The Power of Revealed Preference Tests: Ex-Post Evaluation of Experimental Design*

James Andreoni
Department of Economics
University of California, San Diego

Benjamin J. Gillen
Department of Economics
California Institute of Technology

William T. Harbaugh
Department of Economics
University of Oregon

July 23, 2013

Abstract

Revealed preference tests are elegant nonparametric tools that ask whether choice data conforms to optimizing behavior. These tests present a vexing tension between goodness-of-fit and power. If the test finds violations, is there an acceptable tolerance for goodness-of-fit? If no violations are found, was the test demanding enough to be powerful? This paper complements the many on goodness-of-fit by presenting several new indices of power. By focusing on the underlying probability model induced by sampling, we attempt to unify the two approaches. We illustrate applications of the indices, and provide a field guide to applying them to experimental data.

*We are grateful to Oleg Balashov, David Bjerk, Khai Chiong, Ian Crawford, Federico Echenique, Joseph Guse, Shachar Kariv, Grigory Kosenok, Justin McCrary, Matt Shum, and Hal Varian, as well as seminar participants at the California Econometrics Conference, California Institute of Technology, Stanford University, and the University of California Berkeley for helpful comments. We owe special thanks to Gautam Tripathi for important insights at the early stages of this project. We also acknowledge the financial support of the National Science Foundation.

Contents

1 Introduction

One of the most elegant tools to test theories of optimizing behavior is revealed preference. The core axioms of revealed preference are presented in a remarkable series of papers by Hal Varian (1982, 1983, 1984, 1985), which built on earlier work by Afriat (1967, 1972), Hauthakker (1950) and Samuelson (1938). Given a vector of prices p_t and choices x_t at time t, we know that the bundle x_t is preferred to another bundle x if x was affordable when x_t was chosen, $p_t x_t \geq p_t x$. Relying on transitivity of preferences, we can string together chains of these inequalities to rank bundles, even those that were never directly compared by the consumer, and bound possible indifference curves that could have generated this data. Of course, if these chains of inequalities cannot all be mutually satisfied, then the data fail to conform with a model of utility maximization. Hence, revealed preference is both a descriptive and a diagnostic tool.[1]

We begin with a few definitions:

Definition: Directly Revealed Preferred: x_t is directly revealed preferred to x if $p_t x_t \geq p_t x$, and is *strictly* directly revealed preferred if $p_t x_t > p_t x$.

Definition: Revealed Preferred: x_t is revealed preferred to x if there is a chain of directly revealed preferred bundles linking x_t to x.

The revealed preference relation is thus the transitive closure of direct revealed preference and revealed preference tests evaluate the validity of the following axioms:

Definition: Weak Axiom of Revealed Preference (WARP): If x_t is directly revealed preferred to x, then x is not directly revealed preferred to x_t.

Definition: Strong Axiom of Revealed Preference (SARP): If x_t is revealed preferred to x, then x is not revealed preferred to x_t.

Definition: Generalized Axiom of Revealed Preference (GARP): If x_t is revealed preferred to x, then x is not strictly directly revealed preferred to x_t.

[1] Note the same notions can be applied to optimizing by firms, as Varian (1984) demonstrates. For brevity, we will confine our discussion to consumer theory, but it all can be applied to producer theory as well.

The most commonly applied notion of a revealed preference test focuses on Varian's (1982) Generalized Axiom. If the data is consistent with GARP, then there exists a utility function that would have generated the data. That is, the data conforms with a theory of optimizing behavior. A failure to satisfy GARP, on the other hand, precludes the existence of a utility representation for the observed choices.

Two obvious issues arise in interpreting revealed preference tests from data. The first is that the test is extremely sharp – a single violation of GARP results in a rejection of the model. One can naturally ask whether there is some tolerance that can be applied to the data to account for errors in either measurement or choice that can allow some "minor" violations to be accepted within the theory. This is the notion of goodness of fit of the model. There have been several important attempts in the literature to formalize approaches to goodness of fit, the most prominent of which is the Afriat Critical Cost Efficiency Index (CCEI) proposed by Varian (1990, 1991).[2] These techniques allow researchers to not only identify the event that choices violate GARP, but also characterize the welfare loss due to these violations.

The other issue arises when the data fail to reject GARP, leaving researchers to interpret a negative result. If the optimizing model is not, in fact, the correct model, would the revealed preference test applied be sensitive enough to detect it or is the negative result a consequence of weak design? This is a question of the power of the revealed preference test, one closely related to the empirical content of the theory in the testing environment.

In contrast to substantial efforts characterizing the goodness of fit for GARP studies, there have been few formal attempts to characterize the power of these revealed preference tests. The earliest contribution in this vein was Bronars (1987), who proposes an alternative hypothesis that individual choices are randomly distributed uniformly over the choice set. In investigating the empirical content of GARP tests generally, Beatty and Crawford (2011) incorporate Bronars alternative hypothesis with Selton (1991)'s measure of predictive success

[2]We formally define Afriat's CCEI in Section 5.2, but intuitively, the measure can be thought of as the proportion of an agent's wealth that is preserved despite violations of GARP. Several alternative goodness of fit measures have been proposed in the recent literature, including Echenique, Lee, and Shum (2012)'s Money Pump and Dean and Martin (2011)'s modification of Houtman and Maks (1985)'s goodness of fit measure.

to define the Difference Power Index which is quite similar to our Optimal Placement Index. Dean and Martin (2012) extend Beatty and Crawford's approach to incorporate observed choice information by deploying a bootstrap of budget shares across budget set. In a novel application, Polisson (2012) compares the power of GARP tests over goods and aggregated features of those goods.

When measuring power for a fixed experimental design, the central challenge lies in specifying the alternative hypothesis that characterizes choice behavior. To that end, we introduce a nonparametric panel regression model in Section 3 that provides a flexible characterization of choice behavior without imposing a utility representation. Given an observed sample of choice behavior, our analysis in Section 4 then illustrates ways of estimating the distribution over choice in that regression model using nonparametric methods. We present intuitive sampling strategies to generate these distributions from observed choice behavior and, in so doing, relate these measures to prior approaches adopted by experimental researchers to measure power.

Estimating the probability of observing GARP violations provides a first step toward characterizing the power of an experimental design. Still further information is available by exploiting design features specific to tests for revealed preference to create intuitive measures of the efficiency of the experimental design. For example, we can ask how severely the observations or design would have to be perturbed in order to observe GARP violations. Exploring this sort of question in Section 5 doesn't lead to a probabilistic measure of power, but rather statistics we refer to as power "indices." We introduce three new power indices: the Jittering Index, the Afriat Power Index, and the Optimal Placement Index.

Our analysis presents a series of different approaches to measure the power of an experiment's design. How these different power measures and indices behave depends on the characteristics of choices in the population and the experimental design. For example, if behavior in the cross-section is extremely concentrated around modes (for instance, due to unobserved types that favor equity or fairness), a fixed power measure may behave differently than if that behavior is rather diffuse. Throughout the exposition, we explore these

properties of different power measures by presenting the power indices and measures from two very different experimental settings. Before concluding, we provide some suggestions for how researchers seeking to use these measures should choose the order and depth with which to explore their design's power.

2 Background on Testing Revealed Preference

A vast empirical literature uses revealed preference axioms to build new and better price indices. Mansur and McDonald (1988) examined 27 years of aggregate consumption data. By assuming preferences are homothetic, they improved the power of GARP tests and narrowed the bias in constructing exact prices indices, finding the aggregate data to be broadly consistent with both GARP and with homothetic preferences.

Other researchers have used repeated samples from cross-sectional surveys such as the Consumer Expenditure Survey and the British Family Expenditure Survey. Famulari (1995)'s analysis of the Consumer Expenditure Survey (CEX) from 1982–1985 focused on testing the common preferences assumption. Aggregating choices across representative households to increase power, she found almost all of these "groupings" satisfied GARP. Blundell, Browning and Crawford (2003) note that GARP tests may actually be quite weak when applied to annual data, since incomes expand over time and relative prices are somewhat stable. They propose adopting "flexible parametric models over regions where the nonparametric tests do not fail" to enhance the power of revealed preference tests. Using sophisticated semiparametric methods to estimate expansion paths for preferences, they then project observed choices into an optimal test setting, finding the data largely fails to reject the optimizing model while also deriving much tighter bounds on consumer price indices.

Not all empirical studies show uniform support for revealed preference axioms and numerous violations of the optimizing model have been discovered using disaggregated consumer panels. A recent study by Echenique, Lee, and Shum (2012) uses scanner data from supermarket food expenditures and find individuals' consumption decisions are often inconsistent with WARP. In order to evaluate the magnitude of these violations, they introduce a "money-

pump" measure corresponding to the profits an arbitrageur would be able to generate by exploiting these violations. By this measure, they show that the widespread violations do not impose great costs to the consumer. In another analysis, Dean and Martin (2011) propose a modification of Houtman and Maks (1985) efficiency that finds the least expensive way in which to resolve GARP violations. Using a large panel of household consumption choices, they find pervasive violations of GARP, most of which are of relatively small magnitude. Further, by focusing on substantial heterogeneity in preferences across households, Dean and Martin (2011) underscore that care must be taken in aggregating choices in cross-sectional evaluations of revealed preference.

A parallel literature has developed around controlled laboratory experiments. An important first study is by Cox (1997), who evaluates revealed preference in a field experiment using subjects who were residents at a psychiatric hospital. This hospital had a functioning "token economy" with a local currency that could only be traded for goods at a hospital store. Cox found almost all patients' consumption choices to be consistent with the revealed preference axioms, despite potential problems with some goods being storable.

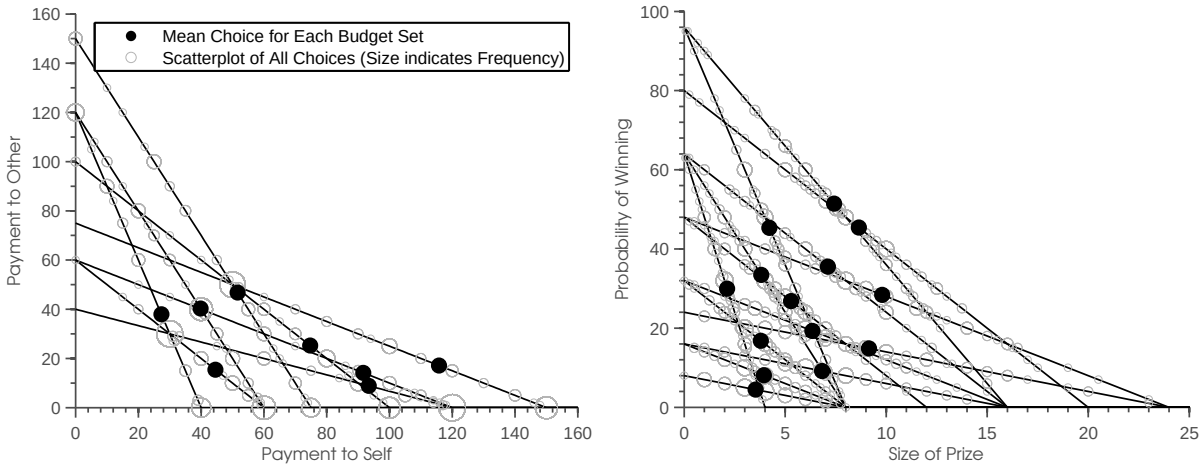
To address issues relating to choices over storable goods, Sippel (1997) provided a test with 10 budget sets over eight commodities, all of which had to be consumed over the course of the experiment. Sippel found 57% of the subjects violated GARP, although very few of these violations were severe in terms of the Afriat Efficiency Index. Fevrier and Visser (2004) used five budgets of six goods, all of which were different varieties of orange juice. They found that 30% of subjects were inconsistent with GARP, with 15% having Afriat Effeciency below 0.95.

Other experimental settings have focused on evaluating the degree to which GARP applies in different populations. Mattei (2000) used 20 budgets with 8 goods (mostly school supplies), conducting the experiment with three different populations, including 20 undergraduates, 100 graduate students, and 320 readers of a consumer affairs magazine. Applying Afriat Efficiency threshold of 0.95, he found fewer than 4% of subjects violated GARP in each of these populations. Harbaugh, Krause and Berry (2001) test the rationality of children and

young adults at different stages of development by offering them choices from budget sets over chips and juice boxes. While second graders performed relatively poorly, sixth graders and college students tended to perform equally well in choosing consumption bundles consistent with GARP. More recent work by Burghart, Glimcher, and Lazzaro (2012) explores the degree to which alcohol impairs an individual's ability to choose consistently with GARP. Surprisingly, even highly inebriated subjects' choice behavior appears to be fairly consistent with axioms of revealed preference.

We illustrate the power measures and indices proposed in the paper using data from two experiments. In the first, Andreoni and Miller (2002) evaluated the extent to which subjects' generosity in a dictator game is consistent with revealed preference. Participants chose from linear budget sets over wealth kept by the dictator and passed to their partner, revealing modal choices corresponding to monotonicity and equity with less than 10% of subjects' choices violating of GARP. The second, Andreoni and Harbaugh (2009), presents subjects with choices over gambles where they faced a linear trade off between the size of the prize and the probability of winning. Behavior in this setting is much more diffuse and, while choices over gains are largely consistent with revealed preference, choices over losses have many more violations of GARP.

The protocols for these two studies are discussed in more detail in Appendix A1. Figure 1's left panel shows the choice sets offered in Andreoni and Miller (2002), as well as a scatterplot of the actual choices and the average allocation chosen on each budget set. The right panel shows a similar perspective of the Andreoni and Harbaugh (2009) experimental treatment. Our objective in the paper is to answer the question of which experiment provided a more powerful test of GARP. The study on risk preferences includes more budget sets and is characterized by more diffuse choices, so we expect to (and do) find more GARP violations in that sample. However, choice behavior in the rational altruism experiment is concentrated around the intersection of budget sets, so the design of that experiment can be considered more efficient even if it reveales fewer ex post violations of GARP in the sample.



(a) Budgets for Testing Altruistic Preferences Andreoni and Miller (2002)

(b) Budgets for Testing Risk Preferences Andreoni and Harbaugh (2009)

Figure 1: Budget Sets and Distribution over Choices from Experimental Data
This figure presents the budget sets and distribution over observed choices in the two experimental treatments we use to illustrate the properties of our ex post power measures and indices. The solid lines present budget sets, the solid black dots represent the mean choice on each budget set, and the light circles illustrate the distribution over choices along each set. Larger circles indicate bundles that were chosen with relatively high frequency.

3 Power and Experimental Design

In this section, we illustrate the tight link between the pattern of choice behavior and the structure of the experiment's design. We define the power of an experiment, conditional on the experiment's design, and develop a nonparametric characterization for the distribution over choices.

