IFS

Institute for Fiscal Studies

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

21/15 Working paper

Ever since Allais

Ever Since Allais*

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

June 8, 2021

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 which adhere to the conventional axioms of ordering and monotonicity. Unlike many existing laboratory experiments designed to test independence, our experiment systematically tests the entire set of axioms, providing much richer evidence against which EUT can be judged. Our within-subjects analysis is nonparametric, using only information about revealed preference relations in the individual-level data. For most subjects we find that departures from independence are statistically significant but minor relative to departures from ordering and/or monotonicity.

JEL Codes: D81, C91.

Keywords: rationality, expected utility, revealed preference, first-order stochastic dominance, experiment.

*We are grateful to David Dillenberger, Federico Echenique, Yoram Halevy, Yusufcan Masatlioglu, Peter Wakker, and William Zame for discussions, 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).

†Dembo: Interdisciplinary Center Herzliya (aluma.dembo@idc.ac.il); Kariv: University of California, Berkeley (kariv@berkeley.edu); Polisson: University of Bristol (matthew.polisson@bristol.ac.uk); Quah: Johns Hopkins University and National University of Singapore (john.quah@jhu.edu).

> "I can tell you of an important new result I got recently. I have what I suppose to be a completely general treatment of the revealed preference problem, which will give a fresh setting for the related work of Samuelson-Houthakker-Uzawa. Calculus methods are unavailable. The methods are set-theoretic or algebraical."
>
> — A letter from Sydney Afriat to Oskar Morgenstern, 1964

1 Introduction

Canonical decision-theoretic models of choice under risk consider a decision-maker who has a *complete* and *transitive* preference relation over the set of lotteries (probability measures) on a set of consequences (outcomes). By Debreu's (1954, 1960) theorem, any *continuous* preference relation can be represented by a continuous utility function, but any such continuous utility representation is admissible. For the utility function to have an expected utility representation, the preference relation must also satisfy the familiar von Neumann and Morgenstern (1947) *independence* axiom.

Expected utility theory (EUT) lies at the very heart of economics, and so it is natural that experimentalists would want to empirically test the axioms which characterize the EUT model. Empirical violations of these axioms generate intriguing questions about the rationality of individual behavior, and specifically raise criticisms of the independence axiom and its status as the touchstone for rational decision-making in the context of risk. In response to these criticisms, various generalizations of EUT have been formulated, and the experimental scrutiny of these theories has led to new empirical regularities in the laboratory.

Considerable effort has been put towards developing alternatives to EUT. Almost all of these models embody ordering (completeness and transitivity) and generalize EUT by weakening the independence axiom, while generally staying within the class of utility functions that are monotone (in other words, increasing) with respect to first-order stochastic dominance (FOSD); 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).[1,2] The accompanying experimental investigations for the most part use pairwise choices, à la Allais, to test EUT and its generalizations, presuming that subjects have well-defined preferences.

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 (and other models) in a choice environment where all features of the model(s) can be simultaneously evaluated. Our experiment and analysis draw upon our prior work (in particular, Choi *et al.* (2007a) and Polisson, Quah, and Renou (2020)). In the experiment, subjects choose an allocation of contingent commodities from a *three-dimensional* budget set through a simple "point-and-click" design. As our power analysis shows, data from three-dimensional budget sets provide a much stronger test—especially of EUT versus non-EUT alternatives—than data from two-dimensional budget lines (as collected by Choi *et al.* (2007a), Choi *et al.* (2014), and Halevy, Persitz, and Zrill (2018), among others).

Afriat's (1967) theorem tells us that if a finite dataset generated by an individual's choices from linear budget sets satisfies the Generalized Axiom of Revealed Preference (GARP), then the data can be rationalized by a well-behaved (by which we mean a continuous and increasing) utility function. This result provides a practical way of checking whether a dataset is *rationalizable* in this minimal/basic sense. There are also extensions of Afriat's 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 utility function that is monotone with respect to FOSD, and whether a dataset is *EUT-rationalizable*, in the sense that it is consistent with the maximization of an expected utility function.

[1] Monotonicity with respect to first-order stochastic dominance is a natural and widely accepted principle in decision theory, so much so that theories of choice under risk have been modified to avoid violations of stochastic dominance, as pointed out by Quiggin (1990), Wakker (1993), and Starmer (2000); for example, cumulative prospect theory (Tversky and Kahneman, 1992) "dominance corrects" the original formulation of prospect theory (Kahneman and Tversky, 1979).

[2] In the choice acclimating personal equilibrium model of Kőszegi and Rabin (2007), monotonicity with respect to FOSD holds if the coefficient of loss aversion is within a certain range (see Masatlioglu and Raymond (2016)).

For datasets that do not satisfy GARP exactly, Afriat (1973) introduces the notion of the Critical Cost Efficiency Index (CCEI), which measures the extent to which budget sets need to be reduced in order to rationalize the data. The CCEI, denoted by e^*, is bounded between 0 and 1; the closer it is to 1, the smaller are the budgetary adjustments required for rationalizability. There are also known procedures to measure the extent to which budget sets need to be adjusted 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).

While other measures of violations of rationalizability are available, we adopt the CCEI since it is straightforward to calculate and interpret (and, partly for those reasons) the most commonly used measure in empirical work. 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 (such as rank-dependent utility) 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 rationalizability (e^*), FOSD-rationalizability (e^{**}), and EUT-rationalizability (e^{***}) *for each subject*, using a purely nonparametric 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 a small fraction of our subjects are perfectly rationalizable (have no violations of GARP), but none are perfectly FOSD-rationalizable and thus EUT-rationalizable. More importantly, the difference between *perfect* rationalizability and FOSD-rationalizability ($1 - e^{**}$) is much larger at all score values than the difference between FOSD-rationalizability and EUT-rationalizability ($e^{**} - e^{***}$). This difference in differences is statistically significant for nearly all subjects. Violations of EUT thus run deeper than

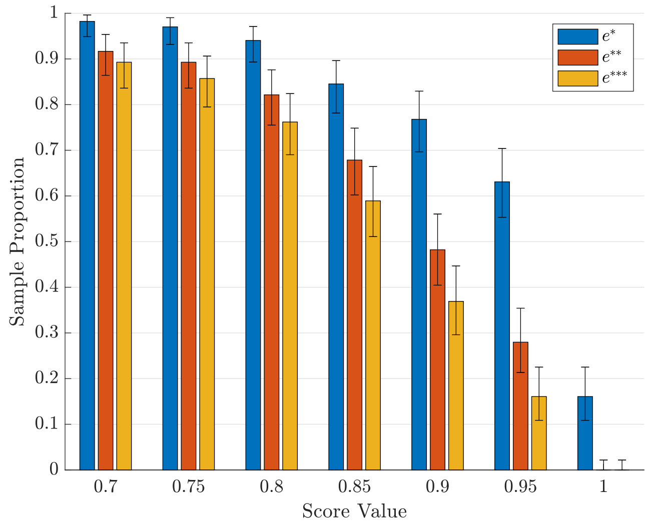


Figure 1: Distributions of Rationalizability Scores

To interpret the bars, consider the score value 0.9. The proportion of subjects in the sample with $e^* \geqslant 0.9$ is 76.8 percent, the proportion with $e^{**} \geqslant 0.9$ is 48.2 percent, and the proportion with $e^{***} \geqslant 0.9$ is 36.9 percent. The braces represent exact 95 percent confidence intervals on the proportions.

violations of independence, challenging the most prominent non-EUT alternatives.

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 minor after accounting for violations of ordering and monotonicity—is what Quiggin (1982) calls an "undesirable result" as ordering and monotonicity are more fundamental principles than the standard independence axiom, and they 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.

Our rich individual-level experimental data involving three states and three associated securities could also be used, in principle, to test each non-EUT theory against the others. The different (weaker) alternatives deliver more empirically testable restrictions on observed behavior in the case of three states than in the case of two states. However, for most subjects there is only a small (or no) difference between FOSD-rationalizability (e^{**}) and EUT-rationalizability (e^{***}), which implies that there is little scope for existing non-EUT alternatives to explain observed behavior.

Looking ahead, we note that an important advantage of our methods and analyses is that they can be transported, with relative ease, to different decision domains. The experiment reported in this paper considers decision making under risk. In related ongoing work, we study decision making under uncertainty/ambiguity, and also intertemporal choice.

The rest of the paper is organized as follows. The next section provides more background and motivation. Section 3 describes our tests of rationalizability, experimental procedures, and the power of the experiment. Section 4 summarizes the experimental results. Section 5 describes how the paper is related to the literature, focusing on recent revealed preference papers on choice under risk. Section 6 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.

2 Background and Motivation

Much of the experimental literature on choice under risk is directed towards finding violations of EUT. To understand the role of each of the axioms on which EUT is based, suppose that there are three mutually non-indifferent outcomes $x_h \succ x_m \succ x_l$ and consider the probability triangle depicted in Figure 2. 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]

Monotonicity with respect to 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). Ordering (completeness and transitivity) plus continuity imply that there exists a map of (non-intersecting) indifference curves. Assuming that these axioms hold, independence 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 (steeper lines imply higher risk aversion).

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

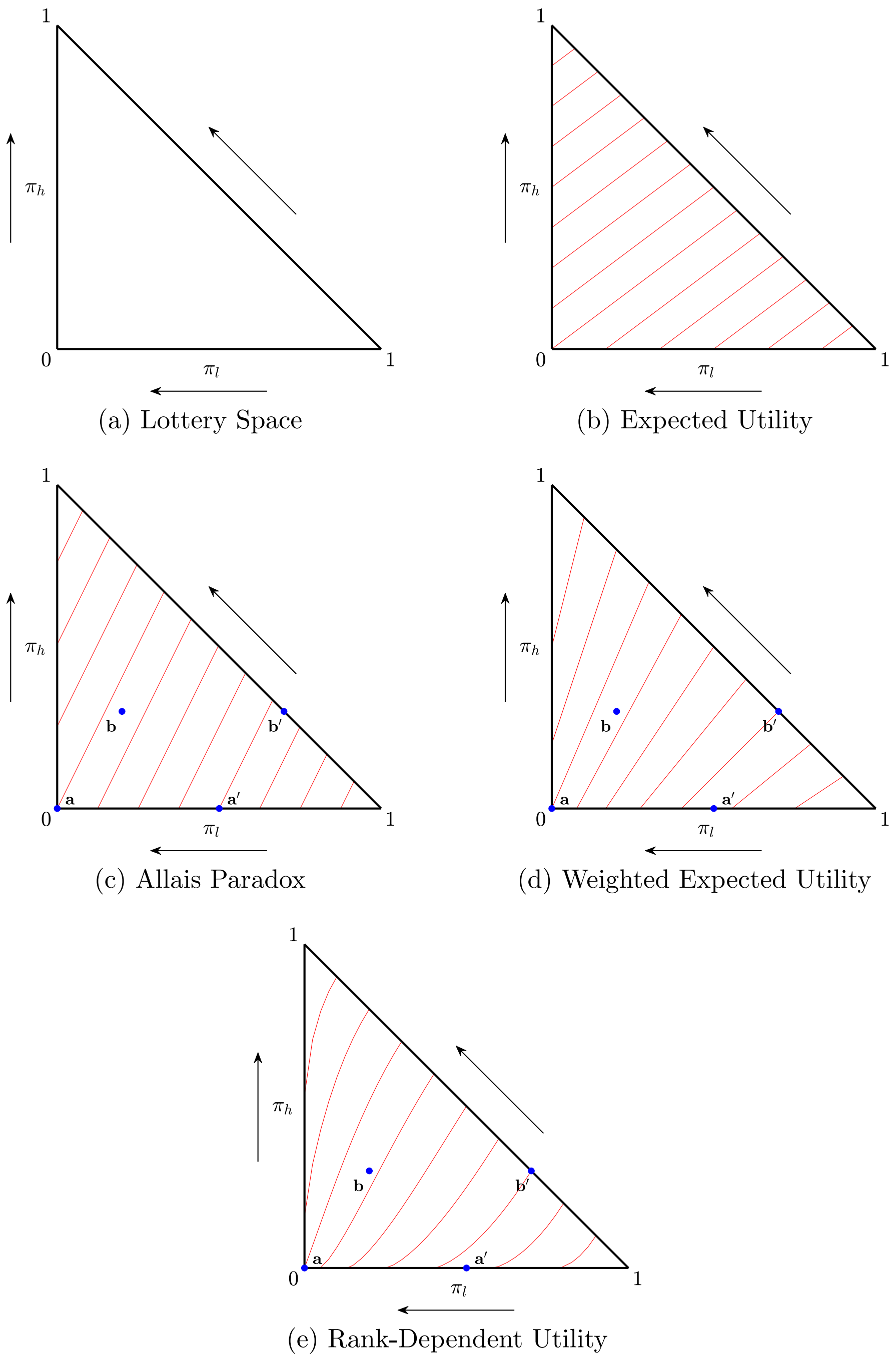


Figure 2: Probability Triangles

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.

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 curves 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]

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).[5] Or in rank-dependent utility (Quiggin, 1982, 1993) and prospect theory (Kahneman and Tversky, 1979; Tversky and Kahneman, 1992) 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 most experimental studies, the criterion used to evaluate a 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 very specific choices that are narrowly tailored to discover violations of independence and its various generalizations. There is less emphasis on ensuring that these decision problems are representative, both in the statistical sense and in the economic sense. As a result, the accumulated experimental evidence against independence that has prompted theorists to develop formal alternatives to EUT consists primarily of Allais-type behaviors—choices inconsistent with linear indifference curves in the probability triangle. Such an approach is unsurprising, given the focus on

[4]Interestingly, while violations of the independence axiom appear to be widespread, in a recent survey on the experimental robustness of the Allais paradox across 83 experiments and 30 studies, Blavatskyy, Ortmann, and Panchenko (2021) concludes that the Allais paradox is a somewhat fragile empirical finding. This survey's conclusion is compatible with our main message.

