}^K = \mathbf{x}^j$, such that $\mathbf{x}^k \, R^D \, \mathbf{x}^{k+1}$ for every $k = 1, \ldots, K-1$.

[7]Models of choice under risk generated by nontransitive preferences were proposed by Bell (1982), Fishburn (1982), and Loomes and Sugden (1982). See Starmer (2000) for a discussion.

to satisfy WARP and violate GARP. We can thus provide a more discriminating test of rationalizability by allowing for the separation of incomplete from complete but nontransitive preferences, a crucial step before jointly testing the additional axioms underpinning EUT.

Since GARP provides an exact test of rationalizability and individual choices almost always involve at least some noise—subjects may optimize incorrectly, or implement optimal choices with imprecision, or err in any other of many possible ways—Afriat (1972, 1973) proposes the notion of the *critical cost-efficiency index* (CCEI) to measure these mistakes/errors. The CCEI is the fraction by which each budget constraint must be reduced in order to remove all violations of GARP. By definition, the CCEI is between 0 and 1: indices closer to 1 mean the data are closer to perfect consistency with GARP and hence to perfect consistency with utility maximization.

Put precisely, given a number $e \in (0, 1]$, we say that a dataset $\mathcal{D}$ is *rationalizable at cost-efficiency* e if there is a well-behaved utility function U such that $U(\mathbf{x}^i) \geqslant U(\mathbf{x})$ for all

$$\mathbf{x} \in \mathcal{B}^i(e) = \{\mathbf{x} \in \mathbb{R}^3_+ : \mathbf{p}^i \cdot \mathbf{x} \leqslant e\}.$$

This concept is a less stringent requirement than (exact) rationalization since $\mathcal{B}^i(e)$ is a subset of $\mathcal{B}^i$ (except when $e = 1$, in which case the concepts coincide). Afriat's (1973) CCEI, denoted e^*, which is associated with the dataset $\mathcal{D}$ is the greatest possible cost-efficiency e at which $\mathcal{D}$ is rationalizable. We can calculate the CCEI e^* based on a modified version of GARP. In the notation introduced above, for any number $e \in (0, 1]$, we define the direct revealed preference relation $R^D(e)$ as $\mathbf{x}^i \, R^D(e) \, \mathbf{x}^j$ if $\mathbf{p}^i \cdot \mathbf{x}^j \leqslant e$ (equivalently, $\mathbf{x}^j \in \mathcal{B}^i(e)$), and we define $R(e)$ to be the transitive closure of $R^D(e)$. We say that $R(e)$ satisfies GARP if, for every pair of allocations $\mathbf{x}^i$ and $\mathbf{x}^j$ where $\mathbf{x}^i \, R(e) \, \mathbf{x}^j$, we have $\mathbf{p}^j \cdot \mathbf{x}^i \geqslant e$. The largest value of e such that the relation $R(e)$ satisfies GARP coincides with e^*.[8]

How should we interpret the CCEI? Suppose a dataset $\mathcal{D}$ is not rationalizable by a well-behaved utility function U. Then, choices must involve at least one error—that is, a budget set $\mathcal{B}^j$ and an allocation $\mathbf{y} \in \mathcal{B}^j$ such that $U(\mathbf{y}) > U(\mathbf{x}^j)$. A money-metric measure of the severity of this error is the fraction of the budget wasted by choosing $\mathbf{x}^j$ instead of $\mathbf{y}$, which is $1 - \mathbf{p}^j \cdot \mathbf{y}$. The CCEI e^* is then obtained by minimizing the budget overspent among *all*

[8] By a variation of Afriat's (1967) theorem, we know that $\mathcal{D}$ is rationalizable at cost-efficiency e if and only if $R(e)$ satisfies GARP (Afriat, 1973).

well-behaved utility functions. In this sense, for every observation of $(\mathbf{p}^i, \mathbf{x}^i)$, there exists a well-behaved utility function U such that $U(\mathbf{x}^i) \geqslant U(\mathbf{x})$ for any allocation $\mathbf{x}$ which is more than $1 - e^*$ percent cheaper than $\mathbf{x}^i$ at the prices $\mathbf{p}^i$ prevailing when $\mathbf{x}^i$ is chosen. The CCEI thus provides the "least costly" adjustment of budget sets that accounts for mistakes/errors. It is a theoretically disciplined measure of noise and it also has a well-established economic interpretation. As Afriat (1973) puts it, the CCEI captures the idea that the decision maker "has a definite structure of wants," but "programs at a level of cost-efficiency e." For a fuller discussion of the interpretation of the CCEI see Polisson and Quah (2022).

3.2 FOSD-Rationalizability

Afriat's (1967) theorem is just the first of a long list of results with the following pattern: an individual-level dataset $\mathcal{D}$ is rationalizable by a well-behaved (continuous and increasing) utility function U belonging to some family if and only if $\mathcal{D}$ obeys some property. For our purposes, two families are particularly important. The first is the family of utility functions that are continuous and increasing with respect to first-order stochastic dominance (FOSD); the latter means that $U(\mathbf{x}'') \geqslant U(\mathbf{x}')$ whenever $F_{\mathbf{x}''} \geqslant F_{\mathbf{x}'}$, where $F_{\mathbf{x}''}$ and $F_{\mathbf{x}'}$ are the resulting payoff distributions, with the inequality being strict if FOSD is strict. We refer to such utility functions as *FOSD-increasing*. Note that every FOSD-increasing utility function is well-behaved since such a utility function must be increasing.[9]

Violations of FOSD might reasonably be regarded as errors in decision making, regardless of underlying risk attitudes—that is, as a failure to recognize that some allocations yield payoff distributions with unambiguously lower returns. Violations of FOSD are therefore compelling as mistakes/errors, and so monotonicity with respect to FOSD is widely embodied in models of decision making under risk.[10] In the Appendix, we cover some prominent

[9]A utility function U that is FOSD-increasing must also be increasing (in the sense that $U(\mathbf{x}'') > U(\mathbf{x}')$ whenever $\mathbf{x}'' > \mathbf{x}'$) but the converse is not true. Suppose that there are two equally likely states. Then $U(1,3) > U(2,1)$ if U is FOSD-increasing because $(1,3)$ strictly first-order stochastically dominates $(2,1)$, but no relationship between $U(1,3)$ and $U(2,1)$ is implied by U being increasing.

[10]This is true, for example, of weighted expected utility (Dekel, 1986; Chew, 1989), rank-dependent utility (Quiggin, 1982, 1993), cumulative prospect theory (Tversky and Kahneman, 1992), and (under certain restrictions) reference-dependent risk preferences (Kőszegi and Rabin, 2007). Notably, the original formulation of prospect theory (Kahneman and Tversky, 1979) allows for violations of monotonicity but, partly for this reason, it was reformulated as cumulative prospect theory (Tversky and Kahneman, 1992) to exclude such behavior. For an exception to this rule see, for example, Manzini and Mariotti (2008).

examples of FOSD-increasing utility functions—including expected utility theory (EUT) and also generalizations of EUT such as the rank-dependent utility (Quiggin, 1982, 1993) and disappointment aversion (Gul, 1991) models—as well as their relationships to one another.

We say that an individual-level dataset $\mathcal{D}$ is *FOSD-rationalizable* if it can be rationalized by an FOSD-increasing utility function. Nishimura, Ok, and Quah (2017) extend Afriat's (1967) theorem to provide a necessary and sufficient condition for $\mathcal{D}$ to be FOSD-rationalizable by strengthening GARP to rule out a larger set of revealed preference cycles. In the case where the states are equally likely (as in our experiment), requiring a utility function to be FOSD-increasing is equivalent to requiring it to be increasing and *symmetric*. An implication of this property is that $\mathcal{D}$ must be free of (within-observation) FOSD violations—clearly, any decision that involves allocating fewer tokens to the cheaper security is incompatible with the maximization of an increasing and symmetric utility function.[11,12] Whenever $\mathcal{D}$ is not exactly FOSD-rationalizable, we can check whether it can be rationalized at cost-efficiency e by an FOSD-increasing utility function and calculate the corresponding CCEI, denoted by e^{**}. Since the family of FOSD-increasing utility functions is contained within the family of well-behaved utility functions, it must be the case that $e^{**} \leqslant e^{*}$.

3.3 EUT-Rationalizability

Within economics there is a vast amount of experimental work which has led to the development of various theoretical alternatives to EUT. The second important family of well-behaved utility functions we analyze therefore contains utility functions that are compatible with EUT. In our experiment with three equally likely states, EUT requires the existence of a continuous and increasing *Bernoulli index* $u : \mathbb{R}_+ \to \mathbb{R}$, such that

$$U(\mathbf{x}) = \tfrac{1}{3}u(x_1) + \tfrac{1}{3}u(x_2) + \tfrac{1}{3}u(x_3),$$

[11]For example, suppose that with two equally likely states, the allocation $(1, 2)$ is chosen at prices $(1, 2)$. However, the allocation $(2, 1)$ is stochastically equivalent to the allocation $(1, 2)$ and yet costs strictly less. This is incompatible with the maximization of an FOSD-increasing utility function since $(2 + \varepsilon, 1 + \varepsilon)$ is affordable for $\varepsilon > 0$ sufficiently small and strictly superior to $(1, 2)$.

[12]Utility functions representing reference-dependent risk preferences (specifically the choice acclimating personal equilibrium model of Kőszegi and Rabin (2007)) can fail to be FOSD-increasing if loss aversion is sufficiently high (see Masatlioglu and Raymond (2016)); however, these preferences are always locally nonsatiated and, in our experimental setting, symmetric. For reasons explained in greater detail in the Appendix, utility functions that are symmetric and locally nonsatiated cannot rationalize any behavior that cannot also be rationalized by a symmetric and increasing utility function. Thus the explanatory power of reference-dependent risk preferences does not extend beyond that of FOSD-increasing utility functions.

for all contingent commodity allocations $\mathbf{x} = (x_1, x_2, x_3)$. We say that an individual-level dataset $\mathcal{D}$ is *EUT-rationalizable* if it can be rationalized by a utility function U taking the expected utility form. Since every such U must be FOSD-increasing, EUT-rationalizability is a stronger requirement than FOSD-rationalizability.