3.1 Characterizing the Distribution over Choice

We begin by presenting a probability model for studies evaluating the rationality of observed choice under differing budget sets. Using this context, we will be able to discuss alternative specifications for the data generating process that are not necessarily consistent with rational choice and how to easily sample from these alternative hypotheses.

Assume the econometrician observes a panel of N individuals choosing allocations among K goods and represent each of individual i's choices of consumption bundles on $t = 1, \ldots, T$

budget sets by the vector $x_{i,t}$. The budget sets are denoted $B_1, \ldots, B_T$ with B_t defined by the price vector p_t and income m_t. Denoting individual i's choice function by r_i and using ε as an error term, we can represent the data generating process using a nonparametric panel regression model:

$$x_{i,t} = r_i(B_t) + \varepsilon_{i,t}, \;\; E[\varepsilon_{i,t} | r_i(B_t)] = 0 \tag{1}$$

Note that there are two underlying sources of randomness in the observed consumption decisions. Across subjects, variation in choices arises from the individual's "true" choice function, r_i, that could result from an underlying random utility model whose error terms are fixed across all budget sets. The second source of randomness is generated by the term $\varepsilon_{i,t}$, which reflects noise in the individual's observed choices, whether due to measurement error, optimization errors, or time-variation in preferences. The realized choices for all individuals in the population forms the outcome for which our probability model is defined. The sampling of individuals and choices induces a population joint distribution for $X_i \equiv \{x_{i,1}, \ldots, x_{i,T}\}$ conditional on the budget sets included in the experimental design, $B = \{B_1, \ldots, B_T\}$. We denote this measure P^* and use it to characterize the power properties of a given test.[3]

3.2 Power under Different Designs

Our power measures explicitly state the probability that the null hypothesis will be rejected when observing the choice behavior of a single subject from the population *for a specified experimental design*. Fixing the budget sets included in the experiment, the null hypothesis here is simply:

$$H_0\text{: } P^*\{X_i(B), \text{ such that } X_i(B) \text{ violates GARP}\} = 0 \tag{2}$$

[3]Additional regularity and exogeneity conditions are needed to ensure identification of the model. Such an exercise is not in the scope of the current work but would be necessary for projecting choices onto unobserved budget sets. An alternative would be to specify a random utility model where an individual receives a preference shock at every budget set. Further, as a referee pointed out to us, a line of research has explored set-valued extrapolation, including Manski (2007, forthcoming) as well as Blundell, Browning, & Crawford (2003). Current research by Blundell, Kristensen, & Matzkin (2013) and Kitamura & Stoye (2013) considers the problem of directly extrapolating choice behavior without complete identification. Our purpose in adopting the regression framework is to frame our analysis in a setting similar to the framework established by Epstein and Yatchew (1985) to study testing in nonparametric models.

This specification of the null hypothesis is rather flexible, in that it allows us to adopt alternative characterizations for a choice profile that "violates GARP." This flexibility will be useful in allowing for different thresholds, perhaps in terms of Afriat's CCEI, to satisfy goodness of fit. While the null hypothesis can satisfy different thresholds of goodness of fit, it applies for any experimental design.[4]

Unfortunately, feasibility of implementation prevents the experimenter from completely spanning the set of all possible budget sets. Consequently, the probability of rejecting the null hypothesis depends both on the data generating process and the experimental design itself. Note that, while conceptually similar notions, it is important to distinguish this definition of power as a feature of the design of an experiment from the classical definition of power characterizing the properties of a statistical test. In particular, we are interested in comparing the power of *potentially different* experimental designs. For a given choice setting, some experimental designs may be more likely to reveal violations of GARP than others. Here, we use the regression model to characterize choice behavior in the experiment under these different designs.

The power of a statistical test states the probability of rejecting the null hypothesis based on a the distribution for a summary statistic for a single experiment when the data is generated according to an alternative specification inconsistent with the null. Evaluating the power of a statistical test requires a probability model, whether or not it satisfies the null hypothesis, for the sampling error of the test statistic.

The power of the experiment states probability of rejecting a behavioral model based on observed behavior within a fixed design. Here the sampling error comes from selection within the population as well as potentially random individual behavior. In order to evaluate experimental power, we need to specify a probability model for that population selection process and the randomness of individual behavior.

The goal of the current exercise is to evaluate the power of a fixed design specification

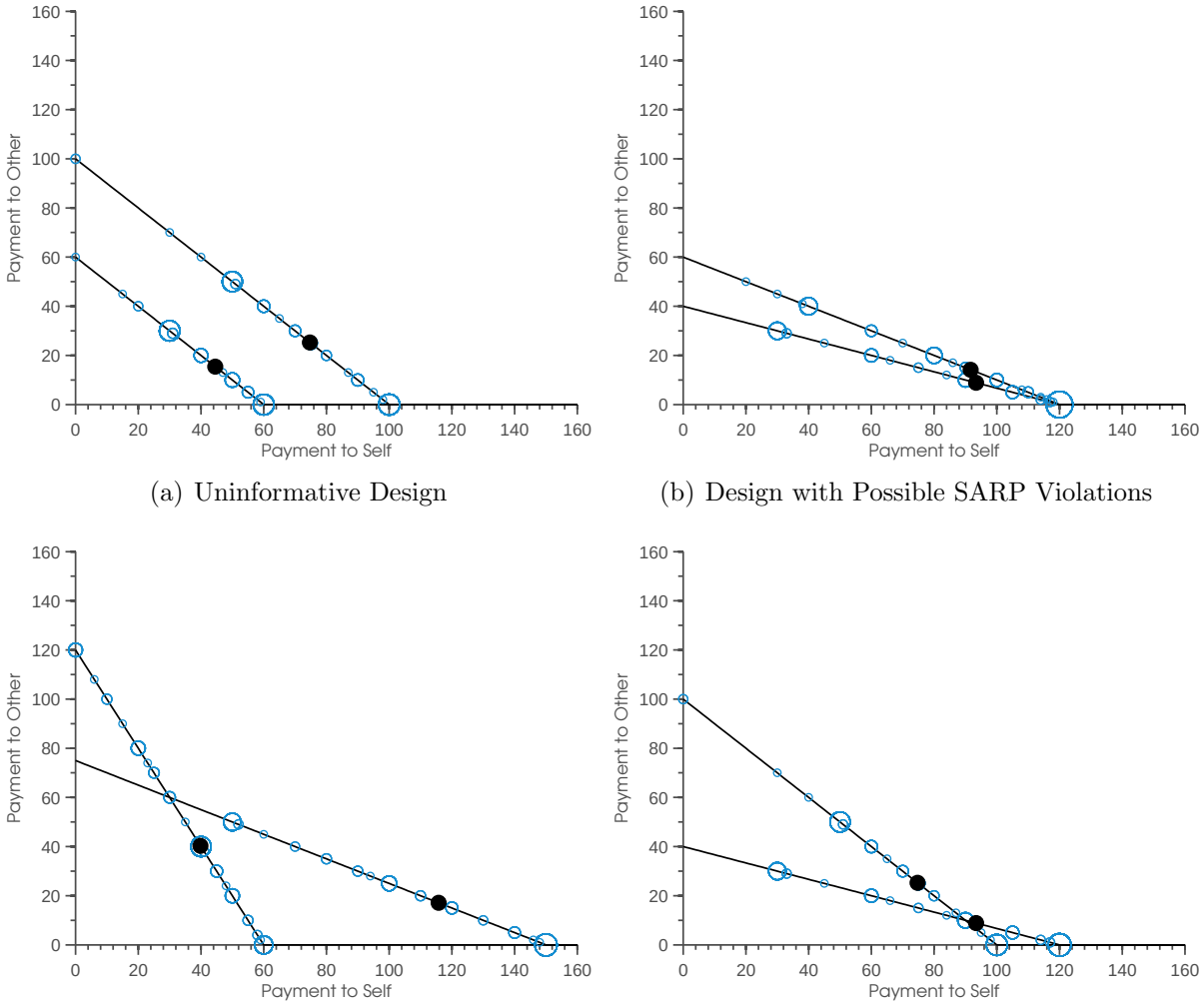
[4]Many natural specifications for the distribution of the error component in the generative choice model, such as the normal error considered in jittering measures below, would lead to observed choice profiles that violate GARP. However, the frequency with which such violations are observed will still depend on the budget sets presented in the experiment.

conditional on the distribution of choices observed in the experiment. This objective corresponds to a counterfactual estimation problem of characterizing the probability that we would observe a violation in other samples or other designs based on what we observe about behavior in a given sample for a fixed design. Under the null hypothesis that every individual of the population makes choices consistent with GARP, any experiment or testing strategy would have zero power by definition. Instead, our goal is to specify the distribution of choices using the ex-post observed choice data. We can then use the structure implied by that distribution over choices to characterize the counterfactual power under different designs.

To illustrate the role of design in experimental power, suppose that, rather than using the full set of budget choices presented in the Andreoni and Miller (2002) experiment presented in Figure 1, the experimenter could include only two budget sets. Figure 2 presents four possible designs. In the design presented by Panel (a), none of the budget sets cross one another, and, as such, the design has no power to generate violations of GARP. That is, no matter what choices a subject makes in Experiment (a), the experimenter will not observe a violation of GARP. In Panel (b), the budget sets meet only at their corners, making GARP violations possible but exceedingly unlikely. The budget sets in Panel (c) cross in the middle of the choice plane, but choices are concentrated away from the crossing. The budget sets in Panel (d) cross away from the mid point of the budget sets, but near where many choices are observed.

3.3 Power Measures and Power Indices

As evidenced by the examples in Figure 2, the placement of budget sets and the distribution over choices on those budget sets jointly determine the power of the experiment. Conditioning on the budget sets, the power of an experiment's design is entirely determined by the distribution over observed choices. The inference objective here is to characterize this distribution over choices to evaluate ex post the degree of probability of rejecting revealed preference. Since distributional estimation typically involves some form of smoothing the



(a) Uninformative Design

(b) Design with Possible SARP Violations

(c) Budget Sets Cross Away From Most Choices

(d) Budget Sets Cross Near Many Choices

Figure 2: Experimental Design and Ability to Detect GARP Violations

Panels (a) through (d) present hypothetical experiments in which subjects faced only two budget sets. The probability of these hypothetical experiments detecting violations of GARP in the population depends both on the point at which the budget sets cross and the distribution of choices in the population.

observed sample distribution, different assumptions on that smoothness will generate different distributions over choice and different measures of power. Defining the alternative hypothesis distribution over choices as a function of the smoothing parameter, we refer to such direct probabilities of revealed preference violations as power measures.

Power indices differ from direct power measures by evaluating ex post how the data generating process must be perturbed to generate violations of the null hypothesis. For

example, we may wish to find the "smallest" change that would be necessary for a set of observed choices, which are consistent with the restrictions of revealed preference, to generate violations. Here, indices will vary in terms of how one defines the distance used to characterize "smallest." While they are not probabilistic statements, power indices provide ordinal measures of power by exploiting features of the model that characterize the degree of rigor in the test.

4 Power Measures

In this section, we characterize different nonparametric approaches to estimating the ex post distribution of choices implied by model 1 for use as the alternative hypothesis in measuring experimental power. Perhaps the simplest formulation of this alternative hypothesis is to adopt the sample distribution of choices observed from the $N \subset \mathcal{I}$ subjects in the experiment, which we'll denote P_N. We can then define:

$$H_{\text{A, Sample}}\colon\ r_i\left(B_t\right) = x_{it};\ \varepsilon_{it} = 0 \tag{3}$$

The key limitation of a purely sample-based measure is that it may be overfit in finite samples, understating the true variation in choice behavior, and is difficult to extend to choices outside the observed budget sets, limiting its application in counterfactual analysis. To this end, we propose nonparametric strategies for smoothing the observed measure P_N along with simple sampling algorithms that facilitate calculating the power properties under these smoothed measures.

The ex post results for observed choices from the two experimental samples are displayed in the bottom row of each panel for Table 1 along with the Bronars power measures that we present in the next subsection. The rational altruism study uncovered a small frequency of violations with 9% of subjects selecting choice profiles inconsistent with GARP and only 2% of subjects choices implying an Afriat Critical Cost Efficiency Index (CCEI) less than 0.95 and an average CCEI of 0.998. The study evaluating risk preferences over gains reveals a higher frequency of violations, with 44% of the subjects including at least one violation and

Table 1: Sample Ex Post Results and Bronars' Power Measures

Panel A: Altruistic Preferences

	Budget Share	Violation	CCEI	CCEI	Frequency of CCEI <					
	Avg St Dev	Frequency	Average	St Dev	0.50	0.75	0.90	0.95	0.99	1.00
Bronars M1	0.289	75%	0.88	0.122	0%	16%	46%	59%	69%	72%
Bronars M2	0.238	59%	0.93	0.095	0%	7%	27%	40%	52%	55%
Bronars M3	0.220	44%	0.95	0.082	0%	4%	18%	27%	37%	40%
Sample	0.278	9%	1.00	0.017	0%	0%	1%	2%	3%	3%

Panel B: Risk Preferences over Gains

	Budget Share	Violation	CCEI	CCEI	Frequency of CCEI <					
	Avg St Dev	Frequency	Average	St Dev	0.50	0.75	0.90	0.95	0.99	1.00
Bronars M1	0.289	91%	0.82	0.133	1%	27%	68%	80%	88%	90%
Bronars M2	0.238	80%	0.89	0.112	0%	11%	44%	60%	74%	78%
Bronars M3	0.237	86%	0.87	0.120	1%	15%	52%	68%	81%	84%
Sample	0.185	44%	0.98	0.057	0%	2%	7%	14%	25%	27%

This table reports the Sample and Bronars Power Measures for choices observed in the experimental studies by Andreoni and Miller (2002) and Andreoni and Harbaugh (2009). The cross-sectional standard deviation of budget shares is averaged across budgets and the Violation Frequency reports the frequency with which a subjects' choice profile violates GARP. The table also presents the cross-sectional average, standard deviation, and quantiles for the distribution of the Afriat Critical Cost Efficiency Index (CCEI).

14% of subjects' CCEI's falling below 0.95. Interestingly, while there was more variation in the realized CCEI values for individuals in this experiment, these were generated with a smaller average variance in the budget shares.

4.1 Bronars' Power Measures

Bronars (1987) developed the first and most lasting index for the power of revealed preference tests, specifying an alternative hypothesis based on Becker's (1962) notion that individual choices are made at random and uniformly distributed on the frontier of the budget set. In a data generating model consistent with this behavior, the alternative hypothesis can be stated as:

$$H_{\text{A, Bronars M1}}\colon\ r_{i(k)}\left(B_t\right) = \frac{1}{K}\frac{m_t}{p_{t(k)}}, k = 1, \ldots, K;\ \varepsilon_{i,t} \sim \mathcal{U}\left(B_t\right) \tag{4}$$

where $\mathcal{U}\left(B_t\right)$ denotes the uniform distribution over the frontier budget set B_t recentered at zero.

