A Kinky Consistency: Experimental Evidence of Behavior under Linear and Non-Linear Budget Constraints*

Emiliano Huet-Vaughn
Pomona College

Ethan M. L. McClure
UC Berkeley

Juan Carlos Suárez Serrato
Duke University & NBER

March 2021

Abstract

Individuals face non-linear budget constraints in myriad situations. We test a fundamental assumption of economic analysis in such settings: that individuals display stable preferences when facing linear and non-linear incentives. We use a laboratory experiment to characterize how revealed preferences are affected by changes in the complexity of the budget set environment. We find that choices under kinked (piece-wise linear and convex) budgets exhibit a similar degree of rationality as choices under linear budgets—with very high levels of internally consistent behavior in each setting. However, for about half the subjects, individual choices across settings are inconsistent with the maximization of a stable preference. Relative to those who act consistently across settings, subjects displaying such arbitrary consistency exhibit large and significant changes in utility parameters, risk premiums, and price elasticities across settings. Finally, we show that subjects with initially more sophisticated decision rules are most susceptible to changes in complexity.

*This research benefited greatly from the advice of Shachar Kariv, to whom we are extremely grateful. Thanks are also due to Alan Auerbach, Youssef Benzarti, Javier Birchenall, Henry Brady, Raj Chetty, Tom Davidoff, John Duffy, Jim Hines, Max Kasy, Botond Koszegi, Damon Jones, Rob MacCoun, Ulrike Malmendier, Daniel Martin, Katherine Meckel, Denis Nekipelov, Matthew Rabin, Dan Sacks, Emmanuel Saez, Alisa Tazhitdinova, Xiaxin Wang, Melanie Wasserman, and Philippe Wingender as well as seminar audiences at Chapman University, UCLA, UCSD, UC Berkeley, University of Queensland, and UTS and conference participants at the AEA, AEA Pipeline, AGEP, ESA, and NTA annual conferences. We benefited tremendously from the research assistance of Yamaç Isik. Remaining errors are our own. Financial support from an NSF IGERT fellowship, UC MEXUS, Mexico's CONACyT and SEP, and the UC Berkeley XLab is gratefully acknowledged. *Email:* Huet-Vaughn: `emiliano.huet-vaughn@pomona.edu`. McClure: `ethanmlmcclure@berkeley.edu`. Suárez Serrato: `jc.suarez@duke.edu`.

Individuals face non-linear incentives in myriad contexts including retirement savings decisions (e.g., employer contributions and the Social Security earnings test), tax schedules on labor supply (e.g., the earned income tax credit and progressive income taxes), and service rates that vary with usage (e.g., mobile phone contracts and electric power). A central assumption of economic analyses in these settings is that individual preferences, and the decision rules that they imply, are invariant to whether individuals face linear or non-linear incentives.

This paper tests this fundamental assumption and asks three inter-related questions. First, do the choices individuals make in non-linear settings satisfy axioms of rational choice? Second, do the choices individuals make when facing linear and non-linear incentives satisfy the same set of preferences? Finally, do measures of economic behavior—such as risk aversion and price responsiveness—depend on whether individuals face linear or non-linear incentives?

To answer these questions, we designed a laboratory experiment that elicits a large number of decisions from each individual subject in both linear and kinked budget sets (i.e., those that are convex with piece-wise linear constraints). Our experimental design builds on the work of Choi et al. (2007b), Choi et al. (2007a), Ahn et al. (2014), and Choi et al. (2014), who use a graphical interface that presents subjects with a series of randomly generated linear budget sets. We extend their toolkit to consider non-linear budget sets as well. The experiment generates rich, individual-level data that allow powerful tests of rationality and preference stability. The data also allow us to characterize how key features of utility and demand functions vary depending on whether individuals face linear or kinked budget sets.

Our first result is that the choices experimental subjects make when facing kinked budget sets display a high degree of rationality. Specifically, we test whether subjects' choices satisfy the Strong Axiom of Revealed Preference (SARP). Because a subject's choices either satisfy SARP or do not, we measure the degree of rationality in terms of the smallest number of choices that need to be removed for the remaining data to satisfy SARP. We find that the choices of nearly 75% of subjects satisfy SARP when we remove 4 or fewer choices (out of 50) and 90% of subjects satisfy SARP when we remove 6 or fewer choices. Indeed, we find the same result when analyzing decisions from linear budget sets. These findings confirm the result of Choi et al. (2007a) that experimental subjects display a high degree of rationality under linear budgets while also exploring the previously unanswered question of whether individuals choices remain rational when marginal incentives are no longer constant.