[5]The indifference curves corresponding to disappointment aversion (Gul, 1991) 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). See Gul (1991), Figure 2 (p. 679).

the independence axiom and that, apart from a few notable exceptions,[6] 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 performance (or underperformance) of EUT and therefore that all of the assumptions which underpin the model deserve closer scrutiny.

In this paper, we develop tests of rationalizability that are *comprehensive*, in the sense that we check 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 the performances of progressively restrictive models using a common measure of model performance, we can compare the relative impact of the different axioms which make up EUT. 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. The overall objective of our experiment and analysis is to provide a positive account of choice under risk in natural economic environments.

3 Framework for Analysis

In this section, we describe the theory on which the experimental design is based, the design itself, and the power of the experiment. All technical details that are not essential for the experimental results are relegated to Appendix I.

3.1 Rationalizability

We consider a portfolio choice framework with S states of nature, each state denoted by $s = 1, \ldots, S$. For each state s, there is an Arrow (1964) security that pays one in state s and zero in the other state(s). Let $x_s \geqslant 0$ denote the demand for the security that pays off in state s and $p_s > 0$ denote the corresponding price, so that $\mathbf{x} = (x_1, \ldots, x_S)$ is a demand allocation and $\mathbf{p} = (p_1, \ldots, p_S)$ is a price vector. Let $\mathcal{D} := (\mathbf{p}^i, \mathbf{x}^i)$ be the data generated by a subject's choices from linear budget sets, where $\mathbf{p}^i$ denotes the i-th observation of the price vector and $\mathbf{x}^i$ denotes the associated allocation. We say that a data set $\mathcal{D}$ is *rationalizable*

[6] For generalizations of EUT that allow for nontransitivity, see, for example, Bell (1982), Fishburn (1982), and Loomes and Sugden (1982).

if there is a utility function $U : \mathbb{R}^S_+ \to \mathbb{R}$ such that $U(\mathbf{x}^i) \geqslant U(\mathbf{x})$ for all

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

In other words, the utility of $\mathbf{x}^i$ is weakly higher than that of any alternative that is weakly cheaper at the price vector $\mathbf{p}^i$.

Note that rationalizability, as defined, has no empirical content, since any dataset $\mathcal{D}$ can be rationalized by a constant utility function. For this concept to be meaningful, some restriction has to be imposed on U. A well-known result, due to Afriat (1967), tells us that $\mathcal{D}$ can be rationalized by a *well-behaved* (in the sense of being continuous and increasing) utility function if and only if the data satisfy the Generalized Axiom of Revealed Preference (GARP). GARP is an intuitive and (more importantly from the perspective of empirical application) easy-to-check condition on $\mathcal{D}$.

To account for data that are not exactly rationalizable, Afriat (1972, 1973) proposes the notion of the Critical Cost Efficiency Index (CCEI). Given a number $e \in (0, 1]$, a dataset $\mathcal{D}$ is said to be *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

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

Clearly, approximate rationalizability weakens the notion of rationalizability since $\mathcal{B}^i(e)$ is a subset of $\mathcal{B}^i$. As Afriat (1973) notes, this definition captures the idea that while the consumer "has a definite structure of wants," she "programs at a level of cost-efficiency e." The approach is otherwise agnostic about the deeper nature of the "errors" which may arise in individual choices.

It is not difficult to see that *every* dataset $\mathcal{D}$ could be rationalized by a well-behaved utility function at an efficiency level e for some $e \in (0, 1]$ that is sufficiently close to zero. The CCEI, denoted by e^*, of a dataset $\mathcal{D}$ is *the greatest* e *for which* $\mathcal{D}$ *is rationalizable.* For example, if $e^* = 0.95$, then we can find U such that $U(\mathbf{x}^i)$ is greater than $U(\mathbf{x})$ for any bundle $\mathbf{x}$ that is more than 5 percent cheaper than $\mathbf{x}^i$ at the prevailing prices $\mathbf{p}^i$. Alternatively, the decision maker is effectively "wasting" as much as 5 percent of his income by making "irrational" choices. Just as GARP characterizes rationalizabilty by a well-behaved utility function, so too is there a modified version of GARP that can be used to check whether a dataset is rationalizable by a well-behaved utility function at some efficiency level e. It follows that one could easily obtain e^*.

Afriat's Theorem is just the first of a long list of results developed by various authors with the following pattern: $\mathcal{D}$ is rationalizable by a well-behaved utility function 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 well-behaved utility functions that are monotone with respect to FOSD. In our framework, the probability of state s is commonly known to be $\pi_s > 0$, so that $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_S)$ is a vector of probability weights with $\pi_1 + \cdots + \pi_S = 1$. Then we say that U is monotone with respect to FOSD if $U(\mathbf{x}'') \geqslant U(\mathbf{x}')$ whenever $\mathbf{x}''$ (considered as a distribution through $\boldsymbol{\pi}$) first-order stochastically dominates $\mathbf{x}'$ (with the inequality being strict if the dominance is strict).[7] It is straightforward to check that, in the case where the states are equiprobable (as in our experiment), a well-behaved utility function is monotone with respect to FOSD if and only if it is symmetric. A dataset $\mathcal{D}$ is said to be *FOSD-rationalizable* (with respect to a given $\boldsymbol{\pi}$) if it can be rationalized by a utility function that is well-behaved and monotone with respect to FOSD. Relying on Nishimura, Ok, and Quah (2017), we provide an easy-to-implement (necessary and sufficient) test of whether $\mathcal{D}$ is FOSD-rationalizable; furthermore, one could also check whether $\mathcal{D}$ can be rationalized at cost efficiency e by a utility function in this family and thus the corresponding CCEI, denoted by e^{**}, can easily be calculated. Since this family of utility functions is contained within the family of well-behaved utility functions, it must be the case that $e^{**} \leqslant e^*$.

The second important family is the family of well-behaved utility functions that satisfy expected utility. These are utility functions U taking the form

$$U(\mathbf{x}) = \pi_1 u(x_1) + \cdots + \pi_S u(x_S),$$

[7] A utility function U that is monotone with respect to FOSD is 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 just two equiprobable states. Then $U(1,3) > U(2,1)$ if U is monotone with respect to FOSD because $(1,3)$ first-order stochastically dominates $(2,1)$, but no relationship between $U(1,3)$ and $U(2,1)$ is implied by U being increasing.

where the Bernoulli index $u : \mathbb{R}_+ \to \mathbb{R}$ is continuous and increasing. Recently, Polisson, Quah, and Renou (2020) have developed a procedure called the Generalized Restriction of Infinite Domains (or GRID) method that could be employed to test whether a dataset is rationalizable (at cost efficiency e) by a well-behaved expected utility function, or *EUT-rationalizable*. 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 well-behaved utility functions which respect FOSD, it must be the case that $e^{***} \leqslant e^{**}$.

To recap, given any dataset $\mathcal{D}$ we could 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, specific families of utility functions (such as rank-dependent utility) which generalize expected utility and respect FOSD 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 departure from EUT. A subject who is perfectly EUT-rationalizable will have $1 = e^* = e^{**} = e^{***}$. More generally, e^{***} will be strictly less than one, and the corresponding values of e^* and e^{**} will then allow us to say something about why that has occurred. For example, if $1 = e^* = e^{**} > e^{***}$, then it would be plausible to believe that the subject is indeed violating the independence axiom and her behavior could potentially be explained by a utility model that relaxes the independence axiom, while retaining monotonicity with respect to FOSD. On the other hand, a subject for whom $1 = e^* > e^{**} = e^{***}$ could be utility-maximizing, but her choices could only be explained by a model that departs from monotonicity with respect to FOSD. Last but not least, the choice behavior of a subject with $1 > e^*$ is not consistent with the maximization *any* utility function; she may or may not also be violating the independence axiom, but understanding her behavior would require a more radical departure from the classical framework.

In Appendix I, we provide more details on GARP and the other conditions for checking rationalizability (or rationalizability at a given cost efficiency) with respect to specific families of utility functions.

3.2 Experiment

In this paper, we employ the same experimental methodology as in Choi *et al.* (2007a, 2014) and Halevy, Persitz, and Zrill (2018), except that instead of having just two states of nature ($S = 2$) and two associated Arrow securities, the new experiment incorporates three states ($S = 3$) and three associated Arrow securities, with a price for each security. Choices from three-dimensional budget sets provide more rigorous tests of rationalizability than choices from two-dimensional budget sets, in particular when it comes to testing EUT (see more on this below in our discussion of the power of the experiment).

We conducted the experiment at UC Berkeley and UCLA. The subjects in the experiment were recruited from undergraduate classes at these institutions. In the experiment, subjects choose an allocation from a three-dimensional budget set presented using the graphical interface introduced by Choi *et al.* (2007b). Subjects make choices by using the computer mouse to move the pointer on the computer screen to the desired point, and are restricted to allocations on the budget constraint. The full experimental instructions, including the computer program dialog windows, are reproduced in Appendix II.[8]

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 risky assets, except that each choice involved choosing a point on a three-dimensional (instead of two-dimensional) graph representing the set of possible allocations. In the experimental

[8]We are building on the expertise that we have acquired in previous work using the experimental method across different types of individual choice problems. Choi *et al.* (2014) introduces the graphical interface of Choi *et al.* (2007b) into 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 (2021), and Echenique, Imai, and Saito (2021). Fisman, Kariv, and Markovits (2007), Fisman *et al.* (2015), Fisman, Jakiela, and Kariv (2015, 2017), and Li, Dow, and Kariv (2017) employ a similar experimental methodology to study social preferences across a number of different samples, including a nationally representative sample. Three-dimensional budget sets have been used by Fisman, Kariv, and Markovits (2007) to study preferences for giving, and also by Ahn *et al.* (2014) to study ambiguity aversion, but so far have not been used to study risk. Other related work by Zame *et al.* (2020) develops theoretical tools and experimental methods for testing the linkages between preferences for personal and social consumption and attitudes toward risk and inequality.

task, there are three *equally likely* states denoted by $s = 1, 2, 3$ and three associated securities, each of which promises a payoff of one token (the experimental currency) in one state and nothing in the others. Recall that $x_s \geqslant 0$ denotes the demand for the security that pays off in state s and $p_s > 0$ denotes the corresponding price. Without loss of generality, we assume that the budget is normalized to 1. The budget set is then given by $\mathcal{B} = \{\mathbf{x} : \mathbf{p} \cdot \mathbf{x} = 1\}$, where $\mathbf{x} = (x_1, x_2, x_3)$ denotes the portfolio of securities and $\mathbf{p} = (p_1, p_2, p_3)$ denotes the vector of security prices.

Each experimental subject faced 50 independent decision rounds. For each subject, the computer selected 50 budget sets randomly from the set of planes that intersect at least one axis at or above the 50 token level and intersect all axes at or below the 100 token level. The budget sets selected for each subject in his/her decision problems were independent of one another and of the budget sets selected for other subjects in their decision problems. 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.

3.3 Power

To show that the three-dimensional budgetary experiment is more powerful than the two-dimensional experiments previously used in the literature—and specifically that it is sufficiently powerful to detect whether or not EUT is the right model of choice under risk—we start by building on the test designed by Bronars (1987) which employs as a benchmark the choices of a simulated subject who randomizes uniformly among all allocations on each budget set. The simulated subject makes 50 choices from randomly generated budget sets, in the same way as do the human subjects.

To focus on EUT-rationalizability, each choice is drawn independently from the uniform distribution over all allocations on the budget set, subject to keeping the data *perfectly* compatible with FOSD-rationalizability, that is $e^{**} = 1$. Figure 3 provides a clear graphical illustration by comparing the distributions of e^{***} generated by such simulated subjects in

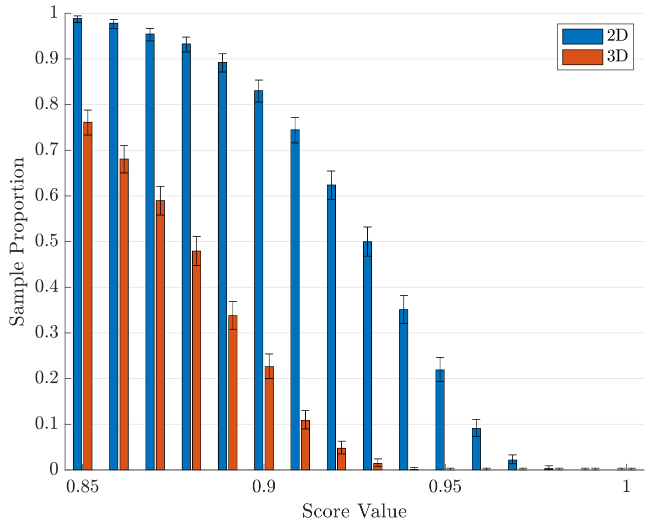


Figure 3: Power of EUT-Rationalizability

The three-dimensional (3D) budgetary experiment is more powerful than the two-dimensional (2D) experiment in detecting violations of EUT. We compare the distributions of EUT-rationalizability scores (e^{***}) in 2D and 3D for simulated subjects who choose randomly conditional on perfect FOSD-rationalizability ($e^{**} = 1$). The proportion of simulated subjects that have e^{***} above 0.9 (conditional on $e^{**} = 1$) is over 80 percent in the 2D experiment but just over 20 percent in the 3D experiment.

the two- and three-dimensional budgetary experiments. The horizontal axis shows the value of e^{***} and the vertical axis measures the fraction of simulated subjects corresponding to each level. If we choose $e^{***} = 0.9$ as our critical value, we find that more than 80 percent of simulated subjects have e^{***} above 0.9 in the two-dimensional experiment, while just over 20 percent have e^{***} above 0.9 in the three-dimensional experiment.