Polisson, Quah, and Renou (2020) develops a test called the *generalized restriction of infinite domains* (GRID) which can be used to characterize individual-level datasets that are EUT-rationalizable. The GRID method replaces the true contingent commodity space (which in our experiment is $\mathbb{R}^3_+$) with a finite set $\mathcal{G}$ of allocations in $\mathbb{R}^3_+$ constructed in a certain manner; a dataset $\mathcal{D}$ is EUT-rationalizable if and only if it can be rationalized by a utility function taking the expected utility form, where only allocations in $\mathcal{G}$ are included in the consumption space (see the Appendix for details). The latter condition gives a viable test of EUT-rationalizability because it can be converted into a problem of solving a finite system of linear inequalities. Using this method, one could also calculate e^{***}, the CCEI corresponding to EUT-rationalizability. Since this family of utility functions is contained within the family of FOSD-increasing utility functions, it must be the case that $e^{***} \leqslant e^{**}$.

3.4 Comparing Scores

To recap, given any individual-level dataset $\mathcal{D}$ we can calculate three rationalizability scores corresponding to three nested models, with

$$1 \geqslant e^* \geqslant e^{**} \geqslant e^{***} > 0.$$

There are, of course, other families of utility functions besides these three, and there will be rationalizability scores corresponding to those families as well. In particular, the families of FOSD-increasing utility functions which generalize EUT will *necessarily* have rationalizability scores between e^{**} and e^{***}. The great advantage of measuring—on the same scale—a dataset's consistency with three increasingly stringent models is that it allows us to determine the *source* of the subject's departure from EUT. A subject who is perfectly consistent with EUT will have $1 = e^* = e^{**} = e^{***}$, while a subject who is perfectly consistent with any of the familiar non-EUT alternatives that respect FOSD will have $1 = e^* = e^{**} > e^{***}$.

Typically, values of e^{***} will be strictly less than one. Crucial to our analysis, the corresponding values of e^{**} will then allow us to make (statistical) comparisons between the difference between *perfect* rationalizability and FOSD-rationalizability and the difference between FOSD-rationalizability and EUT-rationalizability. A subject for whom $1 - e^{**}$ is (much) smaller than $e^{**} - e^{***}$ could indeed be violating the independence axiom, but such behavior could potentially be explained by a FOSD-increasing utility model which relaxes the independence axiom; on the other hand, a subject for whom $1 - e^{**}$ is (much) larger than $e^{**} - e^{***}$ may or may not be violating the independence axiom but would require a more substantial departure from the standard paradigm to explain such behavior.

3.5 Alternative Measures

The CCEI is by no means the only way of measuring departures from exact rationalizability (or FOSD-rationalizability or EUT-rationalizability). We focus on this index in our empirical analysis for two related reasons. First, it has a natural economic interpretation (see Varian (1990)), and it is the most commonly used measure in the revealed preference literature.[13] Second, it is easy to calculate for the three (increasingly narrow) families of utility functions we examine. We are not aware of any other index where there are computationally efficient ways of calculating its values for the three families of utility functions under consideration.[14]

That said, we *do* carry out some analysis using an alternative, distance-based approach to noise/error, which is somewhat more in line with conventional econometric analysis. In this approach, we measure the distance k_U^i between the observed allocation $\mathbf{x}^i = (x_1^i, x_2^i, x_3^i) \in \mathcal{B}^i$ and another allocation $\mathbf{z}^i = (z_1^i, z_2^i, z_3^i) \in \mathcal{B}^i$, where $\mathbf{z}^i$ maximizes the well-behaved utility function U in the budget set $\mathcal{B}^i$. For example, one could measure the distance between $\mathbf{x}^i$ and $\mathbf{z}^i$ by the greatest difference in expenditure among the three goods, in which case

$$k_U^i := \max_{s=1,2,3} |p_s^i(z_s^i - x_s^i)|.$$

We say that an individual-level dataset $\mathcal{D}$ is *rationalizable* at distance k if there is a well-behaved utility function U such that $k_U^i \leqslant k$ for each observation of $(\mathbf{p}^i, \mathbf{x}^i)$ and we define k^* as the smallest distance k at which $\mathcal{D}$ is rationalizable. Put differently, the dataset $\mathcal{D}$

[13] A small subset of the many studies using the CCEI includes Harbaugh, Krause, and Berry (2001) on children's preferences, Andreoni and Miller (2002) and Fisman, Kariv, and Markovits (2007) on social preferences, and Choi *et al.* (2007a, 2014) and Carvalho, Meier, and Wang (2016) on risk preferences.

[14] Several other cost-efficiency indices have been suggested to measure departures from exact rationalizability (see, for example, Varian (1990)). The relationships between the CCEI and other cost-efficiency indices are discussed in Choi *et al.* (2014) and Halevy, Persitz, and Zrill (2018).

can be made rationalizable by perturbing the observed choices by any distance greater than k^*. In a similar way, we can define k^{**} and k^{***} as the smallest distances k at which $\mathcal{D}$ is FOSD-rationalizable and EUT-rationalizable.

There are known procedures for calculating k^* and k^{**} precisely (see Hu *et al.* (2021) for a description). However, there is no known procedure to calculate k^{***} and so we rely on a method that only gives an upper bound on its value; furthermore, this approximate method works only when datasets have sufficiently few observations. These difficulties limit the scope of our empirical analysis using this index. However, the conclusions from our limited empirical exercise using the distance-based index are broadly consistent with what we obtain from our efficiency-based analysis using the CCEI (see the Appendix for the details of our procedures and empirical results).

4 Experiment

The experimental procedures described below are identical to those described by Choi *et al.* (2007b) and used by Choi *et al.* (2007a) to study a portfolio choice problem with two states of nature ($s = 1, 2$) and two associated contingent commodities, except that this experiment incorporates three states ($s = 1, 2, 3$) and three contingent commodities.[15,16,17] We conducted the experiment at UC Berkeley and UCLA. The subjects in the experiment were recruited from undergraduate classes at these institutions. Each experimental subject faced 50 independent decision problems. These decision problems were presented using a graphical inter-

[15] We are building on the expertise that we have acquired in previous work using the experimental method across different types of individual choice problems. The two-dimensional budget lines graphical interface was introduced by Choi *et al.* (2007b), and used by Choi *et al.* (2007a) with student subjects and by Choi *et al.* (2014) with subjects from a nationally representative sample. The datasets of Choi *et al.* (2007a, 2014) have been analyzed in many papers, including Halevy, Persitz, and Zrill (2018), Polisson, Quah, and Renou (2020), de Clippel and Rozen (2023), and Echenique, Imai, and Saito (2023), among others.

[16] A series of papers employ a similar methodology to study social preferences with different pools of subjects: Fisman, Kariv, and Markovits (2007), Fisman *et al.* (2015), Fisman, Jakiela, and Kariv (2015, 2017), Li, Dow, and Kariv (2017), Li *et al.* (2022), and Fisman *et al.* (2023). The two-person budget line dictator experiment of Fisman, Kariv, and Markovits (2007) is identical to Andreoni and Vesterlund (2001) and Andreoni and Miller (2002) except for presenting the choice problems graphically, allowing a much wider range of budget lines than can be tested using a pencil-and-paper questionnaire method.

[17] The experimental method with three-dimensional budget sets has been used by Ahn *et al.* (2014) to study ambiguity aversion, but so far has not been used to study risk. Halevy and Mayraz (2022) introduce an interface to study even higher-dimensional budget sets. In general, the experimental literature involving budgets has shifted towards using user-friendly graphical interfaces that allow for the quick and efficient elicitation of many decisions per subject. See, for example, Andreoni and Sprenger (2012) on time preferences.

face. On a computer screen, subjects saw a graphical representation of a three-dimensional budget set and chose portfolios through a simple "point-and-click."

For each subject, the computer selected 50 budget sets randomly from the set of planes that intersect all axes at or above the 10 token level and at or below the 100 token level, with at least one intercept at or above the 50 token level. The budget sets selected for each subject were independent of one another and of the budget sets selected for other subjects. Subjects were not informed of any state that was actually realized until the end of the experiment. This procedure was repeated until all 50 rounds were completed. At the end of the experiment, the computer randomly selected one of the 50 decision rounds to carry out for payoffs, and token allocations were converted into dollars. The round selected depended solely on chance. Full experimental instructions, including the computer program dialog window, are available in the Appendix.

In order to ascertain the power of an experiment in testing basic rationalizability, Bronars (1987) proposes as a benchmark the choices of a simulated subject who randomizes uniformly among all allocations on each budget set. In our case, we illustrate the power of the experiment in testing EUT-rationalizability by adapting the Bronars (1987) procedure. Specifically, we present the simulated subject with 50 randomly generated budget sets (like the actual human subjects); choices are then drawn randomly from these budget sets, but with the additional restriction that the choice data are perfectly FOSD-rationalizable. Precise details of the simulation procedure can be found in the Appendix.

Figure 3 depicts the distributions of e^{***} generated by 1,000 such simulated subjects in the two-dimensional and three-dimensional budget set experiments. The horizontal axis shows the value of e^{***}; the vertical axis measures the fraction of simulated subjects whose scores are above each value. If we choose $e^{***} = 0.9$ as our critical value, we find that just over 20 percent of simulated subjects are EUT-rationalizable above this threshold ($e^{***} > 0.9$) in the three-dimensional experiment. Thus it is clear that the design of our experiment does *not* guarantee that a dataset is (nearly) EUT-rationalizable simply because it is FOSD-rationalizable. It is also noteworthy that the power of the two-dimensional experiment is significantly lower: in that case, more than 80 percent have $e^{***} > 0.9$.

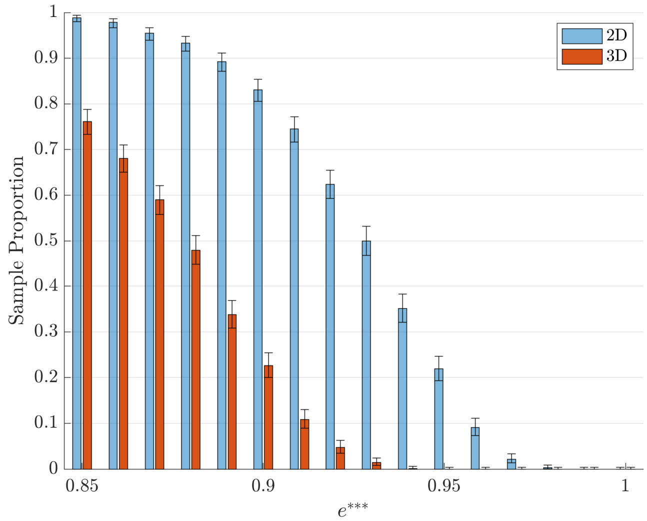


Figure 3: Power of EUT-Rationalizability

The horizontal axis shows the value of e^{***}. The vertical axis measures the fraction of simulated subjects in two-dimensional (2D) and three-dimensional (3D) experiments whose scores are above each value. The braces represent exact 95 percent confidence intervals on the proportions.