With this alternative, one can analytically calculate the exact probability that a random set of choices will violate GARP. Perhaps more sensibly, one can conduct a series of Monte

Carlo experiments on the budgets under the alternative hypothesis and calculate the probabilities of GARP violations. Then the power of a particular GARP test is the chance that random choices will violate GARP. Bronars calls this Method 1.[5] Bronars also considered two modification of Method 1. His Method 2 first derives random budget shares in which the expected share is $1/n$, where n is the number of goods. Method 3 finds random budget shares in which the randomness is centered on actual budget shares. Method 1, however, has come to dominate the literature.[6]

The three Bronars' power measures are presented in Table 1 for the altruism and risk preference experimental designs. Bronars' Method 1 provides the least structured behavioral model and, as such, imparts the highest power to each of the experiments. Under this measure, approximately 75% of the choice samples for altruistic preferences included at least one violation of GARP, generating an average CCEI of 0.88, with 59% of the samples generating a CCEI less than 0.95. The Bronars' power measures rate the risk preference study somewhat higher, mainly due to the larger number of budget sets available, with 91% of the samples including at least one GARP violation and 80% of samples generating a CCEI less than 0.95 for an average CCEI of 0.82. Bronars' Methods 2 and 3 impart more structure on the data and, in doing so, yield weaker power properties. Still, across the two experimental designs, all of the Bronars' measures impart a higher power to the risk preference study.

An advantage of Bronars' approach is that it is both natural and simple, motivated by the representation of the alternative hypothesis as a minimally informative prior in the Bayesian sense. A disadvantage is that the alternative hypothesis is perhaps too unconditional and takes no advantage of the information in observed choices about the distribution

[5]One should also note the paper by Aizcorbe (1991) that argued that using Bronars' method to search for WARP violations in all pairs of observations may misstate power in that violations over pairs is not independent (comparing bundle a vs. b is not independent of the comparison of b vs. c). She then suggests a lower bound estimate of power based on independent sets of comparisons. We propose an alternative method for addressing this dependence using a weighted bootstrap algorithm below.

[6]Famulari (1995) and Cox (1997) offered variants of the Bronars method in which observed prices and quantities were randomly paired, and these pseudo-random budget choices are tested for GARP violations. As these are not formal measures of power, we do not discuss them further here, but mention them for completeness. Both approaches involve sampling from both observed budget sets as well as observed choices in a manner that projects choices from one budget set onto another. While a helpful technique for addressing settings with stochastic budget sets, our focus on a fixed sample of budget sets allows us to avoid such issues.

over behavior. [7] Suppose, for instance, the budgets offered did not intersect near the points where individuals are actually choosing. Then if preferences do not conform to utility maximization, the test would be unlikely to discover it. This is true even if Bronars' analysis shows that randomly made choices provide a high probability of violations. Dean and Martin (2012) present a similar critique in a comment on Beatty and Crawford (2011)'s difference power index, proposing a bootstrap technique loosely related to the approaches we propose in the next section. What would be preferred, though, is an index of power based on an alternative that takes account of the choices exhibited.

4.2 Bootstrapped Power Measures

In the ex post setting where we have choice data from a panel of subjects, we can use the multiple observations to get additional information about the distribution over choices that will actually be made within these budget sets. In particular, we can ask whether the organization put on the data by the subjects themselves—by matching individuals with choices—is superior to another method that would have randomly assigned choices to individuals from the universe of choices actually made.

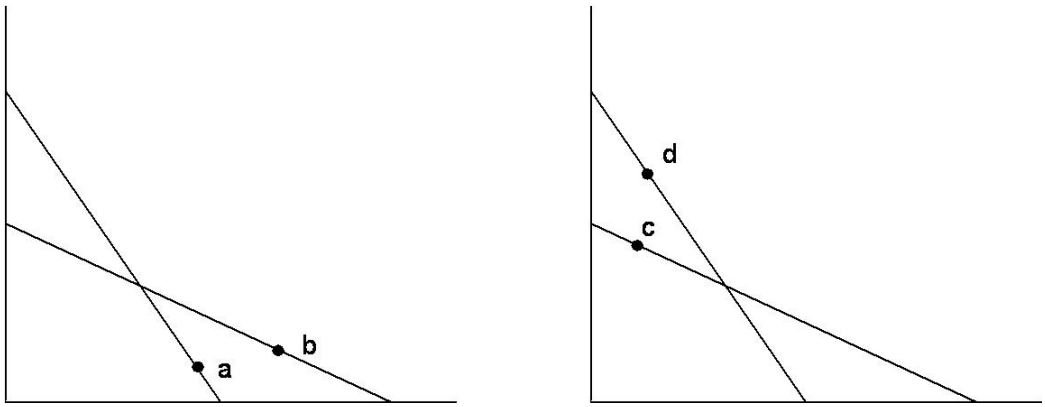


Figure 3: Individual Choices without GARP Violations

For simplicity, consider an example of two experimental subjects given the same two budgets. Suppose the data are like that shown in Figure 3. Here there are no violations of

[7] Note, however, that Bronars' Method 3 does take into consideration the average choices observed in the population at each budget set.

revealed preference. Suppose that, on each budget, we were to pool the choices made by the subjects and then create new synthetic subjects by randomly drawing from the universe of choices actually made. That is, we use bootstrapping techniques to generate a measure of power. In the example of Figure 3, $x_1 \in \{a, d\}$, and $x_2 \in \{b, c\}$. Then there would be a 25% chance that the synthetic subject would be assigned choices a and c, hence violating GARP, which is the maximum probability possible with two budgets and no initial violations of revealed preferences.

Compare these choices to those in Figure 4. Here there would be no chance that we could create a synthetic subject that would violate GARP. In this sense, the test has more power if the study generates data like that in Figure 3 rather than Figure 4.

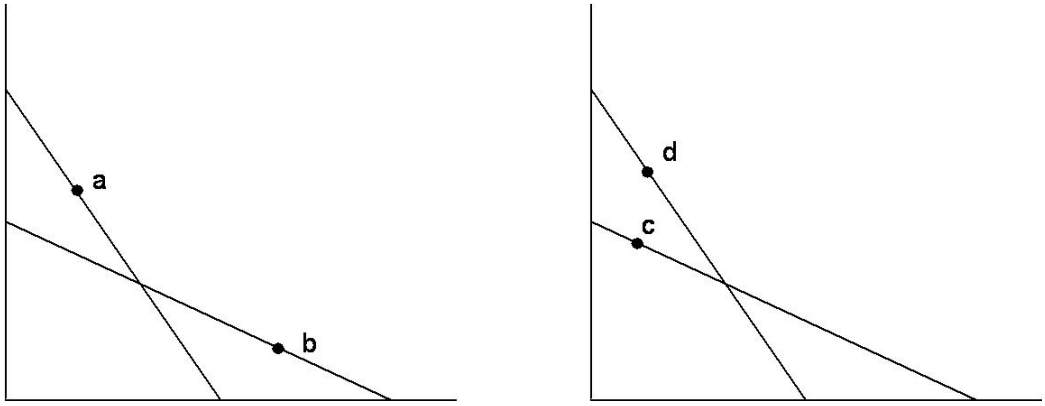


Figure 4: Individual Choices without GARP Violations and Zero Bootstrap Power

Note that this technique can reveal either greater or lesser power than a simple Bronars methods. For instance, in the budgets shown in Figures 3 and 4 a Bronars (Monte Carlo) test would show only about 12% of the cases finding violations, whereas the bootstrapping test will get exactly 25% violations (Figure 3) or 0% violations (Figure 4).

This algorithm maps to an alternative hypothesis that choices are drawn independently across budgets from the empirical marginal distribution of choices on each budget. As such, the bootstrapped alternative hypothesis maintains the Bronars' Method 3 hypothesis that the function $r_i(\cdot)$ is a constant equal to the cross-sectional average consumption bundle chosen on the budget set. However, instead of $\varepsilon_{i,t}$ being uniformly distributed over the

zero-centered budget set, as in Bronars' measures, here $\varepsilon_{i,t}$'s distribution gives rise to the empirical distribution of observed choices on the budget set (which, we recall, is denoted by P). That is:

$$H_{\text{Bootstrap}}\colon\ r_i\left(B_t\right) = \bar{x}_t \equiv \frac{1}{N}\sum_{i=1}^{N} x_{i,t}, \qquad (5)$$
$$\tilde{P}\left(\varepsilon_{i,t} = x_{j,t} - \bar{x}_t\right) = P\left(x_{j,t}\right)$$

With this alternative, the probability of violations among the synthetic subjects is the power of the test.[8]

4.3 Weighted Bootstrap: Sampling from the Conditional Distribution over Choices

The main strength of the unconditional bootstrap as compared to Bronars' method lies in its ability to measure how the test's design was suited to the population studied. Like the Bronars method, however, the alternative hypothesis specified is still subject to the Aizcorbe (1991) critique of ignoring dependence in choices across budget sets and consequently ascribing too much randomness to each subject, especially in populations with heterogeneous preferences.

We can address this shortcoming by exploiting continuity in preferences to group the behavior of different subjects. Consider the example from the previous section but, instead of there being two subjects, there are four subjects who make the choices depicted in Figure 5. Visually, it appears the preferences for subjects A and B and for subjects C and D are similar to one another, but the two sets of subjects appear to have very different preferences. Under the bootstrapped power measure, after drawing an observation of $x_{A,1}$ on budget set 1, we'd be equally likely to impute the selection of $x_{A,2}$, $x_{B,2}$, $x_{C,2}$, and $x_{D,2}$ on budget set

[8]This bootstrapping technique was introduced by Andreoni and Miller (2002) and applied by Harbaugh, Krause, and Berry (2001). Dean and Martin (2012) adopt a bootstrapping strategy where they sample from budget shares across budget sets, effectively implying that the distribution of budget shares on unobserved budget sets is equivalent to the unconditional distribution of chosen budget shares. The Dean and Martin (2012) approach is particularly helpful when projecting observed choices onto budget sets that are not observed in the experiment.

2 as the anticipated decision for that type. However, given the obvious pattern in choices, it seems relatively unlikely that an indvidual who selects $x_{A,1}$ really would select $x_{C,2}$ or $x_{D,2}$ and much more likely that she would select either $x_{A,2}$ or $x_{B,2}$.

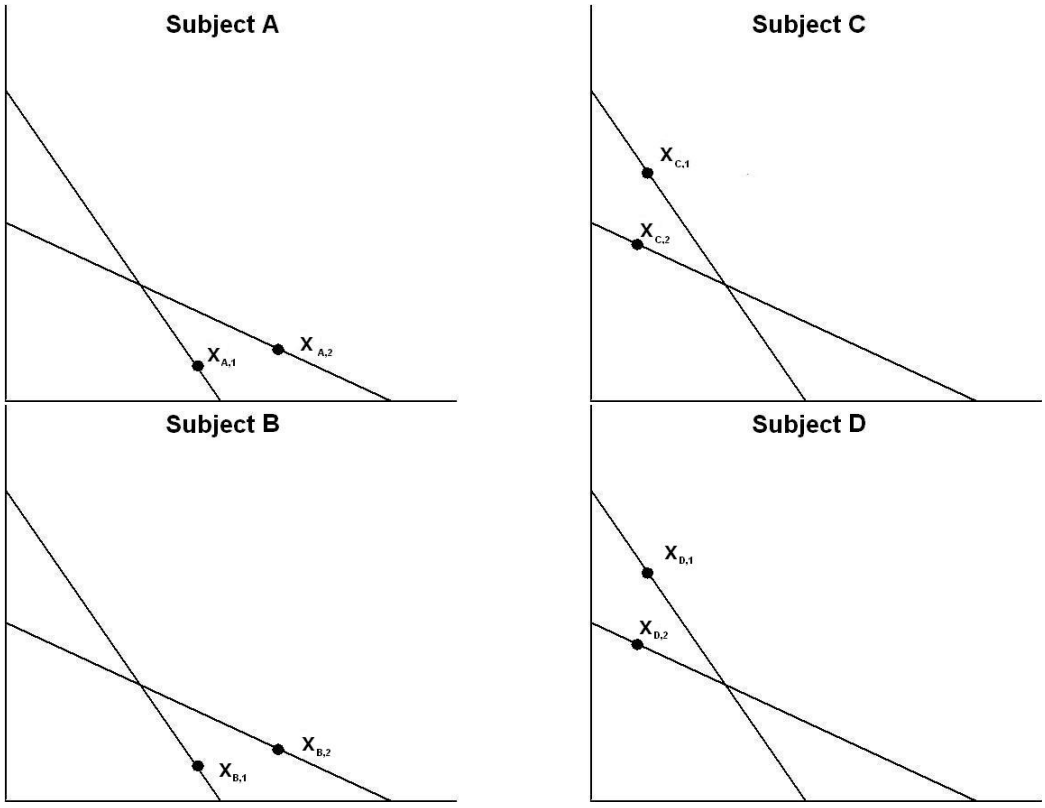


Figure 5: Dependence in Choices Across Budget Sets
When individual choices indicate dependence across budget sets, an unconditional bootstrap can overstate the degree of variation in the data. In contrast, the weighted bootstrap samples from choices in a way that preserves this dependence structure. In this example, individuals tend to prefer one of the two goods, but are unlikely to choose a bundle concentrated in good 1 on one budget and in good 2 on another budget.

For these reasons, we may wish to account for this dependence structure of the choices for an individual across budget sets in the bootstrap. Specifically, for budget set B_1, we want to identify the conditional distribution of selected consumption bundles for subject i given the observed choices subject i has made on the budget sets $2, \ldots, T$. That is, we want to identify the distribution of $x_{i,1}|x_{i,2}, \ldots, x_{i,T}$ and use that distribution to characterize the probability of a WARP violation along the B_1 budget set. Further, denoting by $x_{i,-t}$ the array of observed consumption bundles $x_{i,1}, \ldots, x_{i,t-1}, x_{i,t+1}, \ldots, x_{i,T}$, we can iteratively

identify each of the T_i marginal conditional distributions for subject i. We can then use these marginal distributions to sample from the set of budget choices conditional on drawing subject i from the population, using the frequency of GARP violations in these draws to measure the power of the test for each individual in the study.