Another benchmark against which to compare the power of the two- and three-dimensional designs involves the choices of a simulated subject who maximizes a non-EUT utility function. To illustrate such preferences when there are three states ($S = 3$), consider the rank-dependent utility function:

$$U(\tilde{\mathbf{x}}) = \beta_L u(x_L) + \beta_M u(x_M) + \beta_H u(x_H),$$

where $\beta_L, \beta_M, \beta_H > 0$ are decision weights that sum to unity, $\tilde{\mathbf{x}} = (x_L, x_M, x_H)$ is a *rank-ordered* portfolio with payoffs $x_L \leqslant x_M \leqslant x_H$, and u is the Bernoulli index. This formulation encompasses a number of non-EUT models and reduces to EUT when $\beta_L = \beta_M = \beta_H$ (since each state has an equal likelihood of occurring).[9] When there are two states of nature

[9]As Starmer (2000) points out, although the number of so-called non-EUT models "is well into double

($S = 2$), the rank-dependent utility function takes the simpler form

$$U(\tilde{\mathbf{x}}) = \beta_L u(x_L) + \beta_H u(x_H),$$

where β_L, β_H are the decision weights and $\tilde{\mathbf{x}} = (x_L, x_H)$ is the rank-ordered portfolio with payoffs $x_L \leqslant x_H$. The rank-dependent formula for the rank-ordered portfolio $\tilde{\mathbf{x}}$ can be expressed in terms of the *probability weighting function* w (see more on this below) as follows:

$$\begin{aligned}
\beta_L &= 1 - w\left(\tfrac{2}{3}\right), \\
\beta_M &= w\left(\tfrac{2}{3}\right) - w\left(\tfrac{1}{3}\right), \\
\beta_H &= w\left(\tfrac{1}{3}\right),
\end{aligned}$$

for three states ($S = 3$), and

$$\begin{aligned}
\beta_L &= 1 - w\left(\tfrac{1}{2}\right), \\
\beta_H &= w\left(\tfrac{1}{2}\right),
\end{aligned}$$

for two states ($S = 2$). That is, the cumulative distribution function of the induced lottery assigns to each monetary payoff the probability of receiving that payoff or anything less.[10]

In order to draw a comparison across the two- and three-dimensional experiments using simulated subjects maximizing a rank-dependent utility function, we hold the weighting fixed using the weighting function suggested by Tversky and Kahneman (1992), which distorts each probability $\pi \in (0, 1)$ according to

$$w(\pi) = \frac{\pi^\gamma}{[\pi^\gamma + (1-\pi)^\gamma]^{1/\gamma}}.$$

figures," the preferences generated by rank-dependent utility Quiggin (1982, 1993) is the leading contender. Machina (1994) concludes that rank-dependent utility is "the most natural and useful modification of the classical expected utility formula," and Starmer (2000) argues that "if one is looking to organize the data from the large number of triangle experiments, then the decision-weighting models are probably the best bet." Yaari (1987), Segal (1990), Wakker (1994), and Abdellaoui (2002), among others, provide axiomatizations of rank-dependent utility, and Diecidue and Wakker (2001) discusses its underlying intuition.

[10]The weighting function w, which is increasing and satisfies $w(0) = 0$ and $w(1) = 1$, transforms the distribution function into decision weights. By definition, the decision weight β_H is equal to $w\left(\frac{1}{3}\right)$ in the case of three states and to $w\left(\frac{1}{2}\right)$ in the case of two states.

This formulation takes the familiar (inverted) s-shaped form for $0 < \gamma < 1$, and any $\gamma > 0.279$ guarantees that w is increasing.[11] When $\gamma = 1$ we have $w(\pi) = \pi$, and so we get the standard EUT representation. In our numerical simulation, we set $\gamma = 0.5$ (in order to generate sufficient "pessimism") and we specify $u(x) = \log(x)$. Clearly, for these simulated subjects $1 = e^* = e^{**}$ since their choices are FOSD-rationalizable by construction. However, as a simple indication, while *all* of the simulated subjects have e^{***} above 0.95 in the two-dimensional experiment, *none* have e^{***} above 0.95 in the three-dimensional experiment.

Despite the advantages of the three-dimensional design, we nevertheless complement our analysis of these data by analyzing observations collected from a further 956 subjects, each making 50 choices over two-dimensional budget lines. (These experiments are identical to the (symmetric) risk experiment of Choi *et al.* (2007a).) We discuss these results in Section 4.3; the bottom line is that the major findings in the three-dimensional experiment are replicated across the two-dimensional experiments.

4 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 associated allocation. The experiment provides a large set of data consisting of many individual decisions over a wide range of three-dimensional budget sets. This is an important point, because as our power analysis shows, a large number of individual decisions over three-dimensional instead of two-dimensional budget sets is crucial in order to provide a sufficiently powerful test of the entire set of axioms underlying EUT.

4.1 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 indices of rationalizability (e^*), FOSD-rationalizability (e^{**}), and EUT-rationalizability (e^{***}). But the

[11] The other widely-used (single parameter) probability weighting function was proposed by Prelec (1998).

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.

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 interesting 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 $\bar{x}_1^i \geqslant \bar{x}_2^i \geqslant \bar{x}_3^i$

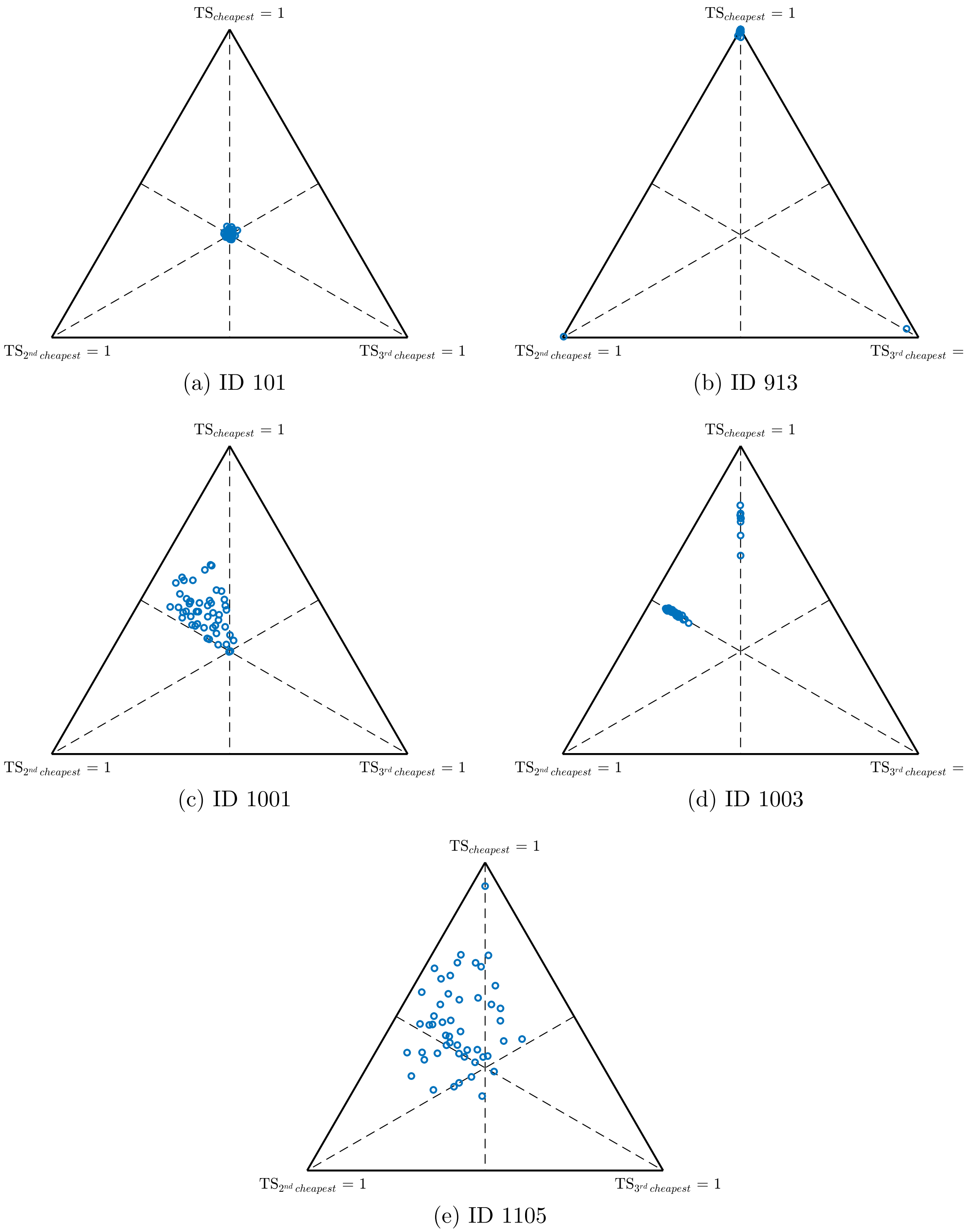


Figure 4: Subject Behavior

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.

where $p_1^i < p_2^i < p_3^i$) but not EUT-rationalizable. As we explain in Appendix I, preferences generated by 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 subset of the simplex where $\bar{x}_1^i \geqslant \bar{x}_2^i \geqslant \bar{x}_3^i$; this behavior is not FOSD-rationalizable (and thus also not EUT-rationalizable). We have obviously shown just a small subset of our full set of subjects, and these are of course special cases where regularities in the data are very clear.[12]

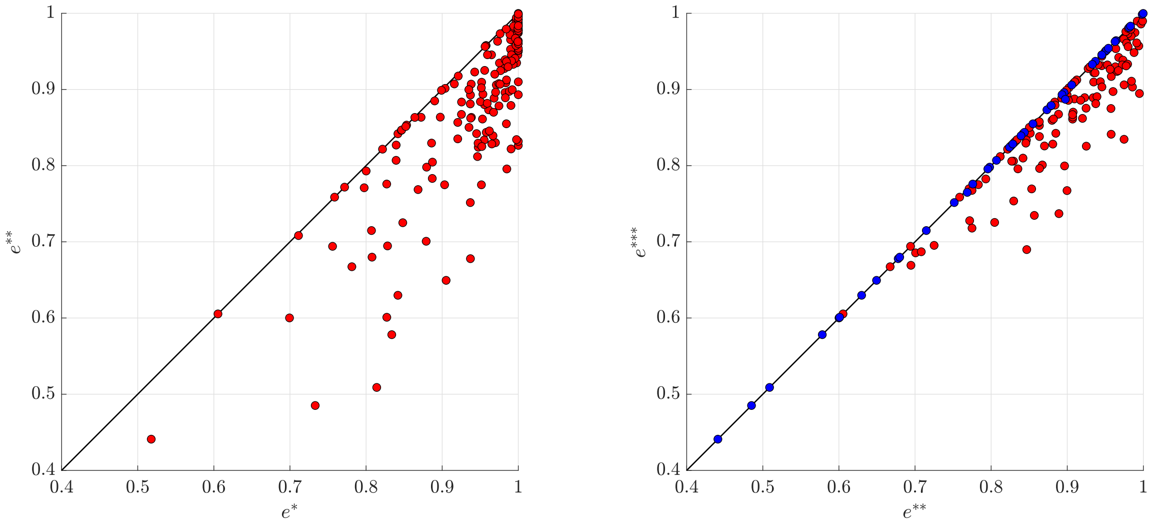
4.2 Rationalizability Scores

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 lines. 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 ($e^* = e^{**} = e^{***}$).

Our rich individual-level data also allow us to make statistical comparisons of rational-

[12] There are many subjects for whom the behavioral regularities are much less clear. However, a review of the full raw dataset reveals both regularities within subjects and heterogeneity across subjects. The scatterplots for the full set of subjects are available upon request.



(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 lines. (a) All individual-level differences between e^* and e^{**} are statistically significant at the 1 percent significance level (red). (b) The individual-level differences between e^{**} and e^{***} are statistically significant for 75.0 percent of the sample (red), but there is also a sizeable minority of subjects for whom this is not the case (blue).

izability (e^*) versus FOSD-rationalizability (e^{**}) and of FOSD-rationalizability (e^{**}) versus EUT-rationalizability (e^{***}) using a purely nonparametric econometric approach. To this end, for each subject, we split the 50 observations into two *non-overlapping* partitions of 25 observations, generating paired subsamples of observations. Clearly, we cannot examine all $\binom{50}{25} > 10^{14}$ possible paired subsamples of the observed individual-level data; instead we draw 1,000 such paired subsamples at random for each subject and construct the sampling distributions of e^* and e^{***} on one subsample and the sampling distribution of e^{**} on the other. Note that given the non-overlapping partitions, the orderings $e^* \geqslant e^{**}$ and $e^{**} \geqslant e^{***}$ are no longer guaranteed. We can then straightforwardly test whether the mean difference between the pairs of e^* and e^{**} and of e^{**} and e^{***} are zero (or not) using a paired t-test.