5 Experimental Results

In this section, we present the experimental results. The data from the experiment contain observations on 168 individual subjects. For each subject, we have a set of 50 observations $\mathcal{D} := \{(\mathbf{p}^i, \mathbf{x}^i)\}_{i=1}^{50}$, where $\mathbf{p}^i = (p_1^i, p_2^i, p_3^i)$ denotes the i-th observation of the price vector and $\mathbf{x}^i = (x_1^i, x_2^i, x_3^i)$ denotes the corresponding allocation. The experiment therefore provides a large amount of data consisting of many individual choices over a wide range of budget sets. This is a crucial point because the (within-subject) power of the experiment depends upon two factors: the frequency of the intersection of the budgets and the number of decisions.

5.1 Illustrative Subjects

In the Introduction, we provide an overview of the important aggregate features of our experimental data, which we summarize by reporting the distributions of our cost-efficiency indices of rationalizability (e^*), FOSD-rationalizability (e^{**}), and EUT-rationalizability (e^{***}). But the aggregate data tell us little about the choice behavior of individual subjects. To get some idea of the wide range of observed behaviors, we present in Figure 4 scatterplots depicting all 50 choices for five illustrative subjects. We have chosen subjects whose behavior corresponds to one of several prototypical choices and illustrates the striking regularity within subjects and heterogeneity across subjects that is characteristic of our data. These selected case studies also help us get a broad sense about some of the common challenges to rationalizability.

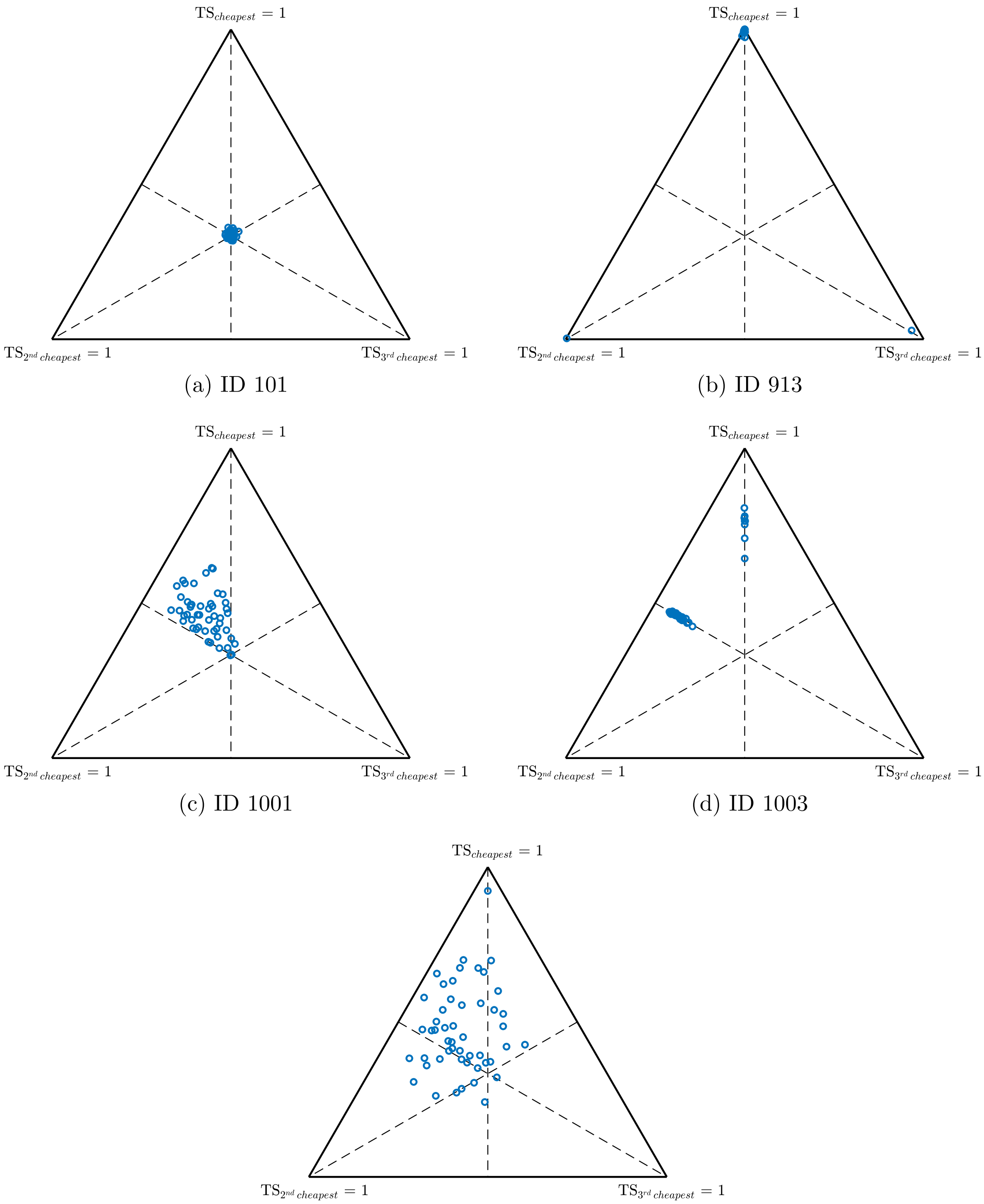
Figure 4 depicts the choices in terms of token shares for the three securities as points in the unit simplex. For each allocation $\mathbf{x}^i = (x_1^i, x_2^i, x_3^i)$, we relabel the states $s = 1, 2, 3$, so that $p_1^i < p_2^i < p_3^i$ and define the *token share* of the security that pays off in state s to be the number of tokens payable in state s as a fraction of the sum of tokens payable across states

$$\bar{x}_s^i = \frac{x_s^i}{x_1^i + x_2^i + x_3^i},$$

and $\bar{\mathbf{x}}^i = (\bar{x}_1^i, \bar{x}_2^i, \bar{x}_3^i)$ is the vector of token shares corresponding to the allocation $\mathbf{x}^i$. Each panel of Figure 4 contains a scatterplot of the token share vectors corresponding to the 50 allocations chosen by one of the five illustrative subjects. The vertices of the unit simplex correspond to allocations consisting of one of the three securities, and each point in the simplex represents an allocation as a convex combination of the extreme points.

The behaviors of the first three subjects are roughly EUT-rationalizable. In the scatterplot for subject ID 101 (Figure 4a), all of the vectors of token shares lie near the *center* of the simplex where $\bar{\mathbf{x}}^i = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$; this behavior is consistent with infinite risk aversion. In the scatterplot for subject ID 913 (Figure 4b), the token shares are all concentrated on (or, in a few cases, adjacent to) the top *vertex* of the simplex where $\bar{\mathbf{x}}^i = (1, 0, 0)$; this behavior is consistent with risk neutrality. A more complex behavior is illustrated in the scatterplot for subject ID 1001 (Figure 4c). The choices of this subject roughly equalize expenditures $p_1^i x_1^i = p_2^i x_2^i = p_3^i x_3^i$, rather than tokens, across the three securities; this behavior is consistent with maximizing a logarithmic von Neumann-Morgenstern expected utility function.

The next two subjects are *not* EUT-rationalizable. In the scatterplot for subject ID 1003 (Figure 4d), all token shares lie roughly along the *bisectors* of the angles of the simplex where $\bar{x}_1^i = \bar{x}_2^i$ or $\bar{x}_2^i = \bar{x}_3^i$; this behavior—equalizing the demands for two out of the three securities for a non-negligible set of price vectors—is FOSD-rationalizable (because $\bar{x}_1^i \geqslant \bar{x}_2^i \geqslant \bar{x}_3^i$ while $p_1^i < p_2^i < p_3^i$) but not EUT-rationalizable. However, preferences generated by, for example, rank-dependent utility (Quiggin, 1982, 1993) could give rise to such choices. Finally, in the scatterplot for subject ID 1105 (Figure 4e), the token shares are not confined to the top left



(a) ID 101 (b) ID 913

(c) ID 1001 (d) ID 1003

(e) ID 1105

Figure 4: Subject Behavior

Each plot shows all 50 choices for a single subject in terms of token shares. Each vertex of the unit simplex corresponds to a full allocation to one of the three securities. Some subjects are roughly EUT-rationalizable: (a) ID 101 is consistent with infinite risk aversion; (b) ID 913 is consistent with risk neutrality; (c) ID 1001 is consistent with the maximization of logarithmic von Neumann-Morgenstern expected utility. Some subjects are distinctly *not* EUT-rationalizable: (d) ID 1003 is FOSD-rationalizable and could be explained by rank-dependent utility; and (e) ID 1105 is not FOSD-rationalizable.

subset of the simplex where $\bar{x}_1^i \geqslant \bar{x}_2^i \geqslant \bar{x}_3^i$; such behavior is not FOSD-rationalizable (and thus also not EUT-rationalizable).[18]

5.2 WARP and GARP

Before analyzing the various measures of rationalizability, we first provide a basic description of the individual-level revealed preference violations. Table 1 reports percentile values of the numbers of WARP violations, GARP violations, and GARP violations that do not contain a WARP violation, alongside percentile values of the CCEI scores required to remove all violations of WARP and GARP. Recall that the number of GARP violations is the number of distinct revealed preference cycles, and that a WARP violation is a special case of a GARP violation involving a pairwise revealed preference cycle (see Section 3.1). With only two goods, all GARP violations must contain one or more WARP violations, which is not the case with three (or more) goods.

		Violations			CCEI Scores	
		WARP	GARP	GARP not WARP	WARP	GARP
Percentile Values	1	0	0	0	0.668	0.668
	5	0	0	0	0.787	0.787
	10	0	0	0	0.828	0.828
	25	1	1	0	0.916	0.913
	50	4	7	0	0.969	0.968
	75	8	318	0	0.996	0.996
	90	14	181,655	1	1.000	1.000
	95	25	$\geqslant 10^7$	7	1.000	1.000
	99	55	$\geqslant 10^7$	24	1.000	1.000

Table 1: WARP and GARP

The table reports percentile values of the numbers of WARP violations, GARP violations, and GARP violations that do not contain a WARP violation, alongside percentile values of the CCEI scores required to remove all violations of WARP and GARP. Due to computational constraints, we are unable to fully compute the number of GARP violations that do not contain a WARP violation for 3 subjects (1.8 percent), each of whom has more than 10^7 GARP violations (even after removing WARP violations); for these subjects, we obtain a lower bound on the number of GARP violations that do not contain a WARP violation.

[18] We have shown just a small and selected subset of the full set of subjects, and these are of course special cases where regularities in the data are very clear. For most subjects, the behavioral regularities are much less clear. However, a full review of the data reveals both regularities within subjects and heterogeneity across subjects. The scatterplots for the full set of subjects are available upon request.