4.3.1 Sampling for an Individual Budget Set

To link choices on budget set B_τ with the choices observed on other budget sets, we add additional structure to the random function $r_i(B_\tau)$. In particular, letting $\eta_{i,t}$ be a mean-zero error term, assume

$$r_i(B_t)\,|x_{i,-t} = g_t(x_{i,-t}) + \eta_{i,t} \tag{6}$$

Importantly, the function g does not depend on the actual individual i, but only the choices made by individual i on other budget sets. Then, subject to exogeneity conditions on $\eta_{i,t}$, we could use nonparametric regression in the cross-section to estimate the g function and characterize the distribution for $r_i(B_t)\,|X_{i,-t}$. Adapting this estimation strategy, however, is not necessary when a weighted bootstrap allows us to sample directly from this distribution.

Define the weighting function $w(x_{i,-t}, x_{j,-t})$ so that, given the sample $x_1, \ldots, x_N$ and a choice profile on the $-t$ budget sets $x_{i,-t}$ we assign as the choice profile on the t budget set the choice $x_{j,t}$ with probability $\frac{w(x_{i,-t},x_{j,-t})}{\sum_{n=1}^{N} w(x_{i,-t},x_{n,-t})}$. The weighting function is analogous to the kernel in nonparametric regression, smoothing out variation across the population, motivating a Gaussian weighting function. Denoting the normal p.d.f. by $\mathcal{N}$, the sample covariance matrix for choices on the budget sets other than budget set t by Σ_{-t}, and a bandwidth parameter by h, we propose the weighting function:

$$w(x_{i,-t}, x_{j,-t}) \propto \mathcal{N}\left(x_{i,-t} - x_{j,-t}, h^2\Sigma_{-t}\right) \tag{7}$$

The weighted bootstrap nests the unconditional bootstrap as h becomes large and each observation is drawn with equal probability. Such a bandwidth would imply individual preferences on a given budget set are not well-characterized by their choices on other budget sets and decision is driven by idiosyncrasies at the individual and budget set level. As h becomes small, the distribution implied by the weighted bootstrap becomes very close to the

sample distribution over choices.[9] The weighted bootstrap implies an alternate hypothesis measure over choices $\tilde{P}$ as:

$$H_{\text{A, WBS}}(h):\ r_i(B_t) = \sum_{j=1}^{N} w(x_{i,-t}, x_{j,-t})\, x_{j,t} \equiv \bar{x}_{i,t}; \tag{8}$$
$$\tilde{P}(\varepsilon_{i,t} = x_{j,t} - \bar{x}_{i,t}) \propto w(x_{i,-t}, x_{j,-t})\, P(x_{j,t})$$

4.3.2 Strengths and Weaknesses of the Weighted Bootstrap

The weighted bootstrap generalizes the unconditional bootstrap method from Andreoni and Miller (2002) for characterizing the distribution over choices. The main benefit to doing so comes by exploiting the cross-section to account for dependence across an individual's choices on multiple budget sets. In this regards, the weighted bootstrap effectively addresses the Aizcorbe (1991) critique of Bronars' and similar power measures, such as the unconditional bootstrap, that ignore this dependence.

The weighted bootstrap also decomposes variation in choices as arising from between subject and within subject variation, sampling from the conditional distribution of $r_i(B_t) + \varepsilon_{i,t}$ given that individual i has been drawn from the population. As such, a standard law of large numbers implies the bootstrap sample average converges to $r_i(B_t)$ and the bootstrap sample variance converges to $Var[\varepsilon_{i,t}]$. Aggregating across individuals, we can take advantage of the exogeneity of errors to decompose the unconditional variance of the choice for a randomly selected individual at budget set t:

$$Var[x_{i,t}] = Var[r_i(B_t)] + Var[\varepsilon_{i,t}] \tag{9}$$

Denoting the total number of bootstrap samples drawn by M, it is straightforward to show the following convergence results hold *within* subjects:

$$\bar{x}_{i,t} \equiv \frac{1}{M}\sum_{m=1}^{M} x_{i,t}^{(m)} \xrightarrow[M,N\to\infty]{} r_i(B_t), \text{ and, } \frac{1}{M}\sum_{m=1}^{M}\left(x_{i,t}^{(m)} - \bar{x}_{i,t}\right)^2 \xrightarrow[M,N\to\infty]{} Var[\varepsilon_{i,t}]$$

[9] In this limit, the only within subject sampling variation in choice for an individual would come from other subjects whose choices matched that individual along each of the $-t$ budget sets but differed at the t-th choice set. In the example above, if $X_{C,2} = X_{D,2}$ but $X_{C,1} \neq X_{D,1}$, then the weighted bootstrap for subject C would draw both $X_{C,1}$ and $X_{D,1}$ with equal probability for the choices on budget set B_1. While minor, aggregating this variation across budget sets can be sufficient to generate choice profiles that do not appear in the sample.

With the following convergence results also holding *between* subjects:

$$\bar{x}_t \equiv \frac{1}{N}\sum_{i=1}^{N} x_{i,t} \xrightarrow[N\to\infty]{} E\left[r_i\left(B_t\right)\right], \text{ and, } \frac{1}{N}\sum_{n=1}^{N}\left(\bar{x}_{i,t}-\bar{x}_t\right)^2 \xrightarrow[N\to\infty]{} Var\left[r_i\left(B_t\right)\right]$$

As in most nonparametric analysis, a central indeterminacy in the weighted bootstrap arises from the need to specify the weighting function and bandwidth. Unfortunately, there is no clear definition of an "optimal" bandwidth in this setting, as such a bandwidth would depend on the true conditional dependence among choices. If the bandwidth is too small, the weighted bootstrap would understate the degree of variation in the data, essentially considering the power of the test to be the frequency of violations in the observed data. Too high of a bandwidth overstates the degree of variation in the data, exaggerating the power for the test. As such, we propose evaluating the weighted bootstrap in terms of its power function determined by the bandwidth.

4.3.3 Empirical Properties of the Unconditional and Weighted Bootstrap

For the altruism and risk preferences experiments, we generate one million samples for each budget using the unconditional bootstrap and 100,000 draws for each subject using the weighted conditional bootstrap for a variety of bandwidths ranging from $h = 0.1$ to $h = 10$. The results summarizing the power properties of the GARP tests using the bootstrapped sample appear in Table 2.

Comparing the two experiments, the unconditional bootstrap illustrates similar power properties for both, the only difference being that we'd expect slightly more severe violations of GARP to be observed in the risk preferences study. Violations occur in approximately 75% of the samples with about 43% (23%) of the risk preference (altruism) samples generating a CCEI of less than 0.95.

As expected for the highest bandwidths, the weighted bootstrap gives almost the exact same results as the unweighted bootstrap. As we decrease the bandwidth, the frequency of violations and the distribution over CCEI's drops from that implied by the unconditional bootstrap to that observed in the sample. Further, as the bandwidth is tightened, the amount of variability in choices and CCEI attributed to variation between individuals increases

Table 2: Bootstrapped Power Measures

	Violation Frequency	CCEI Average	Frequency of CCEI < 0.75	0.90	0.95	1.00	CCEI Variance Analysis St Dev	Within Subject	Between Subjects
Panel A: Altruistic Preferences									
Simple Bootstrap	77%	0.93	6%	29%	33%	34%	0.109	100%	0%
Weighted Bootstrap									
$h = 10$	73%	0.94	5%	24%	29%	30%	0.102	100%	0%
$h = 5$	64%	0.96	3%	17%	22%	22%	0.087	98%	2%
$h = 1$	36%	0.99	0%	3%	6%	6%	0.034	78%	22%
$h = 0.5$	27%	1.00	0%	1%	3%	3%	0.022	59%	41%
$h = 0.1$	15%	1.00	0%	1%	2%	2%	0.017	18%	82%
Sample	9%	1.00	0%	1%	2%	3%	0.017	0%	100%
Panel B: Risk Preferences over Gains									
Simple Bootstrap	75%	0.93	6%	29%	45%	63%	0.094	100%	0%
Weighted Bootstrap									
$h = 10$	66%	0.95	3%	16%	30%	51%	0.072	99%	1%
$h = 5$	61%	0.96	2%	12%	25%	44%	0.063	95%	5%
$h = 1$	49%	0.97	2%	8%	15%	32%	0.058	14%	86%
$h = 0.5$	45%	0.98	2%	7%	14%	28%	0.057	1%	99%
$h = 0.1$	44%	0.98	2%	7%	14%	27%	0.057	0%	100%
Sample	44%	0.98	2%	7%	14%	27%	0.057	0%	100%

This table reports properties of the Bootstrap Power Measures for choices observed in the experimental studies by Andreoni and Miller (2002) and Andreoni and Harbaugh (2009). The Simple Bootstrap corresponds to the alternative hypothesis presented in equation 5 while the Weighted Bootstrap corresponds to the alternative hypothesis in equation 8 for varying bandwidths (h). The cross-sectional standard deviation of budget shares is averaged across budgets and the Violation Frequency reports the frequency with which a subjects' choice profile violates GARP. The table presents the distributional properties for the Afriat Critical Cost Efficiency Index (CCEI), including the percentage of the variance in the CCEI that is due to within- and between-subject variation in choices based on the decomposition in equation 9.

as each individual's choice samples become less variable, with the residual variation being attributed to between subject variation in preferences.

4.4 Jittering Measure: Sampling from the Smoothed Probability Distribution

As the bandwidth of the weighted bootstrap goes to zero, the weighted bootstrap sampling algorithm gives very similar results to the sample measure P_N. However, the collection of atoms representing the empirical distribution P_N will overfit the true measure P^* in the same way that a zero-bandwidth nonparametric kernel density estimator for a sample

overfits the distribution generating that sample. To address this overfit, kernel density estimators smooth the sample distribution, taking a weighted average of the frequency of "nearby" observations to represent the probability of a given observation. Denoting the kernel weighting function (for example, the standard normal p.d.f.) by $\kappa(\cdot)$, the kernel density estimate for the probability that a given choice profile will be drawn from P^* is:

$$\tilde{P}(x_i = x) = \frac{1}{Nh} \sum_{j=1}^{N} \kappa\left(\frac{x_j - x}{h}\right) \tag{10}$$

Analogous to the weighted bootstrap, we could compute the kernel density estimator for $\tilde{P}$ above and use that distribution to characterize the probability of observing a sample that violates GARP. However, an easier calculation is to implement a sampling strategy that "jitters" the data by introducing white noise to the sample. Let ν_m be a T-dimensional vector of independently drawn standard normally distributed noise terms, or jitters, we generate a large number, say M, such jitters and project each of the t noise terms onto their respective budget lines. We then create the jittered sample for subject i by adding this noise to their observed choices. We repeat this process for each of the N subjects in the sample, so that the probability of drawing a given observation in the synthetic sample is identical to that given in equation 10.[10] The frequency of GARP violations observed in this synthetic sample can then characterize the probability of observing GARP violations under the measure P^*. With the sampling interpretation in hand, we can express the jittering alternative hypothesis as:

$$H_{\text{A, Jitter}}(h) : \; r_i(B_t) = x_{i,t}; \; \varepsilon_{i,t} \sim \mathcal{N}_{i,t}(0, h) \tag{11}$$

where $\mathcal{N}_{i,t}(0, h)$ is the truncated normal distribution that ensures $x_{i,t} + \varepsilon_{i,t} \in B_t$.

Using jittering to make inferences about P^* requires addressing two challenges: irregularities in the distribution P^* itself that may require boundary corrections and selecting the

[10] This sampling strategy is equivalent to the smoothed bootstrap, so the distributional equivalence arises from standard results (for example, in Efron and Tibishirani, 1993). Here, we are simply enforcing the uniform frequency of bootstrap draws by holding M constant across each individual in the sample. Note that, while the kernel density estimation strategy could be generalized to project choices onto budget sets not included in the experiment's design, the sampling strategy can only be directly implemented on budget sets for which a cross-section of choices are observed.

kernel and bandwidth parameter for implementation. A priori, one could expect atoms to exist in the distribution P^*, particularly at the corner solutions. The smoothed sample may overstate the degree of variation in choices around such focal points. Also, the constrained support of the budget set requires censoring or truncating the distribution for the jitters to restrict the sample to the support of the budget set. Censoring would tend to overstate the frequency of observed corner solutions, though truncating is known to bias the frequency of corner solutions downward. In our implementation, we consider truncated errors, as these are most consistent with the density estimation strategy described above, although both approaches are feasible.

As in the weighted bootstrap measure, there is room for debate about the choice of kernel, as different kernels will undoubtedly imply slightly different population properties, but the normal kernel is well-suited and widely adopted in the literature. The bandwidth parameter h, however, will have a much more substantial impact on the estimated distribution over choices. For arbitrarily small values of h, the jittered distribution $\tilde{P}$ will converge to the sample distribution P. For extremely large values of h, the jittered distribution will converge to the uniform, Bronars' Method 1 distribution over choices. In this sense, jittering provides a bridge between the Bronars' Method 1, which takes no account of observed choices, and sample based inference, which assigns unobserved choice profiles measure zero.

From an implementation perspective, there are two obvious ways of specifying the standard error for the normal distribution used to jitter the data. The first is to define the standard error in proportion to the length of the budget line ($\tilde{\sigma}_t = \tilde{\sigma}\ell_t$), which we call relative errors. This sampling strategy is equivalent to drawing from a smoothed kernel density over the sample distribution for budget shares in the experiment. The second is to let the distribution be the same for all budget lines regardless of length ($\tilde{\sigma}_t = \tilde{\sigma}$), which we call absolute errors. Jittering with absolute errors is equivalent to drawing from a smoothed kernel density over the sample distribution for choices themselves in the experiment. Because the units for the relative errors bandwidth is constant across budget sets and experiments, the power function from relative jittering is well-suited for comparing experiments. In analyz-

ing the behavior within an experiment, however, jittering with absolute errors provides a measure that can be directly related to the underlying choices.

Table 3 reports the distributional properties of choices and CCEI at variable levels of $\tilde{\sigma}$ for the jittered data. Focusing on the experiment analyzing altruistic prefences, we see a remarkably different effect of the bandwidth in characterizing power using jittering and the weighted bootstrap. In the weighted bootstrap, even minor increases in the bandwidth lead to a substantial increase in GARP violations due to the presence of differing behavioral types that prevented a smooth distribution of choices on the budget set. This same agglomeration of choice requires relatively large jitters in order to bridge the gap between modal choices, so that jittering does not impart a large frequency of GARP violations until the bandwidth exceeds the relevant threshold.

The risk preferences experiment, with its smoother distribution of choices on the budget sets, does not gain much power by relaxing cross-sectional dependence in the weighted bootstrap but does respond to even slight jitters in the data. Interestingly, for the experiment on risk preferences over gains at very low bandwidths, the jittered frequency of violations actually drops compared to the sample violation frequency. This result arises for subjects whose choices violate GARP while maintaining a Critical Cost Efficiency Index of unity. In these cases, jittering the data slightly actually removes violations in 75% of the draws by moving a choice profile off the intersection of two budget sets.