In Figure 5, individual subjects are depicted in red if the two scores—either e^* and e^{**} (Figure 5a) or e^{**} and e^{***} (Figure 5b)—are statistically distinguishable at the 1 percent significance level and depicted in blue otherwise. All individual-level differences between e^* and e^{**} (Figure 5a) are statistically significant, including for those 11 subjects (6.5 percent) falling along the 45-degree line (for whom $e^* = e^{**}$ across all 50 observations). The

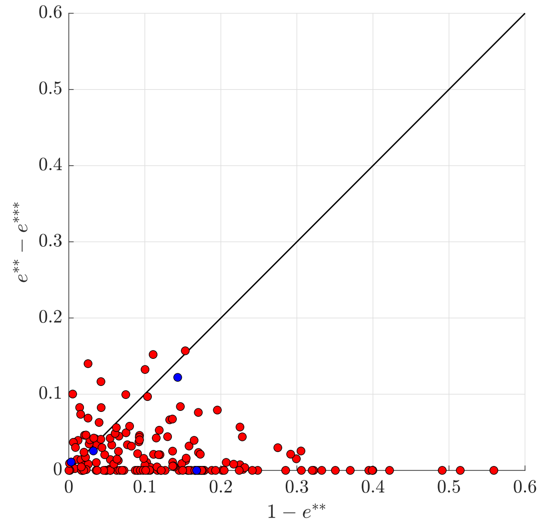


Figure 6: Scatterplot of Score Differences

The plot depicts rationalizability score differences for individual subjects. For the vast majority of subjects, the difference between FOSD-rationalizability and EUT-rationalizability ($e^{**} - e^{***}$) is small (or non-existent), while the difference between perfect rationalizability and FOSD rationalizability ($1 - e^{**}$) is much larger: 85.1 percent of subjects fall below the 45-degree line, and of those 45.5 percent fall along the horizontal axis ($e^{**} = e^{***}$). This difference in differences is statistically significant for 97.6 percent of subjects (red) at both the 1 and 5 percent significance levels.

individual-level differences between e^{**} and e^{***} (Figure 5b) are statistically significant for 126 subjects (75.0 percent), including for 25 of the 65 subjects (38.5 percent) falling along the 45-degree line (for whom $e^{**} = e^{***}$ across all 50 observations). If instead we evaluate at the 5 percent significance level, the individual-level differences between e^{**} and e^{***} are statistically significant for 134 subjects (79.8 percent). Hence, for the majority of subjects the difference between FOSD-rationalizability and EUT-rationalizability ($e^{**} - e^{***}$) is statistically significant, but there is also a sizeable minority for whom this is not the case.

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 ($1 - e^{**} > e^{**} - e^{***}$), and of those 65 subjects (45.5 percent) fall along the horizontal axis ($e^{**} = e^{***}$). 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, such as weighted expected utility, rank-dependent utility, or reference-dependent risk preferences, that relax the independence axiom to explain observed behavior, as they all postulate FOSD-rationalizability ($1 = e^* = e^{**} > e^{***}$).[13]

To provide a statistical test of the difference between $1 - e^{**}$ and $e^{**} - e^{***}$, we again draw 1,000 paired subsamples of observations for each subject and construct the sampling distribution of $1 - e^{**}$ on one subsample and the sampling distribution of $e^{**} - e^{***}$ on the other. We then test whether the mean difference in differences is statistically significant using a paired t-test. We find that it is significant for 164 subjects (97.6 percent) at both the 1 and 5 percent significance levels. These subjects are depicted in red in Figure 6; the other subjects are depicted in blue.

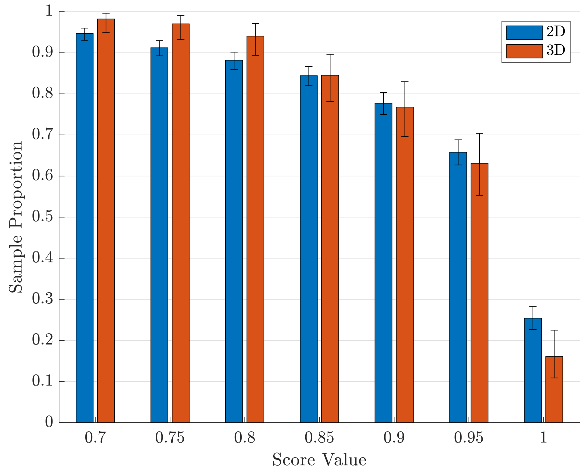
The broad conclusion from our analysis is clear: even for a single subject, the sources of violation of EUT are variegated; furthermore, for many subjects, violations of ordering and monotonicity are more prominent and much larger in magnitude than departures from the independence axiom.

4.3 Two- Versus Three-Dimensional Data

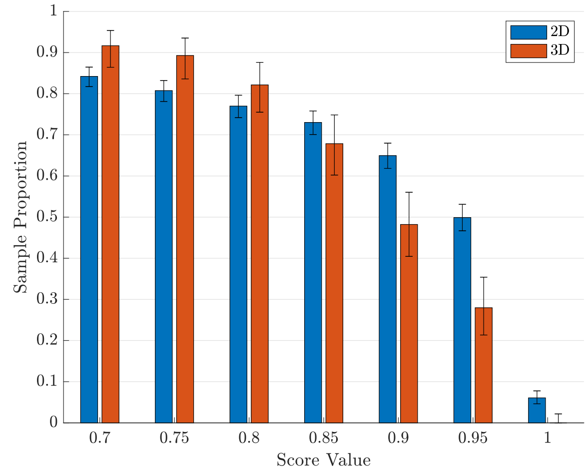
For comparison purposes, in Appendix III we replicate our entire analysis with observations on 956 subjects making choices from two-dimensional budget lines. For each subject, we again have a set of 50 observations $\mathcal{D} := (\mathbf{p}^i, \mathbf{x}^i)_{i=1}^{50}$ where $\mathbf{p}^i = (p_1^i, p_2^i)$ denotes the i-th observation of the price vector and $\mathbf{x}^i = (x_1^i, x_2^i)$ denotes the associated allocation.[14] Figure 7 compares the rationalizability scores across the two- and three-dimensional experiments

[13] Utility functions representing reference-dependent risk preferences (specifically the choice acclimating personal equilibrium model of Kőszegi and Rabin (2007)) can fail to be increasing if loss aversion is sufficiently high (see Masatlioglu and Raymond (2016)); however, these preferences are always locally locally nonsatiated and, in our experimental setting, symmetric. For reasons explained in greater detail in Appendix I, 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 rationalizability score for such preferences cannot improve on e^{**}.

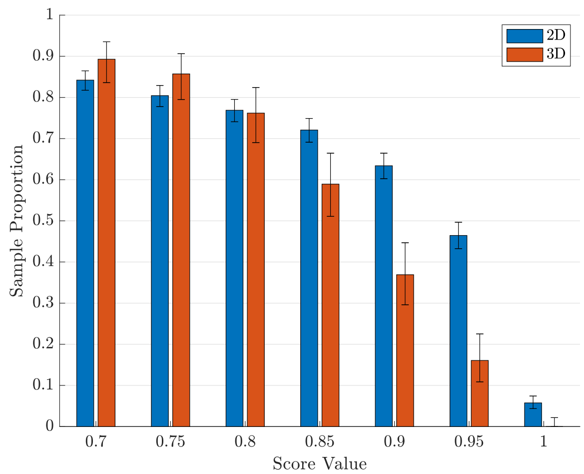
[14] The data include the (symmetric) data collected by Choi *et al.* (2007a) and similar data with different subject pools collected by Zame *et al.* (2020) and Cappelen *et al.* (2021) as well as new data. In all of these experiments, the individual-level data consist of 50 decision problems. We do not include the data of Choi *et al.* (2014) which consist of 25, rather than 50, decision problems. Note that 25 individual decisions provide a rich enough data set to provide a powerful test of GARP. But as our power analysis shows, choices from two-dimensional budget lines provide a much weaker test of EUT, so we omit datasets with only 25 individual decisions, though this number is still higher than is usual in the literature. See, for examples, Cox (1997), Sippel (1997), Mattei (2000), Harbaugh, Krause, and Berry (2001), and Andreoni and Miller (2002), among others.



(a) Distributions of e^*



(b) Distributions of e^{**}



(c) Distributions of e^{***}

Figure 7: Distributions of Rationalizability Scores

The plots depict distributions of rationalizability scores across the two-dimensional (2D) and three-dimensional (3D) experiments for (a) e^*, (b) e^{**}, and (c) e^{***}.

for e^* (Figure 7a), e^{**} (Figure 7b), and e^{***} (Figure 7c). Note that the data from three-dimensional budget sets are at least as rationalizable (e^*) as the data from two-dimensional budget lines, which is an interesting result in its own right. As a practical note, it suggests that subjects did not have any difficulties in understanding the procedures or using the three-dimensional computer program.

On the other hand, the data from three-dimensional budget sets are distinctly less FOSD-rationalizable (e^{**}) and EUT-rationalizable (e^{***}) than the data from the two-dimensional budget lines. In the three-dimensional experiment, 28.0 (resp. 16.1) percent of the subjects have e^{**} (resp. e^{***}) scores above the 0.95 threshold, and 48.2 (resp. 36.9) percent have values above 0.9. In the two-dimensional experiment (also with 50 choices), the corresponding percentages are 49.9 (resp. 46.4) and 65.0 (resp. 63.4). Finally, statistical tests on the two-dimensional data show that the individual-level differences between e^* and e^{**} are statistically significant for 859 (89.9 percent) and 866 (90.6 percent) at the 1 and 5 significant levels, respectively. In contrast, the individual-level differences between e^{**} and e^{***} are statistically significant for only 215 (22.5 percent) and 268 subjects (28.0 percent). This comparison suggests that three-dimensional budget sets (relative to two-dimensional budget sets) considerably improve the power of revealed preference tests of EUT-rationalizability.

In the two-dimensional data, as in the three-dimensional data, the loss of consistency arising from EUT specifically is small, once we account for ordering and monotonicity. Indeed, $1 - e^{**} > e^{**} - e^{***}$ for 827 out of 956 subjects (86.5 percent). These differences in differences are statistically significant for 888 subjects (92.9 percent) and 890 subjects (93.1 percent) at the 1 and 5 percent significance levels, respectively.

5 Related Literature

There is a vast amount of research on decision making under risk and under uncertainty, and laboratory experiments have provided some key empirical guideposts for the development of new ideas in these areas. We will not attempt to review the large and growing experimental literature. Though now somewhat dated, an overview of experimental and theoretical work can be found in Camerer (1995), while Starmer (2000) provides a review of the risk literature that focuses on evaluating non-EUT theories.[15] Following the seminal work of Hey and Orme (1994) and Harless and Camerer (1994)), a number of papers have estimated parametric utility functions. While Harless and Camerer (1994) fits models to aggregate data, Hey and Orme (1994) uses data derived from decisions over a very large menu of binary choices and estimates functional forms at the level of the individual subject.

More recently, Choi *et al.* (2007a) employs graphical representations of budget sets containing bundles of state-contingent commodities in order to elicit preferences; this exper-

[15] Camerer and Weber (1992) and Harless and Camerer (1994) also summarize the experimental evidence from testing the various utility theories of choice under risk and under uncertainty. Kahneman and Tversky (2000) collects many theoretical and empirical papers that have emerged from their pioneering work on prospect theory.

imental approach constitutes the foundation of this paper's contribution as it allows for the collection of a very rich individual-level dataset. For each subject in their experiment, Choi *et al.* (2007a) tests the data for consistency with GARP and estimates preferences in a parametric model with loss or disappointment aversion (Gul, 1991). This formulation encompasses a number of different theories and embeds EUT as a parsimonious and tractable special case. 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. Indeed, Halevy, Persitz, and Zrill (2018) highlights the distinction between the non-parametric and parametric recoverability of preferences.

The most basic question that one could ask about individual-level choice data is whether they are compatible with utility maximization, and classical revealed preference theory (Samuelson, 1938, 1948, 1950; Houthakker, 1950; Afriat, 1967; Diewert, 1973; Varian, 1982) provides GARP as a direct test.[16] Consistency with GARP is implied by—and guarantees—choice from a coherent preference over all possible alternatives, but *any* consistent preference ordering that is locally nonsatiated is 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 consistency and to ask whether the choices made by a subject are 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. To answer these questions properly requires the development of new revealed preference tests.

Originating in the works of Varian (1983a,b, 1988) and Green and Srivastava (1986), some more recent papers which pursue these questions include Diewert (2012), Bayer *et al.* (2013), Kubler, Selden, and Wei (2014, 2017), Echenique and Saito (2015), Chambers, Liu, and Martinez (2016), Chambers, Echenique, and Saito (2016), Nishimura, Ok, and Quah (2017), Echenique, Imai, and Saito (2019, 2021), Polisson, Quah, and Renou (2020), and de Clippel and Rozen (2021). We compare our approach and contribution to existing work along four dimensions—methods, measures, tests, and power.

[16]For overviews of the revealed preference literature, see Crawford and De Rock (2014) and Chambers and Echenique (2016), as well as the papers by Afriat (2012), Diewert (2012), Varian (2012), and Vermeulen (2012), published in a special issue of the *Economic Journal* on the foundations of revealed preference.

Methods. With the exception of the GRID method, all other 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 in our test of the model.[17] 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 $1 \geqslant e^{*} \geqslant e^{**} \geqslant e^{***} > 0$, where an index of 1 implies exact agreement with a given model.