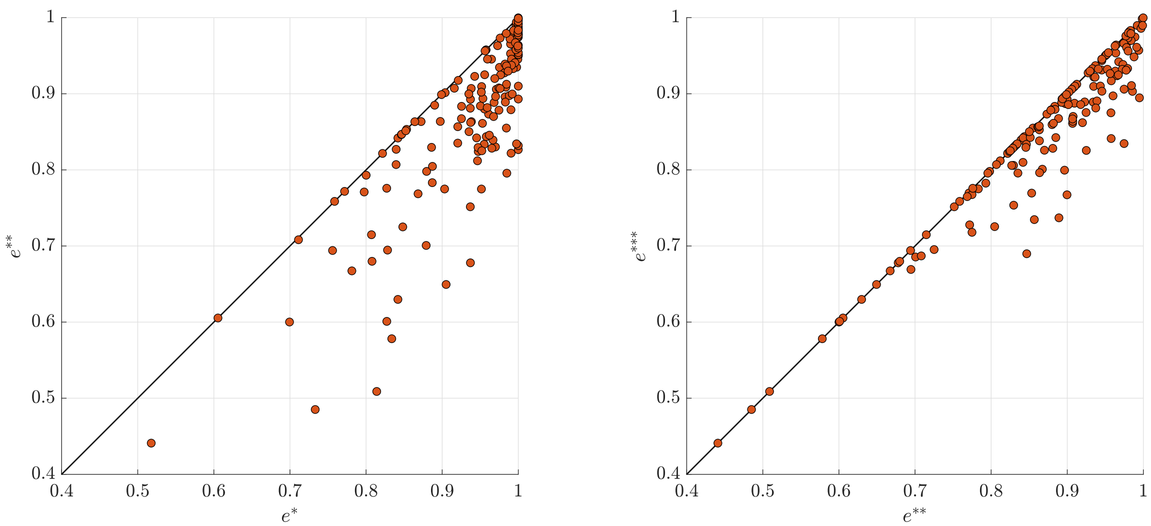
We see from Table 1 that the median number of WARP violations is 4 and the median number of GARP violations is 7. While the number of GARP violations is very high among a small number of subjects, the vast majority of these GARP violations contain WARP violations. In fact, only 23 subjects (13.7 percent) have violations of GARP that do not contain a violation of WARP. Furthermore, *all* of the 27 subjects (16.1 percent) who are free of WARP violations are also free of GARP violations. In other words—and importantly for our analysis—every subject who violates GARP also violates WARP, and so cannot be rationalized by a complete preference, even if it is allowed to be nontransitive. More generally, we can (for each subject) calculate the (usual) CCEI and also the CCEI which measures the amount by which each budget constraint needs to be reduced in order to remove all violations of WARP. The latter must (by definition) be weakly greater than the former, but the scores turn out to be identical for all but 4 subjects (2.4 percent). From the last two columns of Table 1 we see that the distributions of the two CCEIs are almost identical, suggesting once again a limited role for models that allow for nontransitive preferences.

5.3 Rationalizability Scores

As a first basic check on the rationalizability (e^*), FOSD-rationalizability (e^{**}), and EUT-rationalizability (e^{***}) of individual subjects, Figure 5 shows scatterplots of e^* against e^{**} (Figure 5a) and of e^{**} against e^{***} (Figure 5b). By definition, $e^* \geqslant e^{**} \geqslant e^{***}$ so all points in both scatterplots must lie on or below the 45-degree line. An individual subject who is perfectly EUT-rationalizable will have $1 = e^* = e^{**} = e^{***}$. When e^{***} is strictly less than one, the corresponding values of e^* and e^{**} will then allow us to isolate the source of the subject's departure from EUT.

Out of our 168 subjects, the choices of only 27 subjects (16.1 percent) are perfectly rationalizable ($e^* = 1$), but the choices of *none* of our subjects are perfectly FOSD-rationalizable ($e^{**} = 1$), and hence perfectly EUT-rationalizable ($e^{***} = 1$). Most interestingly, only 11 subjects (6.5 percent) fall along the 45-degree line in the scatterplot of e^* against e^{**} (Figure 5a); the choices of these subjects are not necessarily perfectly rationalizable but they are not less FOSD-rationalizable than they are rationalizable ($e^* = e^{**}$). By contrast, 65 subjects (38.7 percent) fall along the 45-degree line in the scatterplot of e^{**} against e^{***} (Figure 5b);

the choices of these subjects are not perfectly FOSD-rationalizable but they are not less EUT-rationalizable than they are FOSD-rationalizable ($e^{**} = e^{***}$). Only 3 subjects (1.8 percent), fall along the 45-degree line in both scatterplots; the choices of these subjects are not less EUT-rationalizable than they are rationalizable ($e^{*} = e^{**} = e^{***}$).



(a) e^{*} versus e^{**} (b) e^{**} versus e^{***}

Figure 5: Scatterplots of Rationalizability Scores

The plots depict rationalizability scores for individual subjects. By definition, $e^{*} \geqslant e^{**} \geqslant e^{***}$, so all points in both scatterplots must lie on or below the 45-degree line.

Furthermore, we compare the *magnitudes* of differences between scores. Figure 6 shows a scatterplot of the difference between *perfect* rationalizability and FOSD-rationalizability ($1 - e^{**}$) against the difference between FOSD-rationalizability and EUT-rationalizability ($e^{**} - e^{***}$). Out of our 168 subjects, 143 (85.1 percent) fall below the 45-degree line in the scatterplot, so the difference between perfect rationalizability and FOSD-rationalizability ($1 - e^{**}$) is larger for these subjects than the difference between FOSD-rationalizability and EUT-rationalizability ($e^{**} - e^{***}$). Furthermore, for 125 subjects (74.4 percent) the difference between perfect rationalizability and FOSD-rationalizability ($1 - e^{**}$) is *twice* as large as the difference between FOSD-rationalizability and EUT-rationalizability ($e^{**} - e^{***}$). Finally, 65 subjects (45.5 percent) fall along the horizontal axis ($e^{**} = e^{***}$). For these subjects, there is no difference between FOSD-rationalizability and EUT-rationalizability.

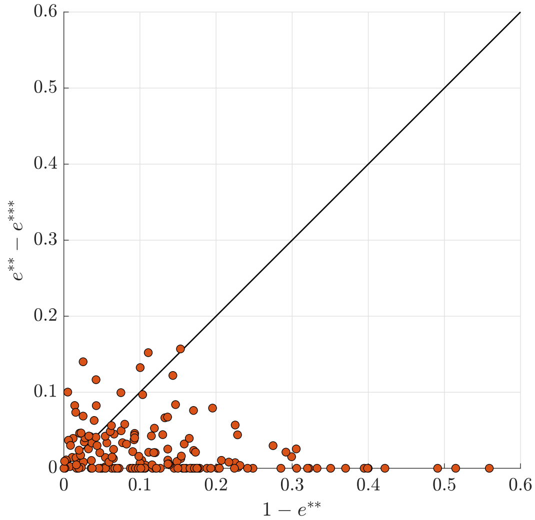


Figure 6: Scatterplot of Rationalizability Score Differences

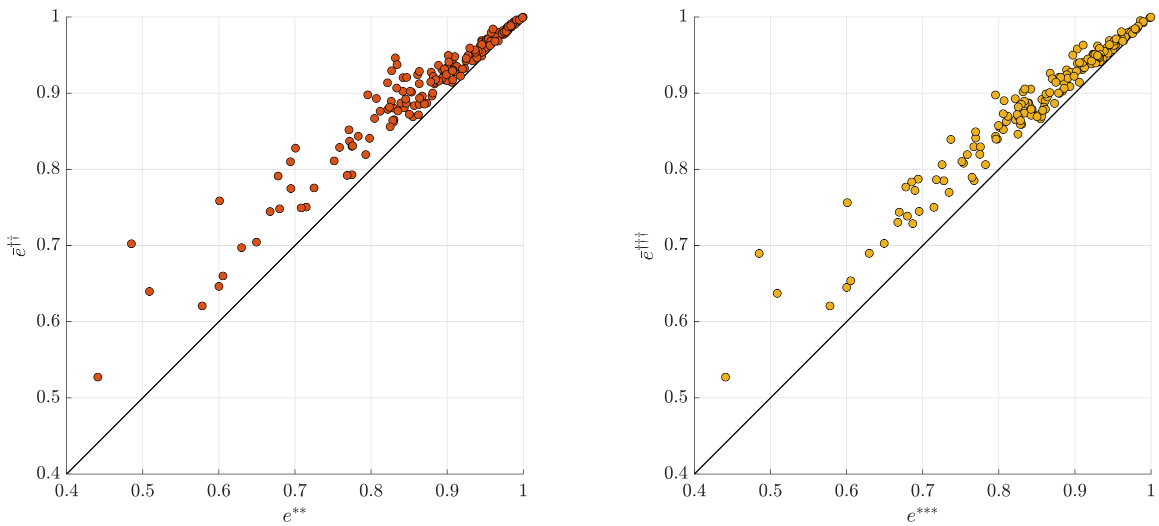
The plot depicts rationalizability score differences for individual subjects: the difference between *perfect* rationalizability and FOSD rationalizability $(1 - e^{**})$ against the difference between FOSD-rationalizability and EUT-rationalizability $(e^{**} - e^{***})$.

Hence, for the vast majority of our subjects there is only a small (or no) difference between FOSD-rationalizability and EUT-rationalizability $(e^{**} - e^{***})$, whereas the difference between perfect rationalizability and FOSD-rationalizability $(1 - e^{**})$ is much larger. For these subjects, there is little scope for the most prominent non-EUT alternatives that relax the independence axiom, such as weighted expected utility, rank-dependent utility, or reference-dependent risk preferences, to explain observed behavior, as they all postulate FOSD-rationalizability $(1 = e^{*} = e^{**} > e^{***})$.

5.4 Difference-in-Differences

Our individual-level data also allow us to compare the difference between *perfect* rationalizability and FOSD-rationalizability to the difference between FOSD-rationalizability and EUT-rationalizability, using a purely nonparametric difference-in-differences approach. For each subject, we first randomly draw 1,000 subsets of 25 (out of 50) observations where each subset is drawn without replacement. We then calculate (across 25 observations) 1,000 pairs of scores $(e^{\dagger\dagger}, e^{\dagger\dagger\dagger})$ for FOSD-rationalizability $(e^{\dagger\dagger})$ and EUT-rationalizability $(e^{\dagger\dagger\dagger})$ and denote the mean scores within this sample by $\bar{e}^{\dagger\dagger}$ and $\bar{e}^{\dagger\dagger\dagger}$, respectively.