5 Power Indices

In this section, we look at power indices characterizing three properties of the experiment's design. We begin with the Jittering Index to characterize the amount of noise that we would need to add to an individual subject's choices to generate GARP violations. Next, we consider the Afriat Power Index, which is defined by the degree to which GARP would need to be strengthened in order to generate violations. We close with the Optimal Placement Index as a measure of the efficiency of the experiment's design relative to a maximally efficient design for testing violations of WARP at each individual budget set.

Table 3: Jittered Measure Properties

	Violation Frequency	CCEI Average	Frequency of CCEI < 0.75	0.90	0.95	1.00	CCEI Variance Analysis St Dev	Within Subject	Between Subjects
Panel A: Altruistic Preferences									
Bronars M1	75%	0.88	16%	46%	59%	72%	0.122	100%	0%
Absolute Jittering									
$\sigma = 50$	58%	0.92	9%	31%	44%	58%	0.104	85%	15%
$\sigma = 25$	38%	0.96	4%	17%	26%	38%	0.083	64%	37%
$\sigma = 15$	27%	0.97	1%	10%	17%	27%	0.059	59%	41%
$\sigma = 5$	22%	0.99	0%	2%	6%	22%	0.024	41%	59%
Relative Jittering									
$\sigma = 1$	73%	0.88	16%	47%	60%	72%	0.120	99%	1%
$\sigma = 0.5$	65%	0.91	11%	38%	51%	65%	0.111	94%	6%
$\sigma = 0.1$	26%	0.98	1%	8%	15%	25%	0.053	56%	44%
$\sigma = 0.05$	23%	0.99	0%	2%	8%	22%	0.028	48%	53%
Sample	9%	1.00	0%	1%	2%	3%	0.017	0%	100%
Panel B: Risk Preferences over Gains									
Bronars M1	91%	0.82	27%	68%	80%	90%	0.133	100%	0%
Absolute Jittering									
$\sigma = 25$	87%	0.85	20%	60%	75%	87%	0.127	96%	4%
$\sigma = 5$	57%	0.95	5%	19%	33%	57%	0.089	53%	48%
$\sigma = 2.5$	45%	0.97	3%	11%	21%	44%	0.072	28%	73%
$\sigma = 0.5$	36%	0.98	3%	8%	15%	35%	0.059	2%	99%
Relative Jittering									
$\sigma = 1$	91%	0.82	26%	68%	81%	90%	0.132	100%	0%
$\sigma = 0.5$	89%	0.84	23%	64%	78%	89%	0.128	99%	1%
$\sigma = 0.1$	58%	0.95	3%	17%	32%	58%	0.079	51%	49%
$\sigma = 0.05$	44%	0.97	2%	10%	19%	43%	0.066	22%	79%
Sample	44%	0.98	2%	7%	14%	27%	0.057	0%	100%

This table reports properties of the Jittered Power Measures corresponding to the alternative hypothesis in equation 11 for choices observed in the experimental studies by Andreoni and Miller (2002) and Andreoni and Harbaugh (2009). The varying bandwidths, σ, correspond to the standard deviation of errors of actual goods (Absolute Jittering) or budget shares (Relative Jittering). The cross-sectional standard deviation of budget shares is averaged across budgets and the Violation Frequency reports the frequency with which a subjects' choice profile violates GARP. The table presents the distributional properties for the Afriat Critical Cost Efficiency Index (CCEI), including the percentage of the variance in the CCEI that is due to within- and between-subject variation in choices based on the decomposition in equation 9.

5.1 Jittering Index

To motivate this approach, suppose a person was offered the five budget constraints pictured in Figure 6 and all the choices involved equal quantities of both goods, as in the left panel of the figure. These choices do not violate GARP and are consistent with preferences that

have a kink at the 45-degree line, however, adding only the slightest shift in choices along the budget constraint could result in a GARP violation. Compare this to the data shown in the right panel of Figure 6 where the data look as though they are consistent with a perfect substitutes utility function and requiring very big perturbations added to the data in order to generate violations of revealed preference. Since both samples imply similar predictability in individual choices we could conclude that the left panel provides a more powerful test of rationality due to its relative sensitivity to perturbations.

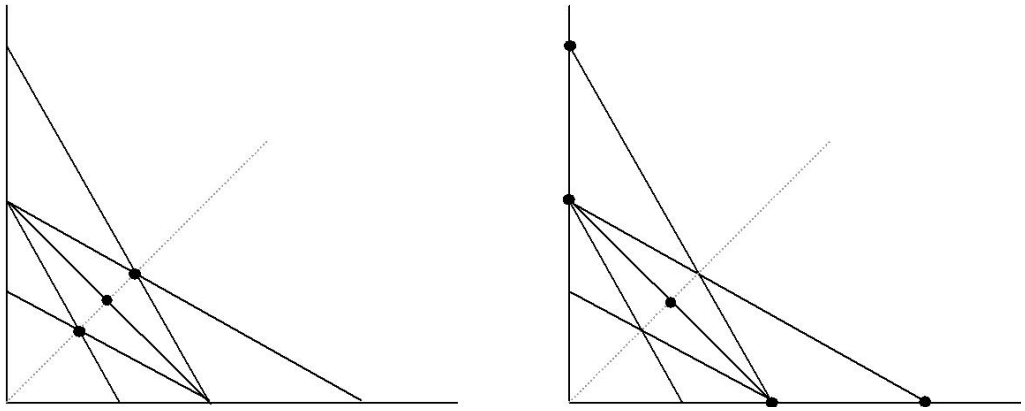


Figure 6: Different Choice Profiles for the Same Budget Sets

To formalize the approach, we need to construct two measures. The first is a measure of how much error we need to add to the data in order to generate a predetermined severity of GARP violations. The second is a measure of the amount of variance or error naturally occurring in the data. By comparing the variation we need to add with the naturally occurring variation, the jittering index can be normalized to account for the number of budget sets in the design so that indices can be compared across experiments.

We can get the first measure from the power function using the jittered power measure. The bandwidth parameter of the jittered sample corresponds to the standard deviation of the noise added to the observed sample. For each subject, we can vary this smoothing parameter to find the minimum bandwidth for which 5% of the jittered experiments find at least one GARP violation. Denoting this bandwidth $\overline{\sigma}$, it gives a direct measure of how close the chosen budgets came to finding a violation of rationality—the closer $\overline{\sigma}$ is to zero,

the sharper the test of rationality.[11]

To normalize the Jittering Index across experiments, consider tests of whether the noise added to create the jittered data, $\tilde{\varepsilon}$, is significantly bigger than the noise naturally occurring in the data, ε, under the null hypothesis that $\tilde{\varepsilon}$ and ε both have the same variance of σ^2. For each individual in the sample, the statistic:

$$\phi = \frac{\frac{1}{T(K-1)}\sum(x_{i,t} - \tilde{x}_{i,t})^2/\sigma^2}{\frac{1}{T(K-1)}\sum(x_t - r_i\left(B_t\right))^2/\sigma^2} \approx \frac{\overline{\sigma}^2}{\sigma^2}$$

is characterized by the F distribution under the null hypothesis. If there are K goods on each of T budgets, then this F–test has $T(K-1)$ degrees of freedom in both the numerator and denominator.[12] Fixing the significance level at the customary 5%, we can find the critical values from the quantiles of the F distribution, denoting these $c_{T,K-1,0.05}$. The Jittering Index is then defined as $\sigma^* = \overline{\sigma} c_{T,K-1,0.05}^{-1/2}$. Then, any $\sigma \geq \sigma^*$ would be enough natural variance to satisfy the desired confidence in the power of our test.

For interpretation, we must appeal to intuitions about whether σ^* is "small." The lack of an objective definition for a "powerful" test presents the greatest limitation for the Jittering Index, though having to specify a σ is tempered by being able to state a needed σ^* threshold for variance in the data. If σ^* is a number that all would agree is small given the nature of the data, then arguments over σ may be avoided.[13]

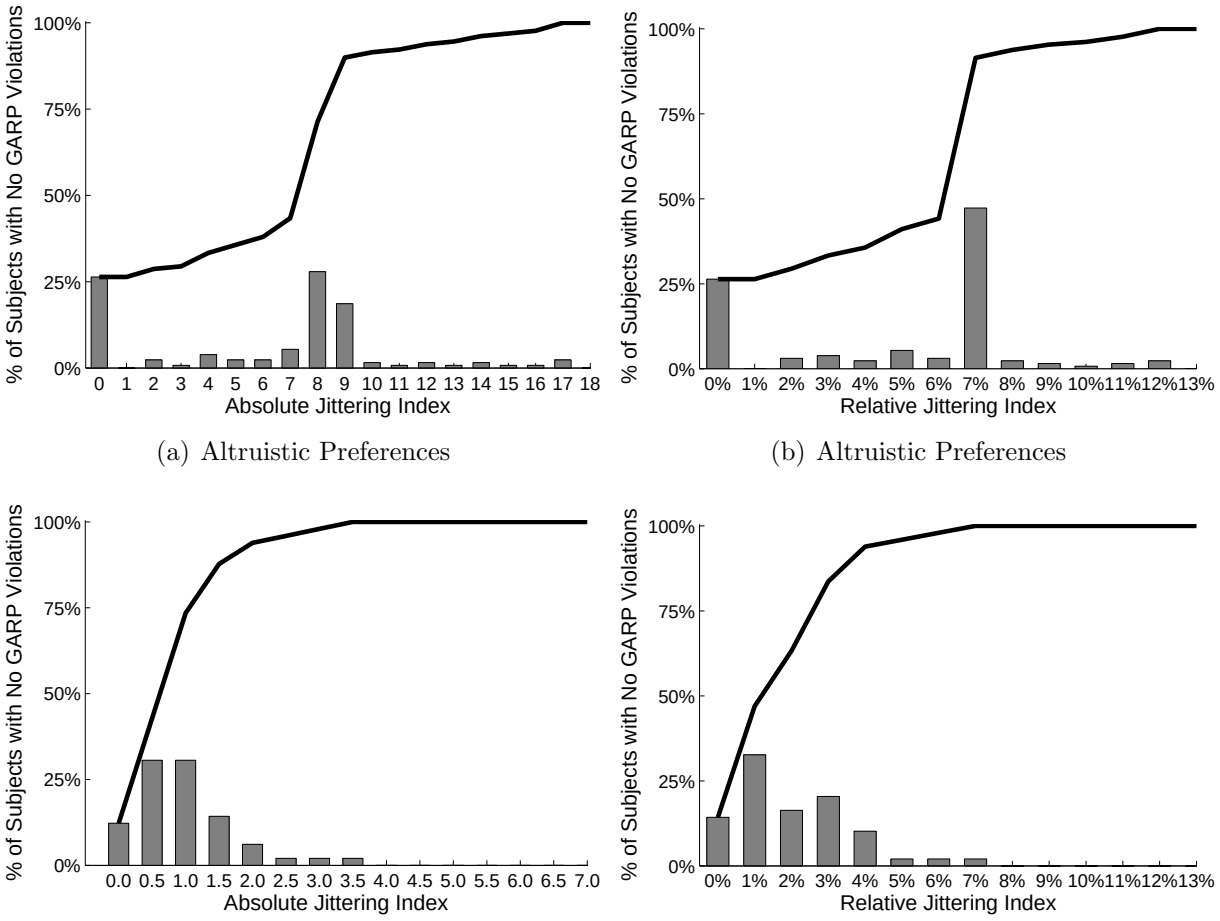
In Figure 7, we show the values of $\overline{\sigma}$ for all subjects who had no violations of GARP. The bars are the marginal frequency and the lines are the cumulative frequency. Panel (a) shows that, under absolute error, a similar degree of power holds if the natural error exceeds

[11]Note that this method even works to find power when there are violations of GARP, but just relatively few. We may still want to jitter the data to see how much noise we need to add to bring violations up to some critical value. However, in an experiment whose design has a Bronars power measure less than 5%, the Jittering Index would be unbounded (though this is not necessarily a bad result as such a design would have very weak power under any data generating process).

[12]Recall that we are thinking of K as a point on a budget plane. Thus there are only $K-1$ independent values in the vector x, and $m-1$ elements in ε. Note also that the vector notation implies that $\sum_{t=1}^{T}(x_t - z_t)^2 = \sum_{t=1}^{T}\sum_{i=1}^{K-1}(x_{ti} - z_{ti})^2$.

[13]This question is reminiscent of that encountered by Varian (1985) in his goodness-of-fit analysis, and the answers are thus similar. One option is to find a parametric estimate of a utility function and let the standard error of the regression stand for σ. This, obviously, dilutes the value of nonparametric analysis with parametric analysis. Moreover, there often may be too few observations from a single agent to estimate such a function, leaving one to postulate σ from some other ad hoc means.

$\sigma = 10$. Panel (b) shows that, under relative errors for the altruistic study, if the natural error in the data exceeds $\sigma_i = 0.08\ell_i$, then 90% of the subjects would have been given significantly powerful tests of GARP. The results for relative jittering in the risk preference experiments imply a slightly lower level of natural error required to generate violations, even after conditioning on those subjects who did not violate GARP.



(a) Altruistic Preferences

(b) Altruistic Preferences

(c) Risk Preferences over Gains

(d) Risk Preferences over Gains

Figure 7: Distribution of Jittering Index for Subjects with No GARP Violations
This figure presents the cross-sectional distribution of the absolute and relative Jittering Indices calculated for the individual subjects in the two experiments whose choices revealed no GARP violations. The Jittering Index corresponds to the maximal standard deviation of within-subject variation in choices that could be rejected by an F-Test when compared to the minimal standard deviation required to generate violations in at least 5% of the jittered samples for that subject.

This leads naturally to the question, how much natural error exists in the data? Looking

at the data from the study of altruistic preferences, one sees immediately that one source of natural error is rounding.[14] Perhaps for cognitive ease, subjects have an overwhelming tendency to choose numbers divisible by 10. This is true for both the hold and pass amounts. In fact, over 85% of all choices had both the hold and pass values divisible by 10. Another 11% were divisible by 5, but not 10. Only 4% of choices made were not divisible by either 10 or 5. Suppose we assume subjects restrict choices to those where both hold and pass amounts are divisible by 10, and that "rational rounding" would choose the point that yields the highest utility. This means that the maximum error would be at least 5, assuming convex preferences. These rounding errors alone would provide enough natural variance in the data to make at least 38% of our GARP tests have sufficient power.[15] If we were to believe that there is some other independent variation in the data, either from measurement, reporting or learning, that is roughly equal to noise from rounding so the expected absolute error was about 5 tokens on each budget, then $\sigma_i \approx 0.1$ for relative and $\sigma_i \approx 13$ for absolute errors. If this were the case, then about 95% of the GARP tests would have sufficient power.