The use of a common index across different models means that we can perform a comprehensive test of each relevant model (in which all the axioms of a model 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 several related reasons: we know how to compute it for the three models under consideration; these computations can be implemented efficiently; and it is the most

[17]For further discussion of this issue, see Polisson, Quah, and Renou (2020). A subject who maximizes expected utility will pass our test and be classified as EUT-rationalizable, even if that subject is *not* globally risk averse. For an example of choice data that are EUT-rationalizable but only with a non-concave Bernoulli index, see Section A4 of the Online Appendix in Polisson, Quah, and Renou (2020). (Note that Polisson, Quah, and Renou (2020) also develops a test for the case where the Bernoulli index is required to be concave.) This empirical distinction runs in contrast with the Afriat (1967) result on basic rationalizability, where concavity of the utility function (not necessarily of the expected utility form) is without loss of generality.

commonly used measure of goodness-of-fit.[18,19]

de Clippel and Rozen (2021) proposes a different index to measure goodness-of-fit which is applicable 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 (2021) proposes essentially the same index as de Clippel and Rozen (2021) for expected utility, albeit with a somewhat different motivation. This index (or collection of indices) relies on a first-order (condition) approach, so they are only applicable to models representable by quasiconcave utility functions (defined on the space of contingent consumption). As such, it is not ideal for our purposes since we want to avoid imposing a concave Bernoulli index (or, more generally, a quasiconcave utility function) as a rationality requirement.

Tests. We create individual-level non-parametric *permutation* (randomization) 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, we obtain the (empirical) distribution functions for the test statistics under the null hypotheses—that choices are as FOSD-rationalizable as they are rationalizable ($e^{**} = e^{*}$) and as EUT-rationalizable as they are FOSD-rationalizable ($e^{***} = e^{**}$)—directly from the individual-level data. We are not aware of similar statistical tests performed in other work.

Power. A number of recent papers—including Polisson, Quah, and Renou (2020), de Clippel and Rozen (2021), and Echenique, Imai, and Saito (2021)—analyze the experimental data from Choi *et al.* (2014). This experiment is identical to Choi *et al.*

[18] 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. Recently, Dziewulski (2020) provides a further behavioral interpretation for the CCEI based on a decision maker's cognitive inability to distinguish between bundles that are sufficiently similar.

[19] The 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 in the case of the EUT model) and therefore it is not practically implementable for us, given the size of our datasets and the scope of our empirical exercise. For more on the computation of this index to measure rationalizability, FOSD-rationalizability, and EUT-rationalizability, see Polisson, Quah, and Renou (2020).

(2007a), except that it consists of 25, rather than 50, decision problems involving two (equiprobable) states of nature and two associated Arrow securities. Echenique, Imai, and Saito (2021) 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 common thread in all these experiments is that there are two states and two securities.

The experiment reported in this paper consists of 50 decision problems involving *three* (equiprobable) states with *three* associated Arrow 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. However, our power analysis indicates that having three states significantly enhances the discriminatory power of the experiment, especially with respect to EUT-rationalizability, when compared to experiments with two states (and 25, or indeed 50, observations). Given that the primary purpose of this paper to reach a robust empirical conclusion on the sources of departure from EUT, our use of a more discriminating choice environment is crucial.

To conclude, Polisson, Quah, and Renou (2020), de Clippel and Rozen (2021), and Echenique, Imai, and Saito (2021) all develop new methodologies and apply their techniques to existing experimental data. Echenique, Imai, and Saito (2021) 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 the analysis cannot separate the empirical validity of each of the axioms on which EUT is based. More closely related to our theme, Polisson, Quah, and Renou (2020) observes a relatively small gap between FOSD-rationalizability and EUT-rationalizability; notwithstanding the use of a different measure, de Clippel and Rozen (2021) draws a similar conclusion. The focus of both Polisson, Quah, and Renou (2020) and de Clippel and Rozen (2021), however, is methodological rather than empirical and both also rely on existing two-dimensional datasets in their empirical analyses; as acknowledged by de Clippel and Rozen

(2021), power issues cast doubts on the robustness of their empirical conclusions. In this paper, our findings rely on new experimental data with three-dimensional budget sets and 50 observations per subject. A thorough analysis of these data allows us to establish conclusively that subjects have multiple sources of EUT violations and, for the vast majority, violations of ordering and/or monotonicity rather than violations of independence are the main sources of departure from EUT.

6 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.[20]

Combining theoretical tools, experimental methods, and non-parametric 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.

[20]Bell (1982), Fishburn (1982), and Loomes and Sugden (1982) (simultaneously) propose a model of nontransitive risk preference. Loomes and Sugden (1987) 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.

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 respecting FOSD), and estimating its parameters leads to model misspecification. As a result, the estimated preference will not be the true underlying preference, if such a preference ordering even exists, and positive predictions and welfare conclusions based on these models will be misleading.[21] 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-defined utility function and are thus of a higher quality (Kariv and Silverman, 2013).

To conclude, by applying the latest revealed preference techniques to an experiment involving three states with three associated securities, we provide strong *comprehensive* and *nonparametric* tests of complete representations of preferences under risk. Our main result is that while the vast majority of our subjects have statistically significant violations of independence, for many subjects these violations are minor when compared against violations of ordering and monotonicity. As EUT lies at the very heart of economics, these results have important implications for both economic theory and economic policy.

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

[21] 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.

References

Abdellaoui, M. 2002. "A Genuine Rank-Dependent Generalization of the von Neumann-Morgenstern Expected Utility Theorem." *Econometrica* 70(2): 717–736.

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.

———. 2012. "Afriat's Theorem and the Index Number Problem." *Economic Journal* 122(560): 295–304.

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.

Arrow, K. J. 1964. "The Role of Securities in the Optimal Allocation of Risk-bearing." *Review of Economic Studies* 31(2): 91–96.

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. 2021. "On the Experimental Robustness of the Allais Paradox." *American Economic Journal: Microeconomics* Forthcoming.

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. 2021. "The Development Gap in Economic Rationality of Future Elites." Unpublished paper.

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. and F. Echenique. 2016. *Revealed Preference Theory.* Cambridge: Cambridge University Press.

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.

Cox, J. C. 1997. "On Testing the Utility Hypothesis." *Economic Journal* 107(443): 1054–1078.

Crawford, I. and B. De Rock. 2014. "Empirical Revealed Preference." *Annual Reviews* 6: 503–524.

de Clippel, G. and K. Rozen. 2021. "Relaxed Optimization: How Close Is a Consumer to Satisfying First-Order Conditions?" Unpublished paper.

Debreu, G. 1954. "Representation of a Preference Ordering by a Numerical Function." In *Decision Processes*, edited by R. M. Thrall, C. H. Coombs, and R. L. Davis. New York: John Wiley and Sons, 159–165.

———. 1960. "Topological Methods in Cardinal Utility Theory." In *Mathematical Methods in the Social Sciences*, edited by K. J. Arrow, S. Karlin, and P. Suppes. Stanford: Stanford University Press, 16–26.

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

Diecidue, E. and P. P. Wakker. 2001. "On the Intuition of Rank-Dependent Utility." *Journal of Risk and Uncertainty* 23(3): 281–298.

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. 2019. "Decision Making under Uncertainty: An Experimental Study in Market Settings." Unpublished paper.

———. 2020. "Testable Implications of Models of Intertemporal Choice: Exponential Discounting and Its Generalizations." *American Economic Journal: Microeconomics* 12(4): 114–43.

———. 2021. "Approximate Expected Utility Rationalization." Unpublished paper.

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., 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, 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.

Houthakker, H. S. 1950. "Revealed Preference and the Utility Function." *Economica* 17(66): 159–174.

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.

———. 2017. “What Are Asset Demand Tests of Expected Utility Really Testing?” *Economic Journal* 127(601): 784–808.

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.

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

———. 1987. “Testing for Regret and Disappointment in Choice Under Uncertainty.” *Economic Journal* 97: 118–129.

Machina, M. J. 1982. “‘Expected Utility’ Analysis without the Independence Axiom.” *Econometrica* 50(2): 277–323.

———. 1994. “Review of ‘Generalized Expected Utility Theory: The Rank-Dependent Model’.” *Journal of Economic Literature* 32(3): 1237–1238.

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.

Mattei, A. 2000. “Full-Scale Real Tests of Consumer Behavior using Experimental Data.” *Journal of Economic Behavior and Organization* 43(4): 487–497.

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

Prelec, D. 1998. “The Probability Weighting Function.” *Econometrica* 66(3): 497–527.

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.

Samuelson, P. A. 1938. "A Note on the Pure Theory of Consumer's Behaviour." *Economica* 5(17): 61–71.

———. 1948. "Consumption Theory in Terms of Revealed Preference." *Economica* 15(60): 243–253.

———. 1950. "The Problem of Integrability in Utility Theory." *Economica* 17(68): 355–385.

Segal, U. 1990. "Two-Stage Lotteries without the Reduction Axiom." *Econometrica* 58(2): 349–377.

Sippel, R. 1997. "An Experiment on the Pure Theory of Consumer's Behaviour." *Economic Journal* 107(444): 1431–1444.

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.

———. 2012. "Revealed Preference and its Applications." *Economic Journal* 122(560): 332–338.

Vermeulen, F. 2012. "Foundations of Revealed Preference: Introduction." *Economic Journal* 122(560): 287–294.

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.

———. 1994. "Separating Marginal Utility and Probabilistic Risk Aversion." *Theory and Decision* 36: 1–44.

Yaari, M. E. 1987. "The Dual Theory of Choice under Risk." *Econometrica* 55(1): 95–115.

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.

Appendix I

Theoretical Framework

Our experimental data are generated by individual subjects solving a series of randomly generated portfolio-choice problems. In our setting, there are S states of nature, denoted by $s = 1, \ldots, S$. The probability of state s is commonly known to be $\pi_s > 0$, with $\sum_{s=1}^{S} \pi_s = 1$, so that $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_S) \gg 0$ denotes the vector of state probabilities.[1] For each state s, there is an Arrow security that pays one token (the experimental currency) in state s and nothing in the other state(s). The amount of consumption in state s is denoted by $x_s \geqslant 0$, and the portfolio of securities may be written as $\mathbf{x} = (x_1, \ldots, x_S) \geqslant \mathbf{0}$.

In the experiment, each subject has a budget of 1, which has to be allocated among the Arrow securities, with $p_s > 0$ denoting the price of security s. Formally the subject chooses a portfolio $\mathbf{x} \geqslant \mathbf{0}$ among those which satisfy the constraint $\mathbf{p} \cdot \mathbf{x} = 1$, where $\mathbf{p} = (p_1, \ldots, p_S) \gg \mathbf{0}$ denotes the vector of state prices. The subject can choose any portfolio $\mathbf{x}$ satisfying the budget constraint.

Let $\mathcal{D} := (\mathbf{p}^i, \mathbf{x}^i)$ be the dataset generated by an individual subject's choices from these linear budget sets, where $\mathbf{p}^i$ denotes the i-th observation of the price vector and $\mathbf{x}^i$ denotes the corresponding demand allocation by the subject. The subject's total expenditure is fixed at 1 throughout, so $\mathbf{p}^i \cdot \mathbf{x}^i = 1$ for all observations i. The experimental design required subjects to solve a sequence of 50 decision problems (so $\mathcal{D}$ has 50 observations) involving three-dimensional budget sets ($S = 3$), and we also compare these results against the results from otherwise identical experiments involving two-dimensional budget lines ($S = 2$). In all of the two- and three-dimensional experiments that we consider, the states are equiprobable, though the theoretical results which we review below do not hinge on this feature.

Rationalizability (e^*) Recall, from the main paper, that we refer to a utility function $U : \mathbb{R}^S_+ \to \mathbb{R}$ as *well-behaved* if it is continuous and increasing, where the latter means that

[1] As a matter of notation, for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^S$, we say that $\mathbf{x} \geqslant \mathbf{y}$ if $x_s \geqslant y_s$ for all s; $\mathbf{x} > \mathbf{y}$ if $\mathbf{x} \geqslant \mathbf{y}$ and $\mathbf{x} \neq \mathbf{y}$; and $\mathbf{x} \gg \mathbf{y}$ if $x_s > y_s$ for all s.

$U(\mathbf{x}'') > U(\mathbf{x}')$ if $\mathbf{x}'' > \mathbf{x}'$. A utility function U rationalizes $\mathcal{D}$ if $U(\mathbf{x}^i) \geqslant U(\mathbf{x})$ for all

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

In other words, the utility of $\mathbf{x}^i$ is weakly higher than that of any alternative that is weakly cheaper at the price vector $\mathbf{p}^i$. When a dataset $\mathcal{D}$ can be rationalized by a well-behaved utility function U, we say that $\mathcal{D}$ is *rationalizable by a well-behaved utility function*, or simply *rationalizable*. Afriat's (1967) Theorem characterizes rationalizable datasets via the Generalized Axiom of Revealed Preference (GARP).