As a prelude to the statistical analysis, Figure 7 shows scatterplots of e^{**} against $\bar{e}^{\dagger\dagger}$ (Figure 7a), and e^{***} against $\bar{e}^{\dagger\dagger\dagger}$ (Figure 7b). Since $\bar{e}^{\dagger\dagger} \geqslant e^{**}$ and $\bar{e}^{\dagger\dagger\dagger} \geqslant e^{***}$, all points in the two scatterplots must lie on or above the 45-degree line. The scores depicted in each panel of Figure 7 are very highly correlated—the correlation coefficient between e^{**} and $\bar{e}^{\dagger\dagger}$ is 0.970, and between e^{***} and $\bar{e}^{\dagger\dagger\dagger}$ the correlation coefficient is 0.978. As a practical note on the experiment, the high concordance in the two sets of scores suggests that subjects' mistakes/errors are unlikely to be due to occasionally lapsing into quasi-random behavior and/or adopting a low-effort heuristic that would generate more dispersion between their scores for half the dataset and for the full dataset.



(a) e^{**} versus $\bar{e}^{\dagger\dagger}$ (b) e^{***} versus $\bar{e}^{\dagger\dagger\dagger}$

Figure 7: Rationalizability Score Correlations

The plots depict rationalizability scores for individual subjects. By definition, $\bar{e}^{\dagger\dagger} \geqslant e^{**}$ and $\bar{e}^{\dagger\dagger\dagger} \geqslant e^{***}$, so all points in the scatterplots must lie on or above the 45-degree line.

We are now in a position to provide our individual-level difference-in-differences test. The test builds on the revealed preference techniques used above. It is fully nonparametric, making no assumptions about the form of the subject's underlying utility function, and allowing for the reality that a subject's behavior is not necessarily perfectly rationalizable. For each subject, the null hypothesis is given by

$$H_0 : D_1 = (1 - \mu_{e^{\dagger\dagger}}) - (\mu_{e^{\dagger\dagger}} - \mu_{e^{\dagger\dagger\dagger}}) = 0.$$

That is, that there is no difference between $1 - \mu_{e^{\dagger\dagger}}$ (the mean difference between *perfect* rationalizability and FOSD-rationalizability) and $\mu_{e^{\dagger\dagger}} - \mu_{e^{\dagger\dagger\dagger}}$ (the mean difference between

FOSD-rationalizability and EUT-rationalizability). For each subject, we use the 1,000 pairs of scores $(e^{\dagger\dagger}, e^{\dagger\dagger\dagger})$ to calculate the difference between $1 - \bar{e}^{\dagger\dagger}$ and $\bar{e}^{\dagger\dagger} - \bar{e}^{\dagger\dagger\dagger}$, denoted by $\bar{D}_1$, which is an estimator for D_1. To obtain the distribution of the estimator $\bar{D}_1$, we re-sample (with replacement) from the original sample consisting of 1,000 pairs of scores $(e^{\dagger\dagger}, e^{\dagger\dagger\dagger})$ to create a bootstrapped sample of the same size; we repeat this procedure 10^6 times.

Out of our 168 subjects, $\bar{D}_1$ is positive and statistically significant at the 1 percent level for 142 subjects (84.5 percent). The $\bar{e}^{\dagger\dagger}$ scores for 112 (66.7 percent) and 58 (34.5 percent) of our subjects are above 0.9 and 0.95, respectively. Even out of these highly FOSD-rationalizable subjects with $\bar{e}^{\dagger\dagger}$ scores above 0.9 and 0.95, $\bar{D}_1$ is positive and statistically significant at the 1 percent significance level for 87 (77.7 percent) and 38 (65.5 percent) subjects, respectively.

Finally, we strengthen our difference-in-differences test by using a "double-differencing" strategy under the null hypothesis is that $1 - \mu_{e^{\dagger\dagger}}$ is *twice* as large as $\mu_{e^{\dagger\dagger}} - \mu_{e^{\dagger\dagger\dagger}}$. We let $\bar{D}_2$ be the mean of this difference-in-differences (which is an estimator for D_2 defined analogously to D_1) and find that it is positive and statistically significant at the 1 percent level for 118 out of 168 subjects (70.2 percent). To summarize our analysis, Figure 8 shows the rationalizability score differences for individual subjects. Subjects are depicted in blue if only $\bar{D}_1$ is positive and statistically significant at the 1 percent level, in red if both $\bar{D}_1$ and $\bar{D}_2$ are positive and statistically significant, and in gray if neither.

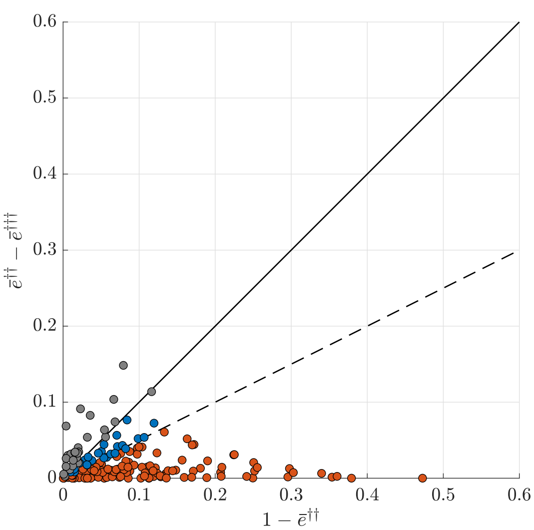


Figure 8: Scatterplot of Rationalizability Score Differences

The plot depicts rationalizability score differences for individual subjects. The solid line corresponds to the D_1 test and the dashed line to the D_2 test.

5.5 Power

A natural question to ask is whether the preceding empirical findings are driven mechanically by the structure of the experimental design. To cast light on this issue, we again generate a benchmark comparison using simulated subjects whose choices are uniformly distributed on the budget sets, but this time conditional on some FOSD-rationalizability score value. The simulated subjects make choices from 25 budget sets drawn from the same distribution as the actual subjects. For each simulated subject, we calculate a difference-in-differences. Precise details of the simulation procedure can be found in the Appendix and the results are shown in Figure 9.

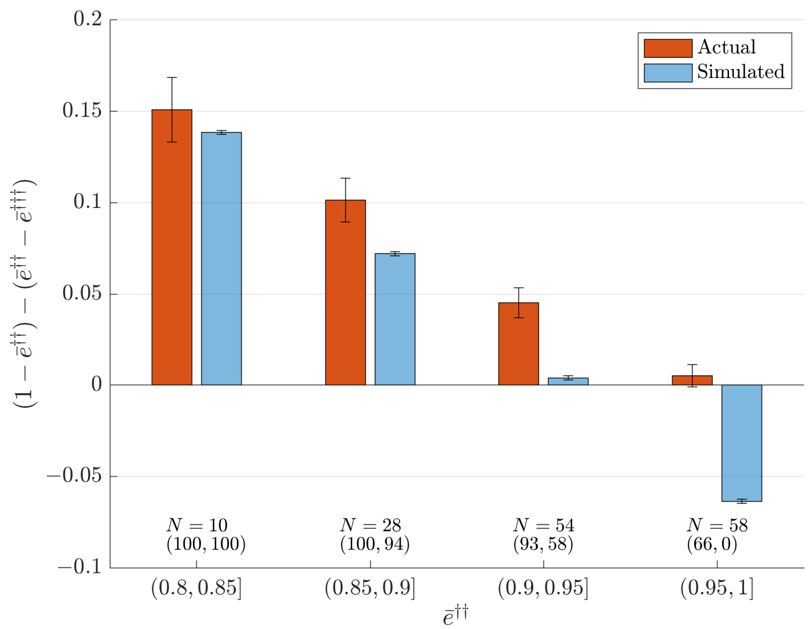


Figure 9: Power of Difference-in-Differences

To interpret the bars, consider the interval $(0.9, 0.95]$. The number of actual subjects with $\bar{e}^{\dagger\dagger}$ in this interval is 54. Among these subjects, the mean value of the difference-in-differences $(1-\bar{e}^{\dagger\dagger})-(\bar{e}^{\dagger\dagger}-\bar{e}^{\dagger\dagger\dagger})$ is 0.045. Among the simulated subjects on the same interval, the mean value is 0.004. The braces represent 95 percent confidence intervals. The proportion of actual subjects for whom the difference-in-differences is positive is 93 percent, as compared to 58 percent among simulated subjects.

Our basic contention is that the pattern of difference-in-differences $(1 - e^{\dagger\dagger}) - (e^{\dagger\dagger} - e^{\dagger\dagger\dagger})$ which occurs among the simulated subjects is very different from that which occurs among the actual subjects. As an example, among the 5,000 simulated subjects with an FOSD-rationalizability score value on the interval $(0.9, 0.95]$, the mean difference-in-differences is

0.004. In contrast, there are 54 actual subjects with $\bar{e}^{\dagger\dagger}$ on the same interval, and the mean difference-in-differences is 0.045. So the difference-in-differences among actual subjects is (on average) larger than among simulated subjects, and (as indicated in Figure 9) that difference is statistically significant. Furthermore, 93 percent of actual subjects have a positive difference-in-differences, as compared to 58 percent among simulated subjects. Across the intervals, we observe a clear pattern: the mean difference-in-differences is significantly higher among actual subjects than simulated subjects. Note that 140 subjects (83.3 percent) have $\bar{e}^{\dagger\dagger}$ above 0.85, which is where the experimental design is most discriminating. We therefore conclude that our empirical results are not merely an artifact of the experimental design.

5.6 Further Analysis

The broad conclusion from our analysis of rationalizability scores is clear: there could be multiple sources of EUT violations even for a single subject and, for a large majority of subjects, violations of ordering and monotonicity are prominent and greater in magnitude than violations of the independence axiom. This finding limits the applicability of the most prominent non-EUT alternatives, such as weighted expected utility or rank-dependent utility, to explain observed behavior since they all postulate FOSD-rationalizability. In this subsection we discuss further analysis that confirms the robustness of these findings.

5.6.1 Distance-Based Indices

In additional to our analysis using the CCEI, we also carry out a difference-in-differences analysis using the distance-based index (explained in Section 3.5). Such an analysis requires us to calculate the index for FOSD-rationalizability and EUT-rationalizability. The former can be obtained following Hu *et al.* (2021), but we are not aware of any computationally efficient method of calculating the index for EUT-rationalizability. To get around this difficulty, we divide each subject's 50 observations into five subsets of 10 observations each. For each of these "mini datasets" we have a way of calculating an upper bound on the index for EUT-rationalizability. We then obtain, for each subject and on each subset of 10 observations, the index for FOSD-rationalizability k^{**} and an *upper bound* on the index for EUT-rationalizability k_u^{***}. Taking an average over the five subsets of 10 observations, we obtain for each subject, $\bar{k}^{**}$ and $\bar{k}_u^{***}$. Obviously, the average value of the index for EUT-rationalizability, $\bar{k}^{***}$, must satisfy $\bar{k}^{***} \leqslant \bar{k}_u^{***}$.