Comparing the Relative Jittering Indices across the two studies, the experiment evaluating risk preferences over gains appears to have greater power than the experiment evaluating altruistic preferences. This enhanced power comes from the difficulty of testing rigid preferences, as evidenced by the spikes in the centers of Panels A & B in Figure 7. These are due to subjects who always chose corner solutions: 28% always kept everything, and 11% had apparently linear preferences (see Andreoni and Miller (2002), Table III). Note also that one would not expect even rounding error to be present at corner solutions, so there's very little naturally occurring variation in the data at these points. As is evident, when testing revealed preference in settings with such stark preferences, it's exceedingly difficult to design

[14]Rounding is commonly observed as a feature of choice in continuous problems in a broad array of forms. In particular, Pollison & Quah (2013) and Cosaert & Demuynck (2012) explore the power of revealed preference in aggregated or discretized choice spaces.

[15]To be conservative, assume a uniform distribution of absolute rounding errors between 0 and 5, and thus an expected absolute error of 2.5 tokens. Under the assumption of relative errors, this implies $E|\varepsilon_i| = 0.43$ and for absolute errors, this implies of $E|\varepsilon_i| = 5.7$. It is easy to show that our assumptions imply the standard error of $\sigma \approx 1.15E|\varepsilon_i|$. For the assumption of relative errors, this means $\sigma_i = 0.049$, while for absolute errors it means $\sigma_i = 6.53$.

an experiment with a viable amount of power.

5.2 The Afriat Power Index

Although the index proposed in this section was not suggested by Afriat, it seems natural to give it his name given its similarity to the Afriat Critical Cost Efficiency Index. To characterize the severity of a violation of revealed preference, Varian (1990, 1991) builds on Afriat (1967, 1972) to construct the Afriat Critical Cost Efficiency Index. Varian first relaxes the directly revealed preferred relation by defining, $R^d(e)$, so that: $x_j R^d(e) x$ iff $ep_j x_j \geq p_j x$, where $0 \leq e \leq 1$. It follows to define $R(e)$, a relaxed revealed preference relation, as the transitive closure of $R^d(e)$. Varian defines a version of GARP, which we call L-GARP(e) ("L" for lower), as

Definition: *L-GARP(e):* If $x_j R(e) x_k$, then $ep_k x_k \leq p_k x_j$, for $e \leq 1$.

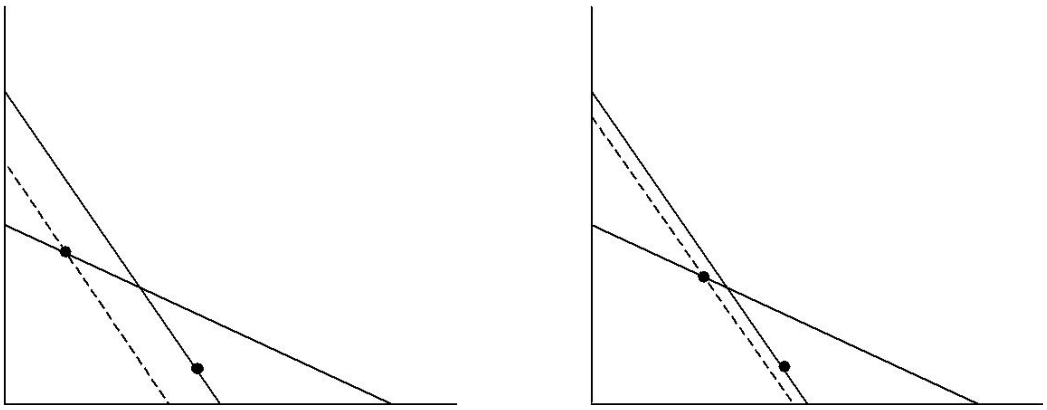


Figure 8: The Afriat Critical Cost Efficiency Index
The Afriat Critical Cost Efficiency Index (CCEI) measures how far budget sets would need to be shifted to remove any violations of GARP from the data. A choice of consumption bundles with a low CCEI (as in the left panel) indicates a more severe violation of GARP than one with a CCEI near unity (the right panel).

Afriat's Critical Cost Efficiency Index, or the Afriat Efficiency Index for short, is the *largest* value of $e \leq 1$, say e^*, such that there are no violations of L-GARP(e). If $e^* = 1$ then there are no violations of GARP in the original data, but for $e^* < 1$ there are violations. Traditionally, researchers begin their analysis of consumer behavior by setting some critical level of e^*, say $\overline{e}$, such that they would consider any $e^* \geq \overline{e}$ a small or tolerable violation of

GARP. Varian (1991), for instance, suggests a value of $\overline{e} = 0.95$.[16]

Suppose a set of choices *does not* violate GARP. If the budget constraints cross near the area that subjects are actually choosing, then we can think of that set of budgets as being more diagnostic than a different set in which the choices are far from the intersections. For instance, Figure 9 shows two budgets without violations of revealed preference. However, the frame on the right gives us more confidence that the person choosing these goods satisfies utility maximization. If there were a violation or rationality, we would be more likely to uncover it in the right panel since even a small change in choices would have been enough to violate GARP. In the frame on the left, by contrast, there would have to be much larger violations of rationality before we could uncover them with this test.

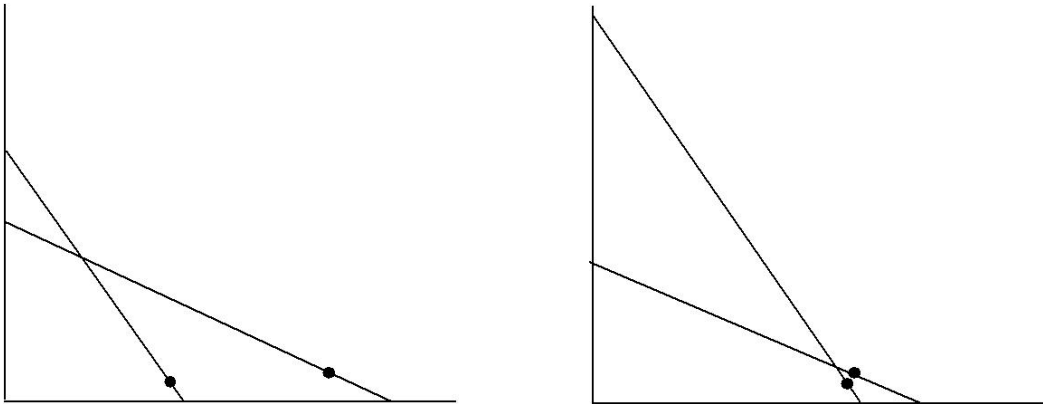


Figure 9: Consumption Choices that do not Violate GARP

To capture the intuition behind Figure 9, define a stronger direct revealed perference relation, $\widetilde{R}^d(g)$, as $x_j \widetilde{R}^d(g) x_k$ iff $g p_j x_j \geq p_j x_k$, where $g \geq 1$. Thus, if $g = 1$ we have the standard notion of directly revealed preferred. Then let $\widetilde{R}(g)$ be the transitive closure of $\widetilde{R}^d(g)$. Given this stronger notion of revealed preference, we can define a new concept H-GARP ("H" for higher) as

Definition: *H-GARP(g):* If $x_j \widetilde{R}(g) x_k$, then $g p_k x_k \leq p_k x_j$, for $g \geq 1$.

[16]Note that, since the Afriat Efficiency Index is just a transformation of individual choices, it can be defined as a random variable on the probability space for the experiment. As such, its distributional properties can be inferred using the sampling techniques presented in section 4. Further, we can extend the variance decomposition from individual choices to separate a priori variation in the Afriat Efficiency Index into individual and cross-sectional components.

Using this inverted notion of the Afriat Efficiency Index, we can define the Afriat Power Index as the *infimal* value of $g \geq 1$, say g^*, such that there is *at least one* violation of H-GARP(g). If $g^* > 1$ there are no violations of GARP in the data, but if g^* is close to 1 the choices are near where the budget constraints intersect. An example of the Afriat Power Index is shown in Figure 10. The choices on the left are less informative about rationality than those on the right, and the Afriat Power Index is closer to 1 in the panel on the right. Hence, while the Afriat Efficiency Index told us how much we need to "relax" the budgets to avoid violations, the Afriat Power Index tells us how much we need to "expand" budgets in order to generate violations.[17]

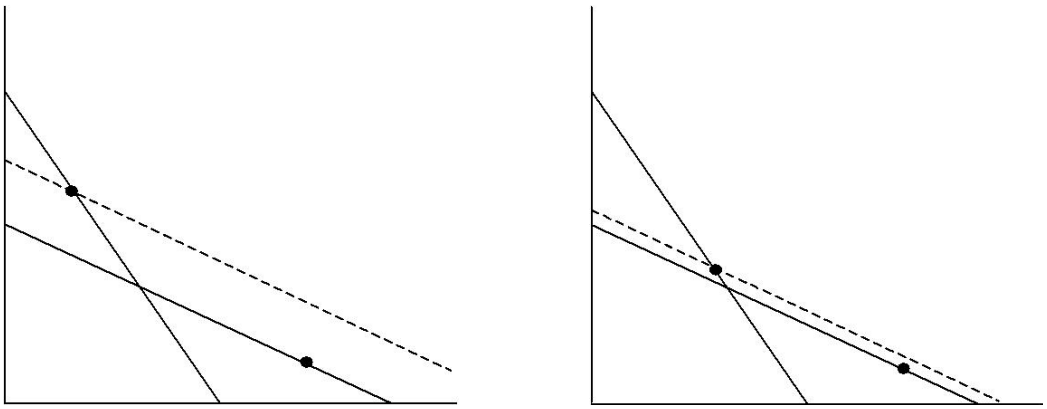


Figure 10: The Afriat Power Index

The Afriat Power Index measures how much budget sets would need to be shifted outward for a violation of GARP to be observed in the choice sample. Subjects whose choices would require greater shifts in the budget set to generate a violation (as in the left panel) face a less demanding test than those whose choices can generate violations with only small shifts in the budget sets (as in the right panel).

When can we say that the g^* found from the Afriat Power Index is "too big" and thus has too little power? One obvious approach is to switch our perspectives. If, under the Afriat Efficiency Index we were willing to accept any $e^* \geq \bar{e}$ as an acceptably small violation of GARP, then any $g^* \leq 2 - \bar{e}$ should also be an acceptably powerful test of GARP.

As the original Afriat Efficiency Index has some shortcomings that have been highlighted

[17] Consider the special case where a single choice is made at the point where two budgets intersect but that there is no violation of GARP. Then the smallest shift in one budget constraint will create a violation, in which case $g^* = 1 + \delta$, where δ is infinitely small. For ease of discussion, we will refer to this as a case of $g^* = 1$.

in the literature, the Afriat Power Index is subject to similar critiques. For instance, the Afriat Efficiency Index is defined by only one violation, and does not give credit to an individual who may otherwise have large numbers of perfectly rational choices. In other words, it is not very forgiving of a single error. By the same token, it can potentially mask the troubling nature of a large number of small errors. Similarly, the Afriat Power Index will score well if there is a single pair of budget constraints which cross near the choices, even if all other budget constraints cross far from the choices. As with the Jittering Index, this limitation of the Afriat Power Index highlights the challenge of testing rationality in settings with sharp preferences.

5.3 The Afriat Confidence Index

Assign an individual i a number $A_i = e_i^* g_i^*$, where e_i^* is the Afriat Efficiency Index and g_i^* is the Afriat Power Index.[18] Call A_i person i's *Afriat Confidence Index*. If $A_i < 1$ the person has at least one violation of GARP and this number can be interpreted as indexing the severity of the violation. If $A_i > 1$ then the person has no violations of GARP, and the number can index the stringency of the GARP test. An $A_i = 1$ corresponds to the ideal data—the person could not have been given a sharper test. By selecting an $\overline{e}$ prior to analysis we gain a "confidence interval" on A_i, that is $\overline{e} \leq A_i \leq 2 - \overline{e}$. An A_i in this interval can be seen as a successful test of GARP.

Figure 11 shows the distribution of Afriat Confidence Indices, A_i, for the experimental samples. The left hand side of the graph presents results for subjects who had at least one violation of GARP (and, as such, an ACI ≤ 1) and the right hand side of the graph presents results for subjects with no violations of GARP (and an ACI ≥ 1).

To focus on the power properties of the experiments' designs, consider those subjects who had no GARP violations. In the rational altruism study, more than two-thirds of these

[18] Alternatively, we could derive A_i from a unified framework. Define $R_A^d(a)$ as $x_j R_A^d(a) x_k$ iff $a p_j x_j \geq p_j x_k$, for some $a > 0$, and let R_A be the transitive closure of R_A^d.
Define *A-GARP(a):* If $x_j R_A(a) x_k$, then $a p_k x_k \leq p_k x_j$, for $a \geq 0$.
Then let $a_i^* = inf\{a$: there exits a single violation of *A-GARP(a)*, or at which the smallest change in a would remove all violations of *A-GARP(a)*$\}$. Then $A_i = a_i^*$.