Let $\mathcal{X} = \{\mathbf{x}^i\}$ be the set of portfolios observed across all observations i. For any $\mathbf{x}^i$, $\mathbf{x}^j \in \mathcal{X}$, we say that $\mathbf{x}^i$ is *directly revealed preferred* to $\mathbf{x}^j$ (and denote this relation by $\mathbf{x}^i \, R^D \, \mathbf{x}^j$) if $\mathbf{p}^i \cdot \mathbf{x}^i \geqslant \mathbf{p}^i \cdot \mathbf{x}^j$. GARP requires that if $\mathbf{x}^i$ is revealed preferred to $\mathbf{x}^j$ (either directly or indirectly via a sequence of other portfolio choices), then $\mathbf{x}^i$ must cost at least as much as $\mathbf{x}^j$ at the prices prevailing when $\mathbf{x}^j$ is chosen. To be precise, we define on $\mathcal{X}$ the *revealed preference* relation, where $\mathbf{x}^i$ is revealed preferred to $\mathbf{x}^j$ (denoted by $\mathbf{x}^i \, R \, \mathbf{x}^j$) if there is a sequence of observations $i_1, i_2, \ldots, i_n$ such that

$$\mathbf{x}^i \, R^D \, \mathbf{x}^{i_1} \, R^D \, \mathbf{x}^{i_2} \, R^D \, \cdots R^D \mathbf{x}^{i_n} \, R^D \, \mathbf{x}^j.$$

In other words, the relation R is the transitive closure of the relation R^D. We also define the *strict direct revealed preference* relation P^D, where $\mathbf{x}^i \, P^D \, \mathbf{x}^j$ if $\mathbf{p}^i \cdot \mathbf{x}^i > \mathbf{p}^i \cdot \mathbf{x}^j$. GARP requires that, for any $\mathbf{x}^i, \mathbf{x}^j \in \mathcal{X}$,

$$\text{if } \mathbf{x}^i \, R \, \mathbf{x}^j, \text{ then } \mathbf{x}^j \, P^D \, \mathbf{x}^i \text{ does not hold.}$$

The term "revealed preference" for the relation R is very intuitive, since if a dataset can be rationalized by some utility function U, then $U(\mathbf{x}^i) \geqslant U(\mathbf{x}^j)$ if $\mathbf{x}^i \, R \, \mathbf{x}^j$. Furthermore, it is not hard to show that if U is locally nonsatiated, then $U(\mathbf{x}^i) > U(\mathbf{x}^j)$ if $\mathbf{x}^i \, P^D \, \mathbf{x}^j$. It follows from these observations that if $\mathcal{D}$ is rationalizable by a locally nonsatiated utility function then it must obey GARP, since it impossible for $U(\mathbf{x}^i) \geqslant U(\mathbf{x}^j)$ and for $U(\mathbf{x}^j) > U(\mathbf{x}^i)$ to

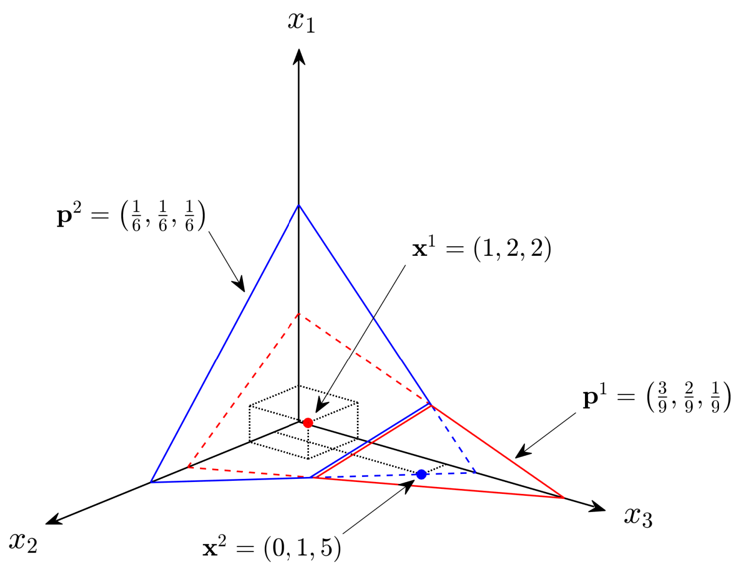


Figure 1: Violation of Rationalizability

hold simultaneously.[2] The substantive part of Afriat's Theorem says that if $\mathcal{D}$ obeys GARP then it is rationalizable by a concave and well-behaved utility function. Notice that the two statements are not completely symmetric: GARP holds whenever a dataset is generated by a locally nonsatiated utility function, but whenever GARP holds on a dataset, it can also be rationalized by a utility function with properties that are stronger than local nonsatiation.

Figure 1 illustrates a simple violation of GARP involving two budget sets $\mathbf{p}^1 = \left(\frac{3}{9}, \frac{2}{9}, \frac{1}{9}\right)$ and $\mathbf{p}^2 = \left(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right)$, and two portfolio allocations $\mathbf{x}^1 = (1, 2, 2)$ and $\mathbf{x}^2 = (0, 1, 5)$. It is clear that $\mathbf{x}^1 \, P^D \, \mathbf{x}^2$ and $\mathbf{x}^2 \, P^D \, \mathbf{x}^1$ since $\mathbf{p}^1 \cdot \mathbf{x}^1 > \mathbf{p}^1 \cdot \mathbf{x}^2$ and $\mathbf{p}^2 \cdot \mathbf{x}^2 > \mathbf{p}^2 \cdot \mathbf{x}^1$.

GARP provides an exact test of utility maximization (either the data satisfy GARP or they do not). To account for the possibility of errors, we assess how close a dataset is to being rationalizable by using Afriat's (1972, 1973) Critical Cost Efficiency Index (CCEI), which we shall now explain.

Given a number $e \in (0, 1]$, a dataset $\mathcal{D}$ is *rationalizable at cost efficiency* e if there is a

[2] A utility function $U : \mathbb{R}^S_+ \to \mathbb{R}$ is locally nonsatiated if, in any open ball around $\mathbf{x} \in \mathbb{R}^S_+$, there is some $\mathbf{x}'$ such that $U(\mathbf{x}') > U(\mathbf{x})$. The eagle-eyed reader may notice that in our experiments each subject at observation i chooses from the budget boundary $\overline{\mathcal{B}}^i = \{\mathbf{x} \in \mathbb{R}^S_+ : \mathbf{p}^i \cdot \mathbf{x} = 1\}$ rather than from the budget set $\mathcal{B}^i$, so that we ought to check that $\mathcal{D}$ satisfies GARP if $\mathbf{x}^i$ is a utility-maximizing choice from $\overline{\mathcal{B}}^i$. This is indeed the case provided that U is continuous and locally nonsatiated; these assumptions on U guarantee that $\arg\max_{\mathbf{x} \in \overline{\mathcal{B}}^i} U(\mathbf{x}) = \arg\max_{\mathbf{x} \in \mathcal{B}^i} U(\mathbf{x})$ so that $U(\mathbf{x}^i) \geqslant U(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{B}^i$ and $U(\mathbf{x}^i) > U(\mathbf{x})$ if $\mathbf{x} \in \mathcal{B}^i \setminus \overline{\mathcal{B}}^i$. In particular, this implies that $U(\mathbf{x}^i) \geqslant U(\mathbf{x}^j)$ if $\mathbf{x}^i \, R \, \mathbf{x}^j$ and $U(\mathbf{x}^i) > U(\mathbf{x}^j)$ if $\mathbf{x}^i \, P^D \, \mathbf{x}^j$.

well-behaved utility function U such that $U(\mathbf{x}^i) \geqslant U(\mathbf{x})$ for all

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

It is not difficult to see that *every* dataset $\mathcal{D}$ could be rationalized by a well-behaved utility function at an efficiency level e for some $e \in (0, 1]$ that is sufficiently close to zero. Afriat's CCEI, denoted by e^*, is the largest value of e associated with the dataset $\mathcal{D}$; formally,

$$e^* = \sup\{e \in (0, 1] : \mathcal{D} \text{ is rationalizable at cost efficiency } e\}.$$

A subject with a CCEI of $e^* < 1$ makes mistakes, in the sense that there is at least one observation k for which $U(\mathbf{x}^k) < U(\mathbf{x})$ for some $\mathbf{x} \in \mathcal{B}^k$, but the cost inefficiency is bounded in the sense that $\mathbf{p} \cdot \mathbf{x} \geqslant e^*$; thus the subject could switch to a bundle $\mathbf{x}$ that gives the same utility as $\mathbf{x}^k$ and spend less, but the savings is no more than $1 - e^*$.

The coefficient e^* can be straightforwardly obtained through a binary search, once there is a way to check if a dataset is rationalizable at cost efficiency e for any given value of e. Very conveniently, rationalizability at cost efficiency e can be characterized by a generalized version of GARP. We define the direct revealed preference relation at efficiency e (denoted by $R^D(e)$) as follows: $\mathbf{x}^i\, R^D(e)\, \mathbf{x}^j$ if $e\,\mathbf{p}^i \cdot \mathbf{x}^i \geqslant \mathbf{p}^i \cdot \mathbf{x}^j$. The revealed preference relation $R(e)$ is the transitive closure of $R^D(e)$. Similarly, the strict direct revealed preference relation at efficiency e (denoted by $P^D(e)$) is defined as follows: $\mathbf{x}^i\, P^D(e)\, \mathbf{x}^j$ if $e\,\mathbf{p}^i \cdot \mathbf{x}^i > \mathbf{p}^i \cdot \mathbf{x}^j$. e-GARP requires that, for any $\mathbf{x}^i, \mathbf{x}^j \in \mathcal{X}$,

$$\text{if } \mathbf{x}^i\, R(e)\, \mathbf{x}^j, \text{ then } \mathbf{x}^j\, P^D(e)\, \mathbf{x}^i \text{ does not hold.}$$

It is straightforward to check that if a dataset $\mathcal{D}$ can be rationalized at cost efficiency e by a locally nonsatiated utility function, then it will satisfy e-GARP; conversely, if $\mathcal{D}$ satisfies e-GARP, then it is rationalizable at efficiency e by a concave and well-behaved utility function (see Afriat (1973)).

FOSD-Rationalizability (e^{**}) Nishimura et al. (2017) shows that a further modification of GARP can be used to test whether a dataset $\mathcal{D}$ is rationalizable (at cost efficiency e) by a continuous utility function that is increasing with respect to a given preorder $\trianglerighteq$ on the choice space. This result is convenient for our purposes because for a utility function U to be monotone with respect to FOSD simply means that it is increasing with respect to the preorder $\trianglerighteq$, where $\mathbf{x}'' \trianglerighteq \mathbf{x}'$ if $\mathbf{x}'$ and $\mathbf{x}''$ (when considered as distributions given the vector of state probabilities $\boldsymbol{\pi}$) have the property that $\mathbf{x}''$ first-order stochastically dominates $\mathbf{x}'$. In our experiments, each state is equally likely; thus, $\mathbf{x}'' \trianglerighteq \mathbf{x}'$ if there is some permutation of the entries in $\mathbf{x}''$ such that the permuted allocation is entry-by-entry weakly greater than $\mathbf{x}'$. For example, $(1, 0, 1) \trianglerighteq (0, 1, 0)$ since $(1, 1, 0) \geqslant (0, 1, 0)$. In this case, a well-behaved utility function is monotone with respect to FOSD if and only it is symmetric.

We say that a dataset $\mathcal{D}$ is *FOSD-rationalizable at cost efficiency* e if it can be rationalized at cost efficiency e by a well-behaved utility function that is monotone with respect to FOSD. The rationalizabily score e^{**} is given by

$$e^{**} = \sup \{e \in (0, 1] : \mathcal{D} \text{ is FOSD-rationalizable at cost efficiency } e\}.$$

The FOSD-rationalizability at cost efficiency e of a dataset $\mathcal{D}$ can be characterized by a generalized notion of GARP which we shall now explain.

We define the direct revealed preference relation at efficiency e (denoted by $R^D_{\trianglerighteq}(e)$) as follows: $\mathbf{x}^i \, R^D_{\trianglerighteq}(e) \, \mathbf{x}^j$ if there exists some $\mathbf{y}$ such that $e\,\mathbf{p}^i \cdot \mathbf{x}^i \geqslant \mathbf{p}^i \cdot \mathbf{y}$ and $\mathbf{y} \trianglerighteq \mathbf{x}^j$. The revealed preference relation $R_{\trianglerighteq}(e)$ is the transitive closure of $R^D_{\trianglerighteq}(e)$. Similarly, the strict direct revealed preference relation at efficiency e (denoted by $P^D_{\trianglerighteq}(e)$) is defined as follows: $\mathbf{x}^i \, P^D_{\trianglerighteq}(e) \, \mathbf{x}^j$ if there exists some $\mathbf{y}$ such that $e\,\mathbf{p}^i \cdot \mathbf{x}^i \geqslant \mathbf{p}^i \cdot \mathbf{y}$ and $\mathbf{y} \trianglerighteq \mathbf{x}^j$ but $\mathbf{x}^j \not\trianglerighteq \mathbf{y}$ (in other words, $\mathbf{y}$ strictly first-order stochastically dominates $\mathbf{x}^j$). e-GARP($\trianglerighteq$) requires that, for any $\mathbf{x}^i, \mathbf{x}^j \in \mathcal{X}$,

$$\text{if } \mathbf{x}^i \, R_{\trianglerighteq}(e) \, \mathbf{x}^j, \text{ then } \mathbf{x}^j \, P^D_{\trianglerighteq}(e) \, \mathbf{x}^i \text{ does not hold.}$$

A dataset $\mathcal{D}$ satisfies e-GARP($\trianglerighteq$) if and only if it is FOSD-rationalizable at efficiency e.

To illustrate in simple terms how the test works, Figure 2 depicts the same two budget sets as in Figure 1, $\mathbf{p}^1 = \left(\frac{3}{9}, \frac{2}{9}, \frac{1}{9}\right)$ and $\mathbf{p}^2 = \left(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right)$, with the portfolio allocations $\mathbf{x}^1 =$

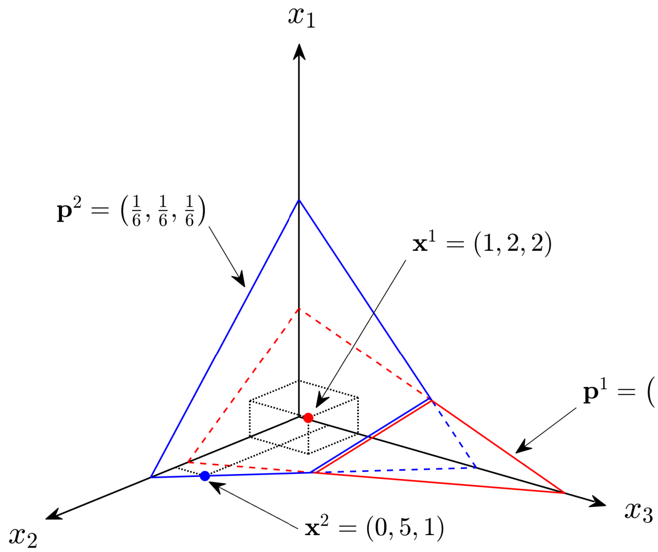


Figure 2: Violation of FOSD-Rationalizability

$(1, 2, 2)$ and $\mathbf{x}^2 = (0, 5, 1)$. These choices are rationalizable but not FOSD-rationalizable (with equiprobable states) because $\mathbf{x}^1\, P^D_{\trianglerighteq}(1)\, \mathbf{x}^2$ and $\mathbf{x}^2\, P^D_{\trianglerighteq}(1)\, \mathbf{x}^1$. It is clear that $\mathbf{p}^2 \cdot \mathbf{x}^2 > \mathbf{p}^2 \cdot \mathbf{x}^1$, but it is also the case that $\mathbf{p}^1 \cdot \mathbf{x}^1 > \mathbf{p}^1 \cdot \mathbf{y}$ where $\mathbf{y} = (0, 1, 5) \trianglerighteq (0, 5, 1) = \mathbf{x}^2$.