Since EUT-rationalizability is more stringent than FOSD-rationalizability, we must have $\bar{k}^{***} \geqslant \bar{k}^{**}$. The central question is about the relative size of the gap—at its most extreme, if $\bar{k}^{***} \gg \bar{k}^{**} \approx 0$, then we conclude that the agent could be maximizing an FOSD-increasing utility function but is not EUT-rationalizable. Out of our 168 subjects, we find that $2(\bar{k}_u^{***} - \bar{k}^{**}) < \bar{k}^{**}$ (hence $2(\bar{k}^{***} - \bar{k}^{**}) < \bar{k}^{**}$ since $\bar{k}^{***} \leqslant \bar{k}_u^{***}$) for 124 subjects (73.8 percent), and $4(\bar{k}_u^{***} - \bar{k}^{**}) < \bar{k}^{**}$ for 99 subjects (58.9 percent). Therefore, for most subjects the additional perturbation required to guarantee EUT-rationalizability beyond FOSD-rationalizability is relatively modest (see the Appendix for details).

5.6.2 Two-Dimensional Data

We also analyze data from 956 subjects making portfolio decisions in two-state experiments with equally likely states.[19] The results we obtain from the two-state experiments are broadly similar to those obtained from the three-state experiment. In particular, for the vast majority of subjects violations of basic ordering and monotonicity are more prominent than violations of independence. Indeed, for 850 out of 956 subjects (88.9 percent), $\bar{D}_1 = (1-\bar{e}^{\dagger\dagger})-(\bar{e}^{\dagger\dagger}-\bar{e}^{\dagger\dagger\dagger})$ is positive and statistically significant at the 1 percent level (see the Appendix for details).

Finally, there is also a small pool of 46 subjects (taken from Choi *et al.* (2007a)) where the two states occur with *unequal* probabilities. For reasons which we provide in the Appendix, this asymmetric environment is not ideal for carrying out our difference-in-differences analysis. However, even in this case, departures from ordering and monotonicity are just as significant as departures from independence (see the Appendix for details).

6 Related Literature

We will not attempt to review the vast theoretical and experimental body of work on decision making under risk.[20] Instead, we focus attention on some recent papers that are particularly

[19]The data include data collected by Choi *et al.* (2007a), similar data using subject pools collected by Zame *et al.* (2020) and Cappelen *et al.* (2023), as well as new data. These experiments are identical to the one in this paper, except that there are two states rather than three.

[20]See Camerer (1995) and Starmer (2000) for excellent surveys. Camerer and Weber (1992) and Harless and Camerer (1994) also summarize the experimental evidence from testing the various utility theories of

relevant to our study. Choi *et al.* (2007a) employs graphical representations of budget lines containing bundles of state-contingent commodities, which allows for the collection of very rich individual-level datasets. In contrast to earlier work, the purpose of Choi *et al.* (2007a) is not to uncover violations of particular axioms, but rather to provide a positive account of choice under risk in a rich choice environment that allows for a general characterization of the patterns of individual behavior.

For each subject in their experiment, Choi *et al.* (2007a) tests the data for consistency with GARP and estimates a parametric utility function, which can be motivated by loss/disappointment aversion (Gul, 1991) and embeds EUT as a parsimonious and tractable special case.[21] But testing EUT as a restriction on a non-EUT utility function has an obvious drawback—it depends on assumptions over functional form and the specification of the error structure, as shown by Halevy, Persitz, and Zrill (2018).

While consistency with GARP is implied by—and guarantees—choice from a consistent preference ordering, *any* such preferences that are locally nonsatiated are admissible. In particular, choices can be compatible with GARP and yet fail to be reconciled with the maximization of a utility function that is monotonic with respect to FOSD, which is not normatively appealing. One is thus naturally led to go beyond basic consistency and to ask whether choices are also compatible with a utility function that has some special structure, in particular one which is monotonic with respect to FOSD and/or adheres to EUT.

Originating in the works of Varian (1983a,b, 1988) and Green and Srivastava (1986), some recent papers that pursue these questions include Diewert (2012), Bayer *et al.* (2013), Kubler, Selden, and Wei (2014), Echenique and Saito (2015), Chambers, Liu, and Martinez (2016), Chambers, Echenique, and Saito (2016), Nishimura, Ok, and Quah (2017), Echenique, Imai, and Saito (2023), Polisson, Quah, and Renou (2020), and de Clippel and Rozen (2023). We compare our approach and contribution to the existing work along four key dimensions—methods, measures, tests, and power.

choice under risk and under uncertainty, and Kahneman and Tversky (2000) collects many theoretical and empirical papers that have emerged from their pioneering work on prospect theory.

[21]Following the seminal work of Hey and Orme (1994) and Harless and Camerer (1994), a number of other papers have estimated parametric utility functions. Harless and Camerer (1994) fits aggregate data, while Hey and Orme (1994) estimates functional forms at the level of the individual subject using decisions from a large menu of binary choices.

Methods. In comparison to previous work, our paper differs in the methods used for the collection and analysis of the data. The data from the experiment reported here involve three states with three associated securities whereas the data used in earlier experiments involve only two states and two associated securities. With only two states, WARP and GARP are observationally equivalent so incompleteness and intransitivity can only be tested jointly. With three states, by contrast, we can separate incompleteness from intransitivity. In the case of three states, prominent non-EUT models generate different structures for the utility function, thus yielding a larger set of empirical restrictions on observed behavior against with EUT can be tested.

Our test of EUT-rationalizability relies on the GRID method developed in Polisson, Quah, and Renou (2020). With the exception of the GRID method, all other revealed preference tests of EUT involve a *concave* Bernoulli index. The GRID method, by contrast, neither assumes nor guarantees concavity. This distinction is by no means cosmetic, since it has empirical implications. Although concavity of the Bernoulli index, which is equivalent to risk aversion under EUT, is widely assumed in empirical applications, we avoid imposing any further requirements that are not, strictly speaking, a part of EUT as such.[22] This feature of our analysis is an important part of our claim that our tests are purely nonparametric, with no extraneous assumptions on the parametric form or shape of the utility function.

Measures. Revealed preference relations generate exact tests while choice data almost always contain some violations. Given this, any serious empirical investigation requires an index to measure a model's goodness-of-fit, or (in other words) the extent to which a subject's choices are (in)compatible with the model. In this paper, we use Afriat's (1973) CCEI to measure a subject's consistency with (basic) rationalizability (e^*), FOSD-rationalizability (e^{**}), and EUT-rationalizability (e^{***}). Since the models are nested, the indices must be ordered for any given subject, with $1 \geqslant e^* \geqslant e^{**} \geqslant e^{***} > 0$. The use of a common index across different models means that we can perform a comprehensive test of each model (in which all of the axioms underpinning the model

[22]Indeed, there are datasets which are EUT-rationalizable but only with a nonconcave Bernoulli index. For such an example, see Section A4 of the Online Appendix in Polisson, Quah, and Renou (2020). A fuller discussion of these distinctions can also be found in Polisson, Quah, and Renou (2020).

are tested in combination) and at the same time cleanly identify the incremental impact of additional axioms.

We employ the CCEI (rather than some other index) for two related reasons (as discussed in Section 3.5): it is straightforward to compute for the models under consideration and it is economically interpretable (see Varian (1990)). See also Dziewulski (2020) for a behavioral interpretation of the CCEI based on a decision maker's cognitive inability to distinguish between similar bundles. For these reasons, the CCEI is the most commonly used measure of goodness-of-fit in the revealed preference literature. Another index proposed by Varian (1990) is closely related to the CCEI and has been used in some important work (see, for example, Halevy, Persitz, and Zrill (2018)). There are known methods for calculating this index for the different models that we consider, but its calculation is much more computationally demanding than the CCEI (especially for EUT-rationalizability) and therefore it is not practically implementable for us, given the size of our datasets and the scope of our empirical exercise.[23,24]

Tests. We implement novel individual-level econometric tests. The approach builds only on revealed preference techniques and it is purely nonparametric, making no assumptions about the form of the subject's underlying utility function or on the error structure. That is, for each individual subject we obtain the (empirical) distribution function for the test statistic under the null hypothesis—that the difference between *perfect* rationalizability and FOSD-rationalizability and the difference between FOSD-rationalizability and EUT-rationalizability are equal—using a purely nonparametric difference-in-differences econometric approach.

[23]For more on the computation of Varian's (1990) index to measure rationalizability, FOSD-rationalizability, and EUT-rationalizability, see Polisson, Quah, and Renou (2020). One advantage of Varian's (1990) index is that it is generally less sensitive to a single errant observation as compared to the CCEI. We address this sensitivity issue through our sub-sampling procedure.

[24]de Clippel and Rozen (2023) proposes a new index to measure goodness-of-fit that applies to different families of utility functions; roughly speaking, the index is based on the size of the departures from the first-order conditions. Building on the methodology in Echenique, Imai, and Saito (2020) within the context of intertemporal choice, Echenique, Imai, and Saito (2023) proposes essentially the same index as de Clippel and Rozen (2023) for expected utility, albeit with a somewhat different motivation. This index (or collection of indices) relies on a first-order (condition) approach, which is only applicable to models representable by quasiconcave utility functions (defined on the space of contingent consumption). We avoid imposing a concave Bernoulli index (or, more generally, a quasiconcave utility function) as a rationality requirement.

Power. A number of recent papers—including Polisson, Quah, and Renou (2020), de Clippel and Rozen (2023), and Echenique, Imai, and Saito (2023)—analyze the experimental data from Choi *et al.* (2014). This experiment is identical to Choi *et al.* (2007a), except that it consists of 25, rather than 50, decision problems involving two (equally likely) states of nature and two associated securities. Echenique, Imai, and Saito (2023) also analyzes the experimental data from Carvalho, Meier, and Wang (2016) and Carvalho and Silverman (2019), which also consist of 25 problems. The Choi *et al.* (2007a) data have also been extensively analyzed, including by Halevy, Persitz, and Zrill (2018) and Polisson, Quah, and Renou (2020).