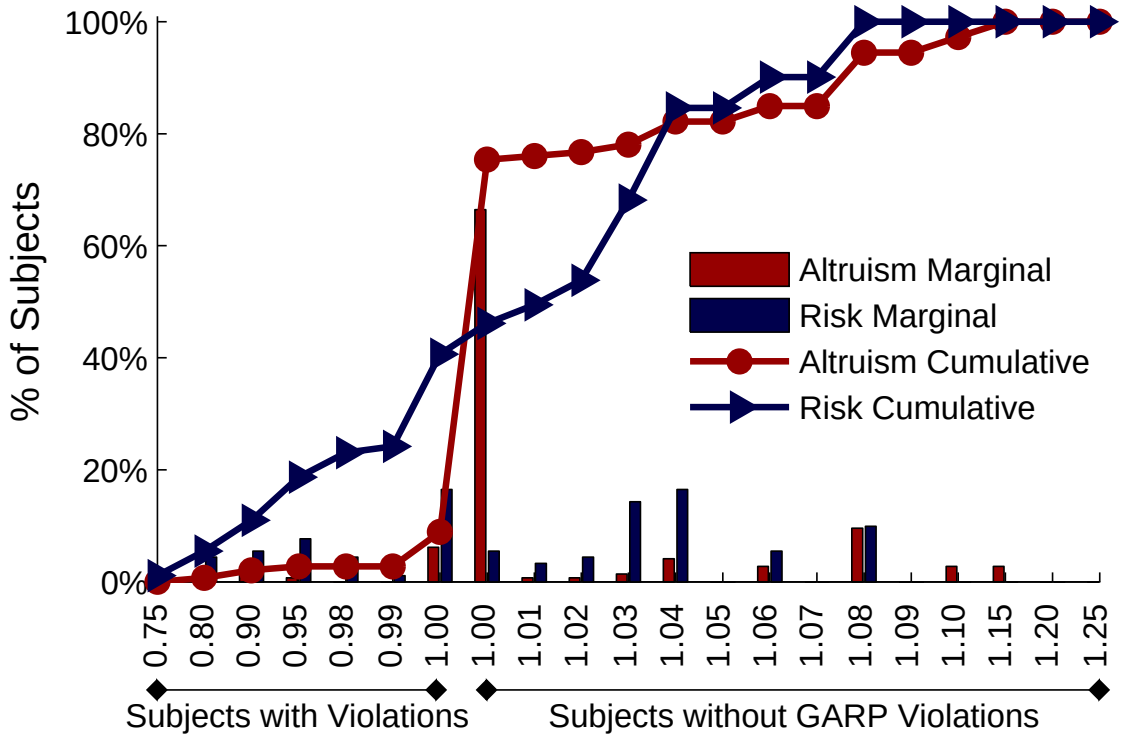


Figure 11: Afriat Confidence Index, A_i

This chart presents the cross-sectional distribution of the Afriat Confidence Index for choices observed in the experimental studies by Andreoni and Miller (2002) and Andreoni and Harbaugh (2009). Subjects whose scores are far above unity did not encounter a sufficiently stringent test of their preferences. Subjects whose scores are far below unity severely failed to pass the GARP test. The population of subjects with A_i near unity either passed a stringent test of their choice behavior or failed within the tolerance of the Critical Cost Efficiency Index.

(71%) had Afriat Confidence Indices of 1, indicating that the GARP test could not have been sharper. If we apply the same criterion for "high power" that we do to "small violation" then 107 (83%) of the non-violators have $A_i \leq 1.05$. Then defining an "acceptably stringent success range" for Afriat Confidence Indices such that $0.95 \leq A_i \leq 1.05$ means that 82% of subjects in the sample were given stringent tests of GARP and passed, 2% of subjects had significant violations of GARP ($A_i < 0.95$) and 16% were given GARP tests that were not sufficiently diagnostic ($1.05 < A_i$).

Comparing the results from altruism study to those from the experiment on risk preferences, the Afriat Confidence Indices have a much smoother distribution, reflecting the fact that the budget line crossings were less "focal" compared to the altruism study. The "ac-

ceptably stringent success range" for rational risk preferences means 75% of subjects can be characterized as passing stringent GARP tests, with 15% presenting significant violations of GARP and 10% revealing the test was not sufficiently diagnostic.

As such, mainly through the higher frequency of violations in the risk preferences experiment, the altruism experiment yielded greater confidence in the validity of GARP. However, the budget sets in the altruism experiment were placed in such a way that most observed choices would fail any stronger notion of GARP. As such, the lower frequency of violations is not due to faulty design, but rather arises from the structure imposed by individuals on their choices.

5.4 The Optimal Placement Index

Consider the choices a on budget A in the left panel of Figure 12 and suppose that, *ex post*, we wanted to alter the placement of budget set B without changing relative prices to test whether the choice a satisfies WARP. In this case, we would obviously choose a budget that would intersect A at point a, corresponding to budget set C. How much more efficient is C than B at testing rationality? As seen on the right panel of Figure 12, on budget B there is a fraction d/D of choices available that would violate WARP, while on budget C there is a fraction e/E of available choices that would violate WARP. Hence, we define the Optimal Placement Index to indicate the relative efficiency of the placement for budget set B to test WARP for choice a as

$$\theta_{a,b} = \frac{d/D}{e/E} = \frac{d}{e}\frac{E}{D}$$

How about choice b on budget B in Figure 12? Here the budget A has no ability to find a violation of WARP, conditional on the observation of b. In this sense, the test has no power to show that b was chosen irrationally. Hence, we can say $\theta_{ba} = 0$. As such, we can calculate two *directed* Optimal Placement Indices for any pair of budget sets. We discuss how to aggregate across these directed power measures in the next subsection.

To relate this calculation to the power measures introduced in the previous section, notice that the numerator of the index $\theta_{\tau,-\tau}$ corresponds to the Bronars' WARP Power Measure

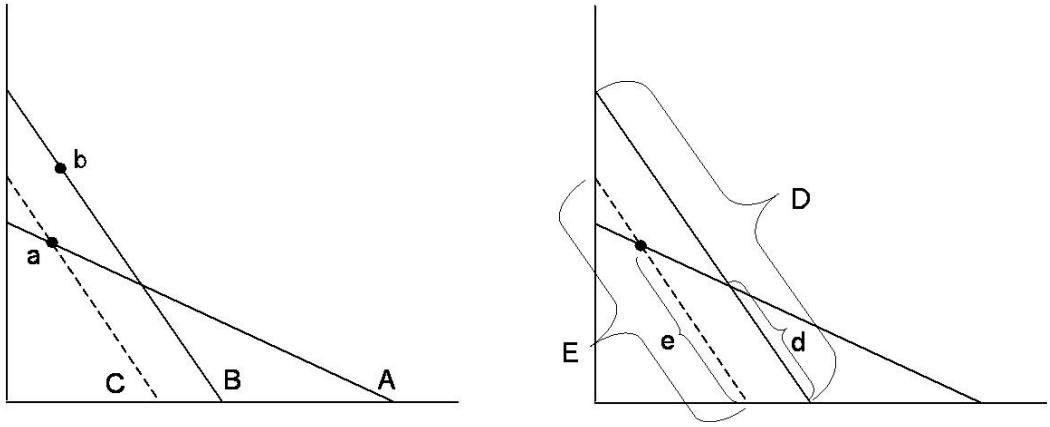


Figure 12: The Optimal Placement Index
The Optimal Placement Index measures the power of the implemented test (d/D) relative to the power of an optimally designed test (e/E) assuming the Bronars' M1 alternative hypothesis of choice behavior on unobserved budget sets.

(BWPM) for budget set $B_{-\tau}$ conditional on the choice x_τ. Further, denoting by $\tilde{B}_{-\tau}$ the budget set with prices $p_{-\tau}$ and wealth $\tilde{m}_{-\tau} = p_\tau x_\tau$, we can see that the denominator of the index $\theta_{\tau,-\tau}$ corresponds to the BWPM for budget set $\tilde{B}_{-\tau}$ conditional on the choice x_τ. The Optimal Placement Index then represents a ratio of the BWPM measure under the experiment as currently designed to the maximum BWPM that would be attainable by shifting the income levels for a single budget set in the experiment.[19]

Note that if a choice lies on two budget constraints, then this approach will assign a directed OPI to that point of $\theta_{\tau,tau} = 1$. As shown above when calculating $\theta(b, a$ in the left hand panel of Figure 12, it is also possible for the index to take on a value of zero. Such a value arises here given choice b on budget set B because, emphex post, any choice along budget set A would be consistent with GARP. However, as with the Afriat Power Index, the Optimal Placement Index will always show power if at least one pair of choices can be ranked by revealed preference.

[19]As formulated, the Optimal Placement Index is only operable for WARP, not GARP. While it could in principle be generalized to GARP, taking the perspective of shifting all budget sets relative to one another would not be feasibly for computational reasons. Moreover, with only two goods it is impossible to have a violation of GARP without also having a violation of WARP, although this is not the case with more goods. As a result, this power index is more demanding of the budgets than GARP requires, implying that the true power of the test is likely to be higher than this index might imply.

Extending this notion to a setting with two budget sets and more than two goods is immediate by comparing the BWPM with the supremum attainable BWPM. By comparing the relative magnitude of the rejection region with the acceptance region, this analysis provides an analog to the difference power index presented in Beatty and Crawford (2011).

5.4.1 Aggregating the Optimal Placement Index Across Budget Sets

Since GARP tests analyze choices at more than two budgets, so we must aggregate the directed placement indices to provide a summary characterization of a given test's efficiency. To begin, for every budget t, calculate the value of $\theta_{t,\tau}$ for all $\tau \neq t$ other budgets (if a budget τ does not cross budget t or has zero power given the relative prices for any income level then $\theta_{t,\tau} = 0$).

There are two natural aggregations for the $\theta_{t,\tau}$ indices across τ for budget t. The first is to define $\theta_t^{MAX} = \max\{\theta_{t,1}, \theta_{t,2}, ..., \theta_{t,t-1}, \theta_{t,t+1}, ..., \theta_{t,T_i}\}$. This θ_t^{MAX} represents the least possible improvement in the test's power at that budget set, or equivalently, the maximum efficiency of those tests given the observed choice x_t. An alternative way to define the index would be to begin by averaging across the other budgets, defining $\theta_t^{AVG} = \frac{1}{T_i - 1} \sum_{\tau \neq t} \theta_{t,\tau}$. This θ_t^{AVG} is the expected ratio of the BWPM under the currently designed experiment to the maximum of the BWPM attainable by changing the income for one randomly selected budget set given the observed choice x_t.

To construct an overall power index for individual i, we propose averaging across θ_t, defining $\theta^{\cdot} = \frac{1}{T_i} \sum_{t=1}^{T_i} \theta_t^i$. This average represents the expected benefit to randomly selecting a reference budget set to use as the basis for shifting the experimental design.[20]

In aggregate, then, we have two definitions for an individual subject's Optimal Placement Index: the Average Budget OPI and the Maximum Budget OPI. The Average Budget OPI

[20] We also considered an aggregation that takes the maximum across θ_t, defining $\theta^{\cdot} = \max\{\theta_t^i\}$. This aggregated $\theta^{\cdot}$ would represent the highest efficiency in terms of the BWPM for the experiment as currently designed compared to the maximum BWPM attainable by shifting one level of income for one budget set conditional on all observed choices. In practice, this maximum is often truncated at unity and is consequently relatively uninformative. Further, the aggregated index's value would be determined by the placement of other budget sets relative to only a single budget set, whereas taking the average incorporates information about the placement of all budget sets.

presents the most diffuse measure of power, but may be a bit too conservative as it requires every budget set to have power against every other budget set, weighing a heavy penalty against designs that include budget sets that are shifted outward. The Maximum Budget OPI balances this by evaluating the efficiency of the most efficient WARP test for all budgets, though at the cost of ignoring the frequency with which such an efficiency is achieved.

5.4.2 Strengths and Weaknesses of the Optimal Placement Index

The analysis above is based on the optimal placement of budget constraints only for predetermined price vectors. As Bronars has shown, the test that will expose the individual to the greatest chance for a GARP violation is the budget that puts the most area under the budget in question. If preferences are normal, then this same conclusion follows naturally from Proposition 1 of Blundell, Browning and Crawford (2003), while Beatty and Crawford (2011) establish a similar result by invoking Selton (1991)'s measure of predictive success. Since the price vector could be shifted to produce a budget set with a greater chance for a GARP violation, this constraint prevents the placement of these budget sets from being truly optimal.[21]

Underlying its definition, the optimal placement index implicitly assumes the Bronars measure for choices over non-observed budget sets. This assumption is maintained in Beatty and Crawford's difference power index. As Dean and Martin (2012) do with a bootstrap approach, we could leverage an approach discussed in the section 4 to take observed choices into account when defining the relative power under different placements.

Lastly, as with the Jittering and Afriat Power Indices, there is no objective magnitude for which the Optimal Placement Index indicates an "efficient test." As such, its interpretation requires an intuitive notion of efficiency, with Optimal Placement Index values across studies

[21] If we were to optimally choose relative prices as well as placement, the maximally efficient budget set would not be well defined as it would be arbitrarily close to the original budget set. We could define the supremum of the power for such an optimally placed budget set as the largest budget share in the consumption bundle chosen on that budget set. We could also define WARP violations subject to a maximum Afriat Confidence Index of some critical value. With more than two goods, such an optimally placed budget set itself would not be uniquely defined as there will be a continuum of budget sets with equivalent power. Despite this multiplicity, the maximum such power will be attained by any of these budget sets and Optimal Placement Index itself would remain well-defined.

being informative solely as an ordinal ranking for the efficiency of test designs.

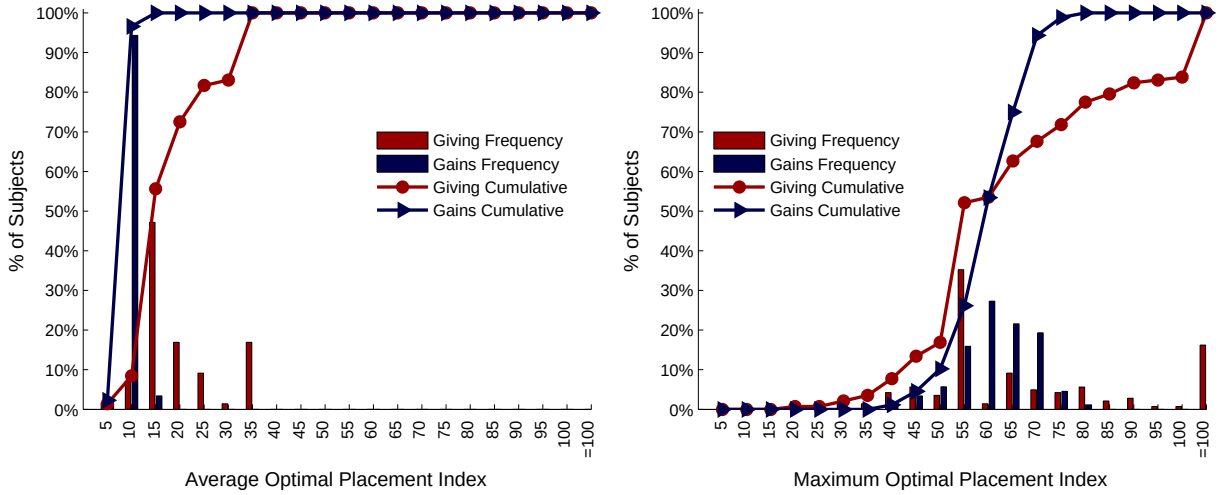
5.4.3 The Distribution of the Optimal Placement Index

Figure 13 shows the sample distributions of the Optimal Placement Index. The Average Optimal Placement Indices for both studies are quite small, with the median Average OPI for both experiments falling under 15%. The Maximum OPI shows the median participant in each experiment faced at least one budget with 55%-60% efficiency. Comparatively, the Average OPI indicates the experiment analyzing altruistic behavior was slightly more efficient than the study of risk preferences. The Maximum OPI is less stark in its ranking, but does have substantially more subjects whose Maximum OPI was above 75%. Again, subjects who keep all of the tokens for themselves appear as a mass in the distribution. With an OPI of 100%, these subjects having been given the most stringent test possible at all budget sets, but still have not violated GARP.