Violations of FOSD could be regarded as errors, regardless of the agent's risk attitude—that is, they represent the agent's failure to account for the fact that some allocations give payoff distributions with unambiguously lower returns than others. As a result, the most prominent non-EUT models have been constructed/amended to avoid violations of FOSD. There are, however, some notable exceptions. For example, Kőszegi and Rabin's (2007) reference-dependent risk preferences may violate FOSD due to (excessive) loss aversion — see Masatlioglu and Raymond's (2016) characterization. However, the Kőszegi and Rabin (2007) utility function U is locally nonsatiated and, in the case where states are equiprobable (as in our experiments), it must respect symmetry. It is straightforward to check that a subject who maximizes a symmetric and locally nonsatiated utility function at cost efficiency e would generate a dataset $\mathcal{D}$ satisfying e-GARP($\trianglerighteq$) (with $\trianglerighteq$ being the preorder corresponding to equiprobable states) and thus $\mathcal{D}$ is FOSD-rationalizable at cost efficiency e. In other words, in the context of our experiments, reference-dependent risk preferences cannot do better in explaining a subject's data than the family of utility functions that are monotone with respect to FOSD.

EUT-Rationalizability (e^{***}) Polisson et al. (2020) develops a revealed preference method to test whether choice data under risk are consistent with maximizing a utility function that has some special structure. The method restricts an infinite choice set to a finite grid, and is thus called the method of Generalized Restriction of Infinite Domains (GRID). GRID tests are mechanically distinct from GARP tests (in the sense that they do not involve constructing revealed preference relations and checking for strict cycles), but they are fully nonparametric (within the specified class of utility functions) and can also be used to measure inconsistencies. This is the approach that we use to test expected utility.

We say that a dataset $\mathcal{D}$ is *EUT-rationalizable at cost efficiency* e if it can be rationalized at cost efficiency e by a well-behaved utility function U taking the expected utility form, i.e., if there is a continuous and increasing Bernoulli index $u : \mathbb{R}_+ \to \mathbb{R}$ such that $U(\mathbf{x}) = \sum_{s=1}^{S} \pi_s u(x_s)$. Following Polisson et al. (2020), let $\mathcal{Y}$ be the set that contains any demand level observed in a given dataset $\mathcal{D}$ plus zero, that is

$$\mathcal{Y} := \{x \in \mathbb{R}_+ : x = x_s^i \text{ for some } (i,s)\} \cup \{0\}.$$

We then form the finite grid $\mathcal{G} = \mathcal{Y}^S \subset \mathbb{R}_+^S$ which is a restriction of the choice space $\mathbb{R}_+^S$ to allocations comprised of demand levels that have been observed in the dataset $\mathcal{D}$. We claim that EUT-rationalizability at cost efficiency e requires the existence of a real number $\bar{u}(y)$ associated with each $y \in \mathcal{Y}$, with $\bar{u}(y') > \bar{u}(y)$ whenever $y' > y$, such that at each observation of $(\mathbf{p}^i, \mathbf{x}^i)$

$$\sum_{s=1}^{S} \pi_s \bar{u}(x_s^i) \geqslant \sum_{s=1}^{S} \pi_s \bar{u}(x_s) \text{ for any } \mathbf{x} \text{ such that } \mathbf{p}^i \cdot \mathbf{x} \leqslant e\, \mathbf{p}^i \cdot \mathbf{x}^i \text{ and } \mathbf{x} \in \mathcal{G},$$

$$\sum_{s=1}^{S} \pi_s \bar{u}(x_s^i) > \sum_{s=1}^{S} \pi_s \bar{u}(x_s) \text{ for any } \mathbf{x} \text{ such that } \mathbf{p}^i \cdot \mathbf{x} < e\, \mathbf{p}^i \cdot \mathbf{x}^i \text{ and } \mathbf{x} \in \mathcal{G}.$$

Indeed, if a dataset $\mathcal{D}$ can be EUT-rationalized at cost efficiency e by a continuous and increasing Bernoulli index u, then these conditions must hold if we choose $\bar{u}(y) = u(y)$ for each $y \in \mathcal{Y}$ since, in the case of the first condition, $\mathbf{x}$ is in $\mathcal{B}^i(e)$ and in the case of the second condition, $\mathbf{x}$ is in the interior of $\mathcal{B}^i(e)$. An important application of the main result of Polisson et al. (2020) is that these conditions are also sufficient for EUT-rationalizality

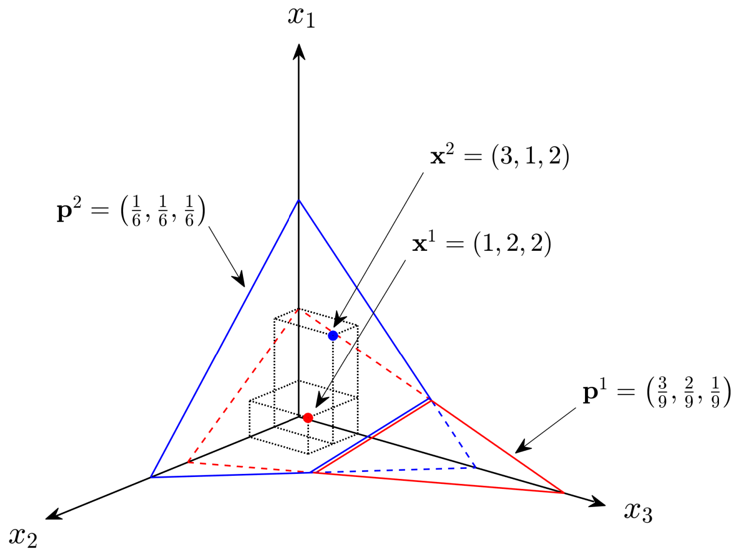


Figure 3: Violation of EUT-Rationalizability

at cost efficiency e. Note that the conditions constitute a finite set of linear inequalities and ascertaining whether or not it has a solution is computationally straightforward. This gives us a way of determining whether a dataset $\mathcal{D}$ is EUT-rationalizable at cost efficiency e and thus allows us to calculate its rationalizability score

$$e^{***} = \sup \{e \in (0, 1] : \mathcal{D} \text{ is EUT-rationalizable at cost efficiency } e\}.$$

To illustrate, Figure 3 depicts the same two budget sets as in Figures 1 and 2, $\mathbf{p}^1 = \left(\frac{3}{9}, \frac{2}{9}, \frac{1}{9}\right)$ and $\mathbf{p}^2 = \left(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right)$, with the portfolio allocations $\mathbf{x}^1 = (1, 2, 2)$ and $\mathbf{x}^2 = (3, 1, 2)$. Assuming that the three states are equiprobable, it is easy to verify that these choices are FOSD-rationalizable, but we claim that they are not EUT-rationalizable. To see this, consider the portfolio allocations $\mathbf{y} = (1, 1, 3)$ and $\mathbf{z} = (2, 2, 2)$ and notice that

$$\mathbf{p}^1 \cdot \mathbf{x}^1 > \mathbf{p}^1 \cdot \mathbf{y} \text{ and } \mathbf{p}^2 \cdot \mathbf{x}^2 = \mathbf{p}^2 \cdot \mathbf{z}.$$

But EUT-rationalizability requires that

$$\tfrac{1}{3}u(1) + \tfrac{1}{3}u(2) + \tfrac{1}{3}u(2) = U(\mathbf{x}^1) > U(\mathbf{y}) = \tfrac{1}{3}u(1) + \tfrac{1}{3}u(1) + \tfrac{1}{3}u(3),$$

$$\tfrac{1}{3}u(3) + \tfrac{1}{3}u(1) + \tfrac{1}{3}u(2) = U(\mathbf{x}^2) \geqslant U(\mathbf{z}) = \tfrac{1}{3}u(2) + \tfrac{1}{3}u(2) + \tfrac{1}{3}u(2),$$

implying that $2u(2) > u(1) + u(3)$ and $u(3) + u(1) \geqslant 2u(2)$, a contradiction. The GRID procedure would also reveal this violation of EUT-rationalizability. To see this, note there must exist real numbers $\bar{u}(1) < \bar{u}(2) < \bar{u}(3)$ satisfying

$$2\bar{u}(2) > \bar{u}(1) + \bar{u}(3) \text{ and } \bar{u}(3) + \bar{u}(1) \geqslant 2\bar{u}(2),$$

which is an impossibility.

Appendix II

Experimental Instructions

Introduction This is an experiment in decision-making. Research foundations have provided funds for conducting this research. Your payoffs will depend partly only on your decisions and partly on chance. It will not depend on the decisions of the other participants in the experiments. Please pay careful attention to the instructions as a considerable amount of money is at stake.

The entire experiment should be complete within an hour and a half. At the end of the experiment you will be paid privately. At this time, you will receive $5 as a participation fee (simply for showing up on time). Details of how you will make decisions and receive payments will be provided below.

During the experiment we will speak in terms of experimental tokens instead of dollars. Your payoffs will be calculated in terms of tokens and then translated at the end of the experiment into dollars at the following rate:

$$2 \text{ Tokens} = 1 \text{ Dollar}$$

A decision problem In this experiment, you will participate in 50 independent decision problems that share a common form. This section describes in detail the process that will be repeated in all decision problems and the computer program that you will use to make your decisions.

In each decision problem you will be asked to allocate tokens between three accounts, labeled x, y and z. Each choice will involve choosing a point on a three-dimensional graph representing possible token allocations, x / y / z. The x account corresponds to the x-axis, the y account corresponds to the y-axis and the z account corresponds to the z-axis in a three-dimensional graph. In each choice, you may choose any combination of x / y / z that is on the plane that is shaded in gray. Examples of planes that you might face appear in Figure 1.

Each decision problem will start by having the computer select such a plane randomly from the set of planes that intersect with at least one of the axes (x, y or z) at 50 tokens or more but with no intercept exceeding 100 tokens. The planes selected for you in different decision problems are independent of each other and independent of the planes selected for any of the other participants in their decision problems.

For example, as illustrated in Figure 2, choice A represents an allocation in which you allocate approximately 20 tokens in the x account, 21 tokens in the y account, and 30 tokens in the z account. Another possible allocation is B, in which you allocate approximately 40 tokens in the x account, 17 tokens in the y account, and 11 tokens in the z account.

To choose an allocation, use the mouse to move the pointer on the computer screen to the allocation that you desire. On the right hand side of the program dialog window, you will be informed of the exact allocation that the pointer is located. When you are ready to make your decision, left-click to enter your chosen allocation. After that, confirm your decision by clicking on the Submit button. Note that you can choose only x / y / z combinations that are on the gray plane. To move on to the next round, press the OK button. The computer program dialog window is shown in Figure 3.

Your payoff at each decision round is determined by the number of tokens in each account. At the end of the round, the computer will randomly select one of the accounts, x, y or z. For each participant, account x will be selected with 1/3 chance, account y will be selected with 1/3 chance and account z will be selected with 1/3 chance. You will only receive the number of tokens you allocated to the account that was chosen.

Next, you will be asked to make an allocation in another independent decision. This process will be repeated until all 50 rounds are completed. At the end of the last round, you will be informed the experiment has ended.

Earnings Your earnings in the experiment are determined as follows. At the end of the experiment, the computer will randomly select one decision round from each participant to carry out (that is, 1 out of 50). The round selected depends solely upon chance. For each participant, it is equally likely that any round will be chosen.

The round selected, your choice and your payment will be shown in the large window that appears at the center of the program dialog window. At the end of the experiment, the tokens will be converted into money. Each token will be worth 0.50 Dollars. Your final earnings in the experiment will be your earnings in the round selected plus the $5 show-up fee. You will receive your payment as you leave the experiment.

Rules Your participation in the experiment and any information about your payoffs will be kept strictly confidential. Your payment-receipt and participant form are the only places in which your name and social security number are recorded.

You will never be asked to reveal your identity to anyone during the course of the experiment. Neither the experimenters nor the other participants will be able to link you to any of your decisions. In order to keep your decisions private, please do not reveal your choices to any other participant.

Please do not talk with anyone during the experiment. We ask everyone to remain silent until the end of the last round. If there are no further questions, you are ready to start. An instructor will approach your desk and activate your program.