The experiment reported in this paper consists of 50 decision problems involving three equally likely states and three associated securities. Collecting 50, or even 25, individual decisions is more than is usual in the experimental literature on choice under risk and, as Choi *et al.* (2014) show, it does provide a rich enough individual-level dataset for a powerful test of (basic) rationalizability. Furthermore, our power analysis indicates that having three states significantly enhances the discriminatory power of the experiment, especially with respect to EUT-rationalizability.

To conclude this section, we compare our empirical findings vis-à-vis a few closely-related recent papers. Both de Clippel and Rozen (2023) and Echenique, Imai, and Saito (2023) develop new methodologies and apply their techniques to existing experimental data which (unlike our newly collected data) is obtained from two-dimensional experiments. Notwithstanding the use of a different measure of rationalizability, de Clippel and Rozen (2023) draws a similar conclusion to ours, namely that the gap between FOSD-rationalizability and EUT-rationalizability is small for many subjects; however, as acknowledged by the authors, power issues cast doubt on the robustness of their empirical conclusions. Echenique, Imai, and Saito (2023) finds that subjects who are more rationalizable (as measured by the CCEI) are not necessarily more EUT-rationalizable (as measured by their index). However, these two rationalizability measures are not formally comparable, so their analysis is not directed at separating the empirical validity of each of the axioms on which EUT is based, which is the principal interest in our exercise.

The principal aim of Polisson, Quah, and Renou (2020) is to develop the GRID method as a nonparametric revealed preference test for a broad class of models (including EUT); to demonstrate its practicality, the GRID method was applied to existing two-dimensional choice data under risk. We share Polisson, Quah, and Renou's (2020) aim to evaluate the empirical validity of the various axioms underlying theories of choice under risk and we also use the GRID method (among others), but our empirical analysis goes much further. The primary contribution of this paper is a combination of richer experimental data and new analytical/statistical methods. This experimental-analytical combination provides much richer guidance for understanding risk preferences and the choices that implement them. Specifically, one key advantage of our three-dimensional experiment over earlier two-dimensional experiments is that with only two goods incompleteness cannot be separated from intransitivity, which is not the case with three goods.[25] Furthermore, there is much greater separation among non-EUT models in the three-dimensional experiment, which thus provides a much stronger test in terms of the power to distinguish between EUT and non-EUT alternatives.

In a separate strand of recent research, Nielsen and Rehbeck (2022) tries to separate the normative from the descriptive value of a theory by allowing subjects to revise their choices; if many subjects choose to revise their decisions after being alerted to their violations of (say) the independence axiom, then the axiom has normative appeal even if it may not be descriptively accurate.[26] The goal of our paper (and indeed of all the papers cited above) is different—we want to evaluate EUT and related models as *descriptive* theories. To the extent that mistakes (in the sense of Nielsen and Rehbeck (2022)) are part of typical choice behavior and their variable severity across subjects is reflected in subjects' variable economic outcomes outside the lab (Choi *et al.*, 2014), it is certainly not our objective to remove such mistakes from the experiment.[27] That said, separating a theory's normative from its descriptive value

[25] In this paper we develop (and implement) a method to find all WARP violations, GARP violations, and GARP violations that do not contain a WARP violation. As a result, we can tell apart incomplete from nontransitive preferences, which cannot be done using data from previous two-dimensional experiments.

[26] In a famous encounter between Allais and Savage in 1952, Savage (in response to questions by Allais), first makes choices that contradict one of the core axioms of his subjective expected utility theory, only to acknowledge the mistake and correct his choices upon reflection. See Dietrich, Staras, and Sugden (2021) and the references therein for an account of this famous interaction and the issues it raises.

[27] In Breig and Feldman (2022) subjects are given random opportunities to revise their decisions but, unlike Nielsen and Rehbeck (2022), violations of axioms are not pointed out. The paper finds that various measures of decision quality improve with revisions. We could modify our experiment to allow for random revision opportunities, and that could improve rationalizability scores, but there is no reason to believe that this modification will upset our findings, which are about the *relative* performance of different models.

is an interesting issue and developing experimental approaches that allow for this separation in the context of budgetary decision making is an important extension.

7 Concluding Remarks

The standard model of choice under risk is based on von Neumann and Morgenstern's (1947) EUT. It is meant to serve as a normative guide for choice and also as a descriptive model of how individuals choose. However, much of the experimental and empirical evidence of "anomalies" in choice behavior suggests that EUT may not the right model. While EUT embodies three important axioms—ordering, monotonicity (with respect to FOSD), and independence—independence is the only axiom which the seminal alternatives to EUT relax.

It is thus natural that experimentalists should want to test the empirical validity of the independence axiom, and the overwhelming body of evidence against independence has raised criticisms about its status as the touchstone of rationality in the context of decision making under risk. In response to these criticisms, various generalizations of EUT have been developed, and the experimental examination of these theories has led to new empirical regularities in the laboratory. Starmer (2000) calls this the "conventional strategy"—theories/experiments designed to permit/test violations of independence (and weakened forms of independence) while retaining the more basic axioms of ordering and monotonicity.[28]

Combining theoretical tools, experimental methods, and nonparametric econometric techniques, our study confronts all of the axioms of EUT with individual-level experimental data that is richer than anything that has heretofore been used. The data are well-suited to purely nonparametric revealed preference tests which allow for the reality that individual behavior is not perfectly consistent with well-behaved preferences.

Why does this matter? It matters because choice data cannot be treated as being generated by a utility function, or by a utility function that is monotone with respect to FOSD, if there are large deviations from rationalizability or FOSD-rationalizability. In these cases, the standard approach of postulating some parametric family of utility functions (typically re-

[28]Bell (1982), Fishburn (1982), and Loomes and Sugden (1982) (simultaneously) propose a model of nontransitive risk preference. Loomes and Sugden (1982) develop a version of this model that involves regret with pairwise choice. Starmer (2000) provides an overview of these models and relates them to other non-EUT alternatives.

specting FOSD), and estimating its parameters leads to model misspecification. As a result, the estimated preference will not reflect the true underlying preference, if such a preference ordering even exists, and positive predictions and normative welfare conclusions based on these models will be misleading.[29]

Our findings also have implications for public policy; for example, in the practice of light paternalism, which is aimed at steering people toward better choices (Camerer *et al.*, 2003; Thaler and Sunstein, 2003; Loewenstein and Haisley, 2008). Clearly, decision makers that only violate independence merit greater deference from policy makers than the more boundedly rational ones that violate ordering and monotonicity because the choices of the former, unlike the latter, maximize a well-behaved utility function and are thus of a higher quality (Kariv and Silverman, 2013).

We hope that our empirical results might also stimulate new theories of how subjects choose contingent consumption bundles from budget sets, given the performance of the standard model with an FOSD-increasing utility function. One possible avenue consists of models where the chosen bundle is optimal among all alternatives in the budget according to some preference, but where the preference is not stable; an example would be a model where the agent's preference is formed around a reference point (such as the risk-free portfolio allocation) that varies with the budget set. Another avenue consists of models where the chosen bundle is optimal (according to a stable preference) among some but not all alternatives in the budget, because the agent is employing simplifying heuristics or has a consideration set (formed according to some rules) which is a strict subset of the budget.[30] We suspect that a more descriptively accurate theory needs to incorporate features from such approaches.

The experimental platform and analytical techniques that we have used are applicable to many other types of individual choice problems and decision domains. One important direction is to study choice under ambiguity. In a separate paper, we apply the GRID method

[29]Halevy, Persitz, and Zrill (2018) parametrically estimates preferences for the dataset collected by Choi *et al.* (2007a) involving two states and two associated securities. They find *significant quantitative and qualitative differences* between the preferences induced by parametric estimation and the revealed preferences implied by choices, due to model misspecification.

[30]Barseghyan, Molinari, and Thirkettle (2021) use a model with expected utility and random consideration sets to help explain households' deductible choices across different types of insurance coverage (which is essentially a choice problem over a finite set of risky alternatives). The use of limited consideration sets in their paper was motivated in part by the frequent observation of households making choices which are dominated (in a certain sense related to FOSD).

and other revealed preference techniques to the analogous data of Ahn *et al.* (2014) which similarly allow for a rigorous test of individual-level decision making under ambiguity. In ongoing work, we study patterns of economic rationality in intertemporal choice—specifically non-constant time discounting—by replacing the state-contingent assets in the experiment reported here with time-dated accounts. To our knowledge, similar evaluations of the rationality of intertemporal choice have not been performed.

References

Afriat, S. N. 1967. "The Construction of Utility Functions from Expenditure Data." *International Economic Review* 8(1): 67–77.

———. 1972. "Efficiency Estimation of Production Functions." *International Economic Review* 13(3): 568–598.

———. 1973. "On a System of Inequalities in Demand Analysis: An Extension of the Classical Method." *International Economic Review* 14(2): 460–472.

Ahn, D., S. Choi, D. Gale, and S. Kariv. 2014. "Estimating Ambiguity Aversion in a Portfolio Choice Experiment." *Quantitative Economics* 5(2): 195–223.

Allais, P. M. 1953. "Le Comportement de l'Homme Rationnel devant le Risque: Critique des Postulats et Axiomes de l'Ecole Americaine." *Econometrica* 21(4): 503–546.

Andreoni, J. and J. Miller. 2002. "Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism." *Econometrica* 70(2): 737–753.

Andreoni, J. and C. Sprenger. 2012. "Estimating Time Preferences from Convex Budgets." *American Economic Review* 102(7): 3333–3356.

Andreoni, J. and L. Vesterlund. 2001. "Which is the Fair Sex? Gender Differences in Altruism." *Quarterly Journal of Economics* 116(1): 293–312.

Banerjee, S. and J. H. Murphy. 2006. "A Simplified Test for Preference Rationality of Two-Commodity Choice." *Experimental Economics* 9: 67–75.

Barseghyan, L., F. Molinari, and M. Thirkettle. 2021. "Discrete Choice under Risk with Limited Consideration." *American Economic Review* 111(6): 1972–2006.

Bayer, R.-C., S. Bose, M. Polisson, and L. Renou. 2013. "Ambiguity Revealed." *IFS Working Papers* W13/05.

Bell, D. E. 1982. "Regret in Decision Making under Uncertainty." *Operations Research* 30(5): 961–981.

Blavatskyy, P., A. Ortmann, and V. Panchenko. 2022. "On the Experimental Robustness of the Allais Paradox." *American Economic Journal: Microeconomics* 14(1): 143–163.

Breig, Z. and P. Feldman. 2022. "Revealing Risky Mistakes Through Revisions." Unpublished paper.

Bronars, S. G. 1987. "The Power of Nonparametric Tests of Preference Maximization." *Econometrica* 55(3): 693–698.

Camerer, C. 1995. "Individual Decision Making." In *Handbook of Experimental Economics*, edited by J. H. Kagel and A. E. Roth. Princeton: Princeton University Press, 587–704.