The Optimal Placement Index reveals very similar power properties to the Afriat Confidence Index. The experiment evaluating preferences over altruism, by taking account of focal features in that space (such as equality, equity, and selfishness), is able to generate a very tight test of GARP. In exploring preferences over risk, however, the lack of focal features results in a "looser" test. Under the Optimal Placement Index, this latter experiment would seem to have less power. However, the heterogeneity in preferences results in much more frequent violations, which is why the Afriat Confidence Index reveals these experiments to have relatively strong power despite this inefficiency.

6 A Field Guide to Characterizing Experimental Power

While the indices presented above all focus on measuring the power of a test's design, each does so from a rather different perspective. Bootstrapping the data helps characterize the role of heterogeneous preferences as potentially driving GARP violations. The Afriat Confidence Index and Jittering Index both characterize how close choice behavior in the experiment is to generating violations, though they differ in the metric that measures this closeness. The



(a) Average Optimal Placement Indices (b) Maximum Optimal Placement Indices

Figure 13: Distribution of Optimal Placement Indices

This chart presents the cross-sectional distribution of the Optimal Placement Indices (OPIs) for choices observed in the experimental studies by Andreoni and Miller (2002) and Andreoni and Harbaugh (2009). The Average OPI corresponds to the average ex post efficiency of the design in terms of BWPM for each budget. The Maximum OPI corresponds to the maximum ex post efficiency for each budget. For each individual, the OPI measures are averaged across all budgets, indicating the average efficiency of the design for that individual across budget sets.

Optimal Placement Indices only incorporates observed choices as an anchor to measure the efficiency of the design.

Considering multiple perspectives in evaluating the power of these tests gives a more detailed characterization of the features of observed choices that result in violations of GARP. However, these different perspectives, while all informative, may be redundant for some experimental contexts. As such, some guidance as to which measures to implement, and when, could help researchers seeking to take advantage of these measures.

The empirical results presented throughout the paper highlight the role of underlying preferences in determining the degree to which a power index accurately reflects the power of the experimental test of GARP. When preferences are concentrated around modal types, power measures that ignore individual heterogeneity (such as the unconditional bootstrap) can overstate the design's power. Similarly, if choices are clustered at corner solutions, the Optimal Placement Index may consider a relatively easy test to be perfectly efficient

whenever the corner solution sits at the intersection of two budget sets. When choice behavior is more diffuse, as is the case with most experimental choice settings, it is more difficult to design "sharp" tests that maximize the power. However, even sight perturbations to observed choices can yield violations of revealed preference axioms, so these designs will still be rated highly by the Jittering Index and the Afriat Confidence Index.

Ideally, the power index should be matched with the source of variation in the data. For choice settings in which decisions are exposed to idiosyncratic noise, perhaps due to rounding or complexity in the problem, the Jittering Index or Afriat Confidence ought to provide similar characterizations of power. In contexts where choice behavior is relatively well structured but not very noisy, perhaps due to the presence of latent types with strong preferences, Optimal Placement Indices can characterize the efficiency of the experimental design and the weighted Bootstrap can illustrate how far you would need to relax this structure to yield violations.

As an initial metric to evaluate an experiment's power, we recommend the Afriat Confidence Index due to its ease of computation and close relationship with the Critical Cost Efficiency Index, a statistic that's already commonly reported for revealed preference tests. If an experiment's power is found to be lacking due to relatively high Confidence Indices, designers may wish to dig further, possibly starting with the Optimal Placement Index to characterize the design's efficiency. While the jittering and bootstrap measures are the most informative characterizations of power, their computational intensity would argue that they are best used in those settings where neither the Afriat Confidence nor the Optimal Placement Indices yield conclusive results.

As ex-post metrics, the techniques we've proposed are designed to evaluate the effectiveness of a design after its completion. To deploy these tools in the process of designing experiments, researchers could apply them to the outcomes of early pilot studies. Indices like the Optimal Placement Index can be particularly helpful in identifying adjustments to budget sets that may help enhance experimental efficiency. Similarly, measures like the Jittering Measure and the Afriat Power Index could indicate when and where additional budget sets

could improve power. In this context, of course, samples will be very small, warranting a caveat that researchers should take care not to infer too much precision from small samples.

7 Discussion and Conclusion

This paper presents, analyzes, and compares several approaches to measuring the power of revealed preference tests. We first characterize different measures of choices by a representative individual in the population as well as conditional on the observed decisions and propose sampling strategies for drawing representative choice profiles from the population. We then distill these measures into statistics that characterize the power of the test both overall and at an individual-subject level.

In terms of measures over choice, our generalized conditional bootstrap and nonparametric kernel-based jittering measures provide novel mechanisms for characterizing the distribution of choice from observed behavior. Several approaches could extend these measures to budget sets that do not occur in the population, for instance, by sampling from budget shares in a manner similar to Dean and Martin (2012) but weighting the sampling measure by characteristics of the budget sets themselves.

In translating these measures into indices, we seek to average over the underlying choice process to develop a more intuitive characterization of power. Our most straightforward index inverts the well-known Afriat Efficiency Index into the Afriat Power and Confidence Indices, allowing us to characterize the degree to which our theory would need to be strengthened to be violated by observed choice behavior. The Optimal Placement Index, which asks how well the experimental design performed relative to the best possible design that could have been dynamically generated after each choice, provides a nice tool for comparing the efficiency of two different GARP tests. The Jittering index adopts a modicum of structure on the distribution of within-subject variation in choice, addressing the question of how noisy the data must be in order for a design to have power. The intuitive appeal of each of these metrics is tempered by the fact that there is no clear guidance on power or a natural threshold to appeal to as "high power." As such, beyond intuitive characterizations of when

a test is "good enough," the metrics are better thought of as ordinal rather than cardinal measures of power.

In sum, the tension in controlling for, and the parallels between measuring, goodness-of-fit and power are clear in revealed preference tests, whether using survey or experimental data. In this paper, we hope to have provided some guidance to researchers to both design and analyze tests that maximize our ability to make the correct inferences about economic models of maximizing behavior.

8 References

Afriat, S. (1967): "The Construction of a Utility Function From Expenditure Data," *International Economic Review*, 8, 67-77.

Afriat, S. (1972): "Efficiency Estimates of Production Functions," *International Economic Review*, 13, 568–598.

Aizcorbe, Ana M. (1991): "A Lower Bound for the Power of Nonparametric Tests," *Journal of Business and Economic Statistics,* 9, 463–467.

Andreoni, James and William T. Harbaugh (2009): "Unexpected Utility: Experimental Tests of Five Key Questions about Preferences over Risk," *Working Paper*.

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

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

Beatty, Timothy K.M. and Ian A. Crawford. (2011): "How Demanding is the Revealed Preference Approach to Demand?" *American Economic Review,*, 101(6), 2782—95.

Becker, Gary S. (1962): "Irrational Behavior in Economic Theory," *Journal of Political Economy,* 70, 1–13.

Blundell, Richard W., Martin Browning, and Ian A. Crawford (2003): "Nonparametric Engel Curves and Revealed Preference," *Econometrica*, 71, 208–240.

Blundell, Richard W., Dennis Kristensen and Rosa Matzkin (2012): "Bounding Quantile Demand Functions using Revealed Preference Inequalities," *Cemmap Discussion Paper*, UCL-IFS.

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

Burghart, Daniel, Paul Glimcher, and Stephanie Lazzaro (2012): "An Expected Utility Maxi-

mizer Walks into a Bar...," *Working Paper*.

Cosaert, Sam and Thomas Demuynck (2013): "Revealed Preference Theory for Finite Choice Sets," *Working Paper*.

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

Dean, Mark and Daniel Martin (2011): "Testing for Rationality with Consumption Data: Demographics and Heterogeneity," *Working Paper.*

Dean, Mark and Daniel Martin (2012): "A Comment on How Demanding Is the Revealed Preference Approach to Demand?" *Working Paper.*

Echenique, Federico, Sangmok Lee, and Matt Shum (2012): "The Money Pump as a Measure of Revealed Preference Violations," *Journal of Political Economy.* Forthcoming.

Epstein, Larry G. and Adonis J. Yatchew (1985): "Non-parametric Hypothesis Testing Procedures and Applications to Demand Analysis," *Journal of Econometrics*, 30, 149–169

Famulari, Melissa (1995): "A Household-Based, Nonparametric Test of Demand Theory," *Review of Economics and Statistics,* 77, 372–382.

Février, Philippe and Michael Visser (2004): "A Study of Consumer Behavior Using Laboratory Data," *Experimental Economics*, 7, 93–114.

Gross, John (1995): "Testing Data for Consistency with Revealed Preference," *Review of Economics and Statistics,* 701–710.

Gross, John (1991): "On Expenditure Indices in Revealed Preference Tests," *Journal of Political Economy*, 99, 416–419.

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

Hauthakker, Henrick (1950): "Revealed Preference and the Utility Function," *Econometrica,* 17, 159–174.

Houtman, M., and J. A. H. Maks (1985): "Determining all Maximal Data Subsets Consistent

with Revealed Preference," *Kwantitatieve Methoden*, 19, 89-104.

Kitamura, Yuichi and Jorge Stoye (2013): "Nonparametric Analysis of Random Utility Models: Testing," *Working Paper*.

Manser, Marilyn E. and Richard J. McDonald (1988): "An Analysis of Substitution Bias in Measuring Inflation, 1959-1985," *Econometrica*, 56, 909–930.

Manski, Charles (2007): "Partial Identification of Counterfactual Choice Probabilities," *International Economic Review*, 48(4), 1393–1410.

Manski, Charles (Forthcoming): "Identification of Income-Leisure Preferences and Evaluation of Income Tax Policy," *Quantitative Economics*.

Mattei, Aurello (2000): "Full-Scale Real Tests of Consumer Behavior Using Expenditure Data," *Journal of Economic Behavior and Organization*, 43, 487–497.

Polisson, Matthew (2012): "Goods versus characteristics: dimension reduction and revealed preference," *IFS Working Paper*, W12/02, 1–22.

Polisson, Matthew and John K.H. Quah (2013): "Revealed Preference in a Discrete Consumption Space," *American Economic Journal: Microeconomics*, 5(1), 28–34.

Samuelson, Paul A. (1938): "A Note on the Pure Theory of Consumer Behavior," *Econometrica*, 5, 61–71.

Selten, Reinhard (1991): "Properties of a Measure of Predictive Success." *Mathematical Social Sciences,* 21(2): 15367.

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

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

Varian, Hal R. (1983): "Nonparametric Test of Models of Consumer Behavior," *Review of Economic Studies*, 50, 99–110.

Varian, Hal R. (1984): "The Nonparametric Approach to Production Analysis," *Econometrica,*

52, 579–597.

Varian, Hal R. (1985): "Non-Parametric Analysis of Optimizing Behavior with Measurement Error," *Journal of Econometrics,* 30, 445–458.

Varian, Hal R. (1990): "Goodness-of-Fit in Optimizing Models," *Journal of Econometrics*, 46, 125–140.

Varian, Hal R. (1991): "Goodness of Fit for Revealed Preference Tests." University of Michigan CREST Working Paper Number 13.

Appendix: Design Details for Experimental Data

In this section, we describe the experimental studies used in the paper to illustrate the properties of different power measures and indices. The first experiment considers individual preferences for altruism, which are characterized by a set of "modal" preferences, where an individual's choices are commonly consistent with heterogeneous types. The second experiment looks at risk preferences, which are much more diffuse, as an individual's choices tend not to be driven by any sense of normative behavior. As such, the two studies illustrate two different settings in which to evaluate measurement of power for GARP: the first where the type of an individual puts a lot of structure on the data, the second where the individual choices are more fungible across budget sets.

Altruistic Preferences

The first sample we use is described in detail in Andreoni and Miller (2002) and Andreoni and Vesterlund (2001). Briefly, the experiment was designed to explore individual preferences for altruism by asking subjects to make a series of choices in a Dictator game, under varying incomes and costs of giving money to another subject. In particular, subjects made eight choices by filling in the blanks in statements like this: "Divide M tokens: Hold ____ at X points, and Pass ____ at Y points (the Hold and Pass amounts must sum to M)," where the parameters M, X, and Y were varied across decisions. The subject making the choice would receive the "Hold" amount times X, and another subject would receive the "Pass" amount times Y. All points were worth \$0.10.

Let π_s be payoff to self, and π_o be payoff to other. The hypothesis is that individuals have well-behaved preferences $U_s = U(\pi_s, \pi_o)$.The experimental parameters imply a budget constraint for any choice of

$$\frac{1}{X}\pi_s + \frac{1}{Y}\pi_o = M.$$

The parameters chosen provided the budgets shown in Figure 9. As can be seen, the pie to be divided ranged from \$4 to \$15 and the relative prices ranged from 3 to 1/3. After subjects made all 8 choices, one choice was selected at random by the experimenter and carried out.

Data was collected on 142 subjects and each subject's choices were tested for violations of GARP.[22] The result was that 13 of the subjects (9.1%) had violations of GARP. Applying the Afriat Efficiency Index, only 3 of these were found to be large violations (as we show below). This is a rather striking failure to contradict the neoclassical model of preferences, but leaves open the question of how discriminating the GARP test was at uncovering potential violations.[23]

Risk Preferences over Gains

In addition to the altruism study, we analyze the power properties for one of the treatments in Andreoni and Harbaugh (2009), who explore rationality of risk preferences and aversiveness over gains and losses. In their experiment, individuals face a lottery that has a probability p of winning $x > 0$, and wins zero otherwise. The subjects are offered to choose p and x from a linear budget, say

$$r_1 p + r_2 x = m, \text{ where } r_1, r_2, \text{ and } m > 0. \tag{12}$$

That is, to get a bigger prize, one has to accept a smaller chance of winning it. If these preferences over risk are rational and well-behaved, then $(p; x)$ choices should satisfy the axioms of revealed preference.

The experiment is split into two treatments. In the first treatment, the prize x is positive (representing gains) and subjects are asked to choose their most prefered combination of probability of winning and magnitude of the prize. In this setting, the standard formulation of revealed preferences and their implications hold and so we focus our analysis on that treatment. In the second treatment, the prize is negative (representing losses) and subjects are asked to choose their least prefered combination of risk and loss. Since the power properties of both treatments are quite similar, our analysis focuses on the first treatment, allowing us to forgo a discussion on the parallel axioms of revealed aversiveness.

[22] Andreoni and Miller (2002) report data on 176 subjects, but their session 5 is set aside here for brevity.

[23] Andreoni and Miller (2002) reported both the Bronars Method 1 power index and the panel index. We repeat them here for completeness.