Figure 1

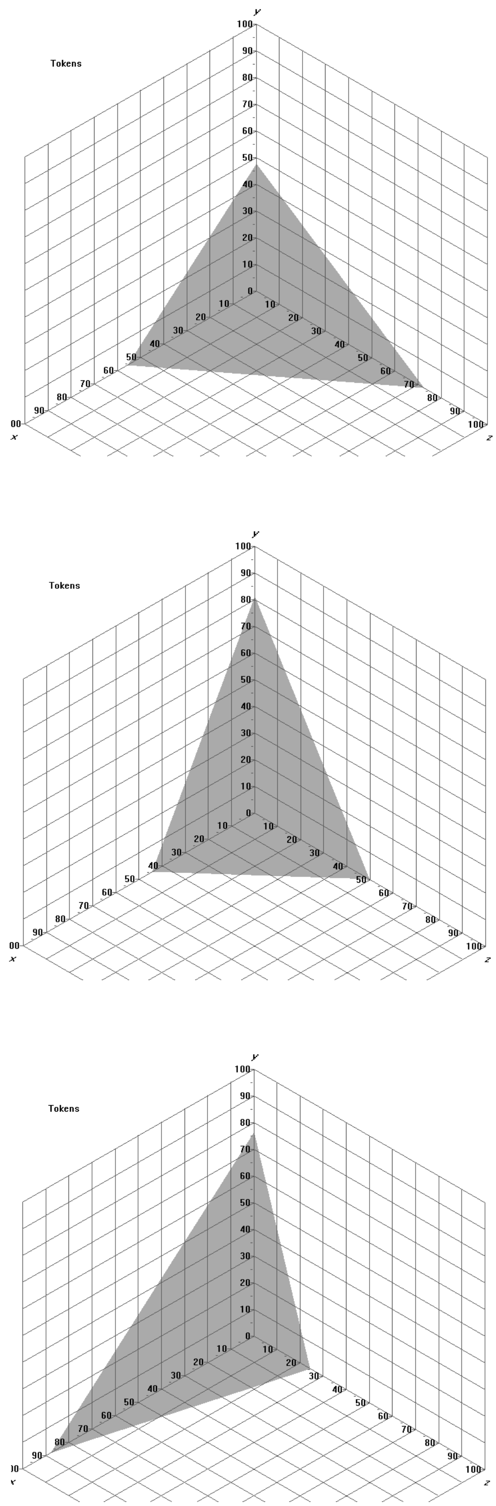
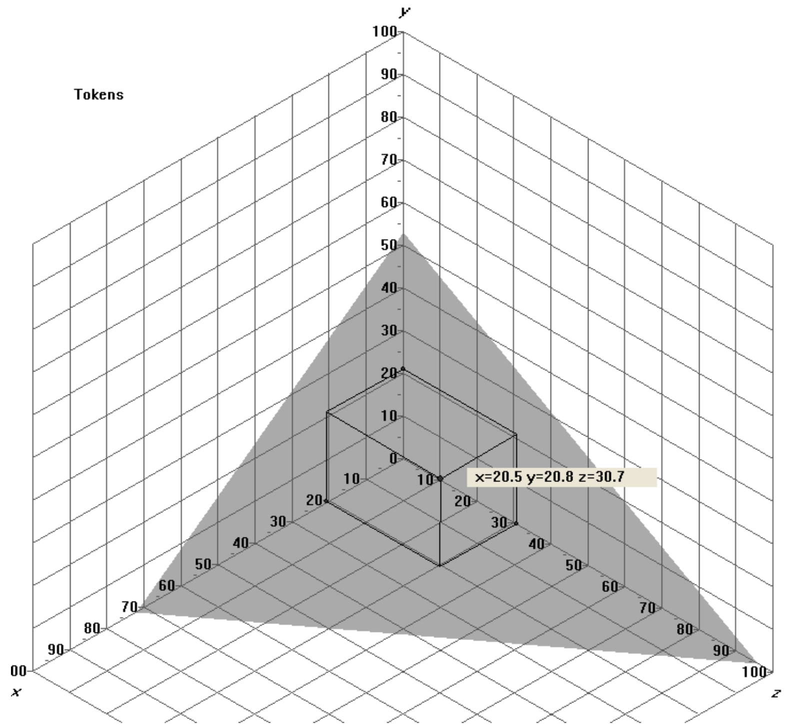


Figure 2

Choice A



Choice B

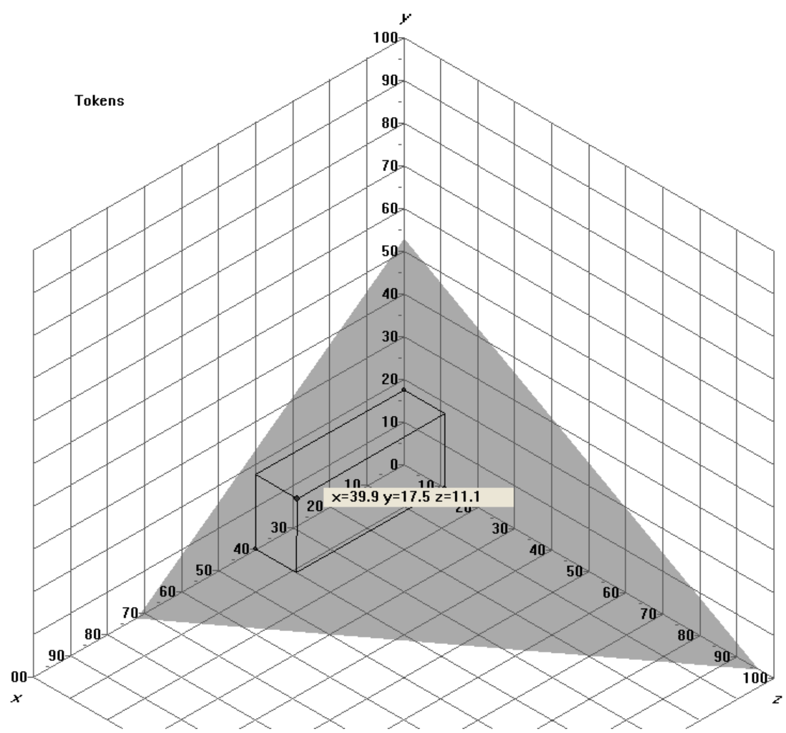
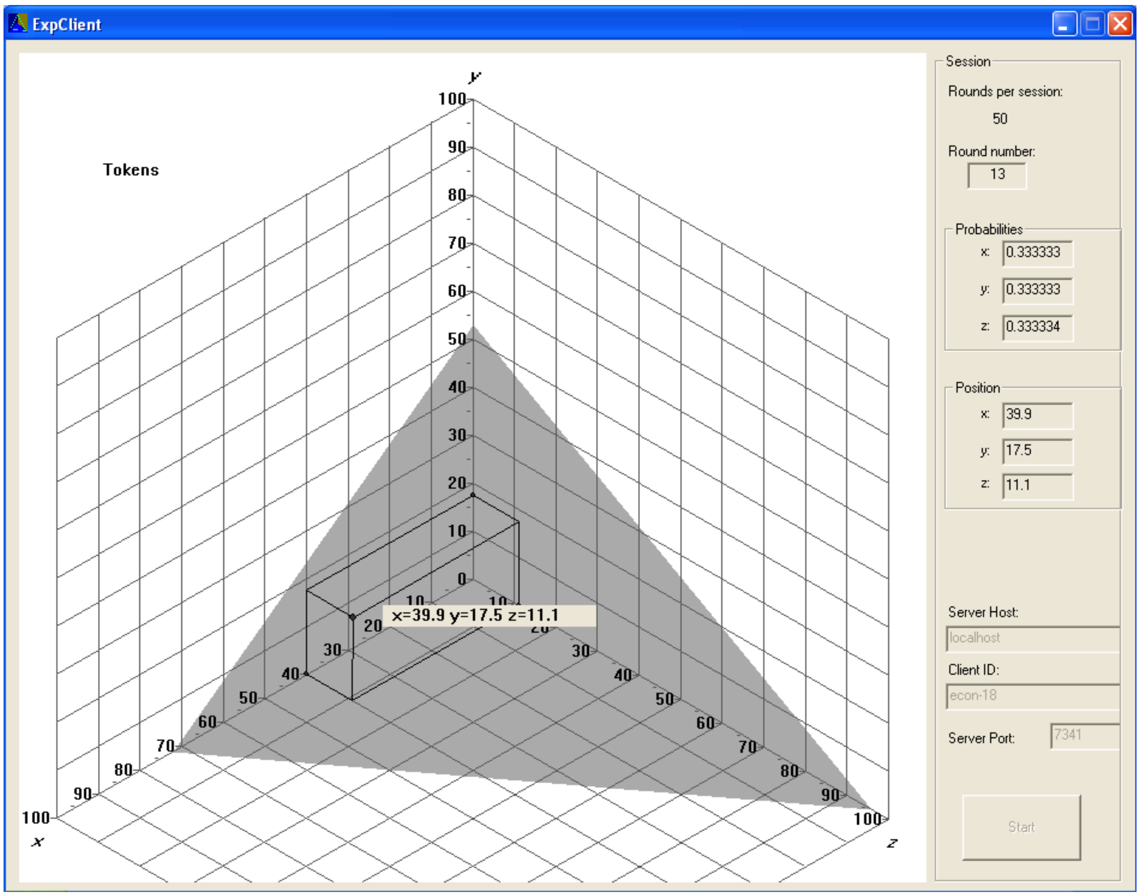


Figure 3


ExpClient
Tokens
x=39.9 y=17.5 z=11.1
y
x
z
Session
Rounds per session:
50
Round number:
13
Probabilities
x: 0.333333
y: 0.333333
z: 0.333334
Position
x: 39.9
y: 17.5
z: 11.1
Server Host:
localhost
Client ID:
econ-18
Server Port:
7341
Start

Appendix III

Two-Dimensional Results

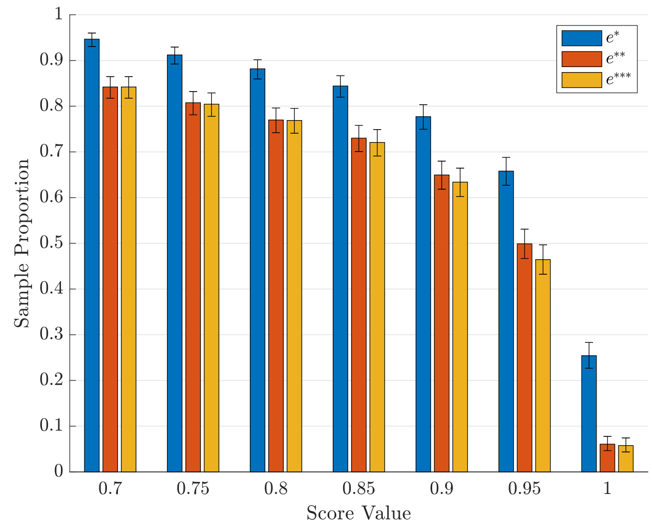
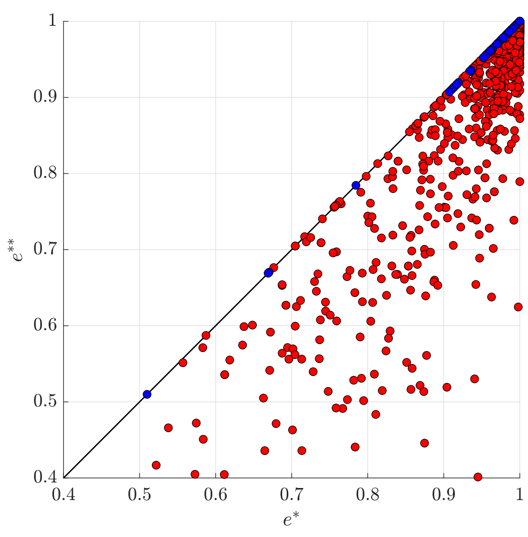
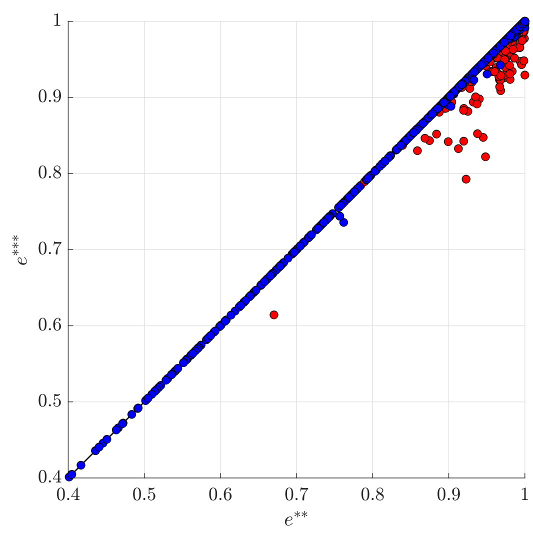


Figure 1: Distributions of Rationalizability Scores



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



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

Figure 2: Scatterplots of Rationalizability Scores

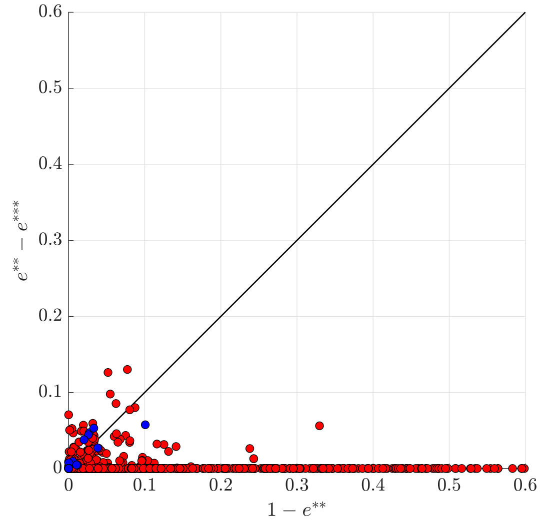


Figure 3: Scatterplots of Score Differences