Camerer, C., S. Issacharoff, G. Loewenstein, T. O'Donoghue, and M. Rabin. 2003. "Regulation for Conservatives: Behavioral Economics for 'Asymmetric Paternalism'." *University of Pennsylvania Law Review* 151: 1211–1254.

Camerer, C. and M. Weber. 1992. "Recent Developments in Modeling Preferences: Uncertainty and Ambiguity." *Journal of Risk and Uncertainty* 5(4): 325–370.

Cappelen, A. W., S. Kariv, E. Ø. Sørensen, and B. Tungodden. 2023. "The Development Gap in Economic Rationality of Future Elites." *Games and Economic Behavior* 142: 866–878.

Carvalho, L. and D. Silverman. 2019. "Complexity and Sophistication." *NBER Working Paper Series* Working Paper 26036.

Carvalho, L. S., S. Meier, and S. W. Wang. 2016. "Poverty and Economic Decision-Making: Evidence from Changes in Financial Resources at Payday." *American Economic Review* 106(2): 260–284.

Chambers, C. P., F. Echenique, and K. Saito. 2016. "Testing Theories of Financial Decision Making." *Proceedings of the National Academy of Sciences* 113(15): 4003–4008.

Chambers, C. P., C. Liu, and S.-K. Martinez. 2016. "A Test for Risk-Averse Expected Utility." *Journal of Economic Theory* 163: 775–785.

Chew, S. H. 1989. "Axiomatic Utility Theories with the Betweenness Property." *Annals of Operations Research* 19(2): 273–298.

Choi, S., R. Fisman, D. Gale, and S. Kariv. 2007a. "Consistency and Heterogeneity of Individual Behavior under Uncertainty." *American Economic Review* 97(5): 1921–1938.

———. 2007b. "Revealing Preferences Graphically: An Old Method Gets a New Tool Kit." *American Economic Review: AEA Papers and Proceedings* 97(2): 153–158.

Choi, S., S. Kariv, W. Müller, and D. Silverman. 2014. "Who Is (More) Rational?" *American Economic Review* 104(6): 1518–1550.

de Clippel, G. and K. Rozen. 2023. "Relaxed Optimization: How Close Is a Consumer to Satisfying First-Order Conditions?" *Review of Economics and Statistics* 104(4): 883–898.

Dekel, E. 1986. "An Axiomatic Characterization of Preferences under Uncertainty: Weakening the Independence Axiom." *Journal of Economic Theory* 40(2): 304–318.

Dietrich, F., A. Staras, and R. Sugden. 2021. "Savage's Response to Allais as Broomean Reasoning." *Journal of Economic Methodology* 28(2): 143–164.

Diewert, W. E. 1973. "Afriat and Revealed Preference Theory." *Review of Economic Studies* 40(3): 419–425.

———. 2012. "Afriat's Theorem and Some Extensions to Choice under Uncertainty." *Economic Journal* 122(560): 305–331.

Dziewulski, P. 2020. "Just-Noticeable Difference as a Behavioural Foundation of the Critical Cost-Efficiency Index." *Journal of Economic Theory* 188: 105071.

Echenique, F., T. Imai, and K. Saito. 2020. "Testable Implications of Models of Intertemporal Choice: Exponential Discounting and Its Generalizations." *American Economic Journal: Microeconomics* 12(4): 114–43.

———. 2023. "Approximate Expected Utility Rationalization." *Journal of the European Economic Association* 21(5): 1821–1864.

Echenique, F. and K. Saito. 2015. "Savage in the Market." *Econometrica* 83(4): 1467–1495.

Fishburn, P. C. 1982. "Nontransitive Measurable Utility." *Journal of Mathematical Psychology* 26(1): 31–67.

Fisman, R., P. Jakiela, and S. Kariv. 2015. "How Did the Great Recession Impact Social Preferences?" *Journal of Public Economics* 128: 84–95.

———. 2017. "Distributional Preferences and Political Behavior." *Journal of Public Economics* 155: 1–10.

Fisman, R., P. Jakiela, S. Kariv, and D. Markovits. 2015. "The Distributional Preferences of an Elite." *Science* 349(6254): 1300.

Fisman, R., P. Jakiela, S. Kariv, and S. Vannutelli. 2023. "The Distributional Preferences of Americans, 2013–2016." *Experimental Economics* 26: 727–748.

Fisman, R., S. Kariv, and D. Markovits. 2007. "Individual Preferences for Giving." *American Economic Review* 97(5): 1858–1876.

Green, R. C. and S. Srivastava. 1986. "Expected Utility Maximization and Demand Behavior." *Journal of Economic Theory* 38(2): 313–323.

Gul, F. 1991. "A Theory of Disappointment Aversion." *Econometrica* 59(3): 667–686.

Halevy, Y. and G. Mayraz. 2022. "Identifying Rule-Based Rationality." *Review of Economics and Statistics* (Forthcoming).

Halevy, Y, D. Persitz, and L. Zrill. 2018. "Parametric Recoverability of Preferences." *Journal of Political Economy* 126(4): 1558–1593.

Harbaugh, W. T., K. Krause, and T. R. Berry. 2001. "GARP for Kids: On the Development of Rational Choice Behavior." *American Economic Review* 91(5): 1539–1545.

Harless, D. W. and C. F. Camerer. 1994. "The Predictive Utility of Generalized Expected Utility Theories." *Econometrica* 62(6): 1251–1289.

Hey, J. D. and C. Orme. 1994. "Investigating Generalizations of Expected Utility Theory Using Experimental Data." *Econometrica* 62(6): 1291–1326.

Hu, G., J. Li, J. K.-H. Quah, and R. Tang. 2021. "A Theory of Revealed Indirect Preference." Unpublished paper.

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

Kahneman, D. and A. Tversky, editors. 2000. *Choices, Values, and Frames*. Cambridge: Cambridge University Press.

Kariv, S. and D. Silverman. 2013. "An Old Measure of Decision-Making Quality Sheds New Light on Paternalism." *Journal of Institutional and Theoretical Economics* 169(1): 29–44.

Kőszegi, B. and M. Rabin. 2007. "Reference-Dependent Risk Attitudes." *American Economic Review* 97(4): 1047–1073.

Kubler, F., L. Selden, and X. Wei. 2014. "Asset Demand Based Tests of Expected Utility Maximization." *American Economic Review* 104(11): 3459–3480.

Li, J., L. P. Casalino, R. Fisman, S. Kariv, and D. Markovits. 2022. "Experimental Evidence of Physician Social Preferences." *Proceedings of the National Academy of Sciences* 119(28): 1–11.

Li, J., W. Dow, and S. Kariv. 2017. "Social Preferences of Future Physicians." *Proceedings of the National Academy of Sciences* 114(48): 10291–10300.

Loewenstein, G. and E. Haisley. 2008. "The Economist as Therapist: Methodological Ramifications of 'Light' Paternalism." In *The Foundations of Positive and Normative Economics: A Handbook*, edited by A. Caplin and A. Schotter. Oxford: Oxford University Press, 210–245.

Loomes, G. and R. Sugden. 1982. "Regret Theory: An Alternative Theory of Rational Choice Under Uncertainty." *Economic Journal* 92(368): 805–824.

Machina, M. J. 1982. "'Expected Utility' Analysis without the Independence Axiom." *Econometrica* 50(2): 277–323.

Manzini, P. and M. Mariotti. 2008. "On the Representation of Incomplete Preferences over Risky Alternatives." *Theory and Decision* 65(4): 303–323.

Marschak, J. 1950. "Rational Behavior, Uncertain Prospects, and Measurable Utility." *Econometrica* 18(2): 111–141.

Masatlioglu, Y. and C. Raymond. 2016. "A Behavioral Analysis of Stochastic Reference Dependence." *American Economic Review* 106(9): 2760–2782.

Nielsen, K. and J. Rehbeck. 2022. "When Choices Are Mistakes." *American Economic Review* 112(7): 2237–2268.

Nishimura, H., E. Ok, and J. K.-H. Quah. 2017. "A Comprehensive Approach to Revealed Preference Theory." *American Economic Review* 107(4): 1239–1263.

Polisson, M. and J. K.-H. Quah. 2022. "Rationalizability, Cost-Rationalizability, and Afriat's Efficiency Index." Unpublished paper.

Polisson, M., J. K.-H. Quah, and L. Renou. 2020. "Revealed Preferences over Risk and Uncertainty." *American Economic Review* 110(6): 1782–1820.

Quiggin, J. 1982. "A Theory of Anticipated Utility." *Journal of Economic Behavior and Organization* 3(4): 323–343.

———. 1990. "Stochastic Dominance in Regret Theory." *Review of Economic Studies* 57(3): 503–511.

———. 1993. *Generalized Expected Utility Theory: The Rank-Dependent Model.* Dordrecht: Kluwer.

Rose, H. 1958. "Consistency of Preference: The Two-Commodity Case." *Review of Economic Studies* 25(2): 124–125.

Starmer, C. 2000. "Developments in Non-Expected Utility Theory: The Hunt for a Descriptive Theory of Choice under Risk." *Journal of Economic Literature* 38(2): 332–382.

Thaler, R. H. and C. R. Sunstein. 2003. "Libertarian Paternalism." *American Economic Review* 93(2): 175–179.

Tversky, A. and D. Kahneman. 1992. "Advances in Prospect Theory: Cumulative Representation of Uncertainty." *Journal of Risk and Uncertainty* 5(4): 297–323.

Varian, H. R. 1982. "The Nonparametric Approach to Demand Analysis." *Econometrica* 50(4): 945–973.

———. 1983a. "Non-Parametric Tests of Consumer Behaviour." *Review of Economic Studies* 50(1): 99–110.

———. 1983b. "Nonparametric Tests of Models of Investor Behavior." *Journal of Financial and Quantitative Analysis* 18(3): 269–278.

———. 1988. "Estimating Risk Aversion from Arrow-Debreu Portfolio Choice." *Econometrica* 56(4): 973–979.

———. 1990. "Goodness-of-Fit in Optimizing Models." *Journal of Econometrics* 46(1-2): 125–140.

von Neumann, J. and O. Morgenstern. 1947. *Theory of Games and Economic Behavior*. Princeton: Princeton University Press, 2nd ed.

Wakker, P. 1993. "Savage's Axioms Usually Imply Violation of Strict Stochastic Dominance." *Review of Economic Studies* 60(2): 487–493.

Zame, W. R., B. Tungodden, E. Ø. Sørensen, S. Kariv, and A. W. Cappelen. 2020. "Linking Social and Personal Preferences: Theory and Experiment." Unpublished paper.