Our second result is that in spite of the fact that choices in kinked budget sets are close to being internally consistent, these choices are often not consistent with the choices subjects make when facing linear budget sets. We establish this result by testing for compliance with SARP in data that pool choices from linear and kinked budget sets. Using a cutoff value of 4 removals, we find that 46% of subjects have choices that, while consistent in either linear or kinked settings considered separately, are not consistent with the maximization of a single stable preference. This finding is robust to alternative cutoffs.[1] We call this type of behavior *arbitrary consistency*. While a minority of subjects make choices that are consistent within and across both budget settings, arbitrary consistency is the most common type of behavior in our data.

Our third result is that subjects displaying arbitrary consistency demonstrate different patterns of economic behavior in linear and kinked settings. We show this result by estimating a structural utility model for each individual in each treatment. The model is flexible enough to incorporate multiple behavioral features and does a good job of matching observed behavior from the experiment. We benchmark this estimation approach using both data from simulated subjects and data from subjects who display consistent preferences throughout. We find that relative to these fully consistent subjects, who see very small changes in estimates, the arbitrarily consistent exhibit large changes in structural parameters, risk premiums, and price elasticities when moving across budget settings. Interestingly, we find that the direction of change is not uniform: some subjects

[1]To ensure comparable and robust results, we hold the total number of observations in our pooled test at 50, and we repeat the test 10 times using different, randomly selected data from linear and kinked choices. The fraction of subjects displaying arbitrarily consistent preferences grows to 50% when we use a cutoff value of 6 removals. The fraction of subjects with consistent choices across settings is 14% and 30% when we use cutoff values of 4 and 6, respectively.

become more sensitive to price changes or risk-taking, and others become less sensitive. This result highlights the value of our experimental design, as it allows us to characterize individual-level changes in behavioral parameters that result from exposure to different types of incentive structures.

Overall, our results show that the fundamental assumption that preferences are stable across linear and non-linear budgets is violated by a large number of experimental subjects. Importantly, these violations are not driven by potentially careless mistakes. Indeed, we show that the presence of kinked incentives leads about half of subjects to adjust their behavior to different yet internally consistent decision rules. Moreover, these alternative decision rules are markedly different in both their responsiveness to price changes and the implied attitudes toward risk.

This paper contributes to several literatures. First, our finding of arbitrary consistency is related to work identifying context-dependent shifts between alternative yet apparently rational rules of decision making. Ariely et al. (2003) demonstrate evidence of this type of behavior in the coherent arbitrariness literature.[2] Our experiment contributes to this literature in two ways. First, by using explicit tests of rationality, our approach rules out an irrationality hypothesis, according to which the greater complexity of the kinked budget constraint leads otherwise-rational decision makers to cease behaving rationally. This hypothesis is ruled out by our finding that behavior in kinked settings is as internally consistent as behavior in linear settings. Moreover, our approach ensures that subjects switch between alternative and rational decisions rules, as opposed to between decision rules that merely appear rational. Second, our experiment documents systematic differences in revealed preferences along the foundational domain of whether incentives faced by decision makers are linear or non-linear.

A second set of related papers studies heuristic explanations that aim to explain irregularities of choices under non-linear incentives. One prominent hypothesis put forth by Liebman and Zeckhauser (2004) is that of "ironing," which has been shown to be relevant

[2] Ariely et al. (2003) use the term *coherent arbitrariness* for the general phenomena of context-dependent shifting between apparently rational decisions rules. We prefer the term *arbitrary consistency* and use it throughout.

in other contexts (e.g., Ito, 2014; Feldman et al., 2016; Rees-Jones and Taubinsky, 2016). Under this heuristic, individuals mentally average a set of marginal price schedules (each applying at different points in the commodity space range) and then act upon this average price. We explore this possibility by estimating models of demand based on average prices in the kinked budgets and show that this heuristic fails to explain our results. The data also rule out two additional alternative hypotheses. First, we rule out the possibility that subjects in the experiment who change their behavior are driven by a greater propensity to choose the kink point due to its salience. Second, we rule out the notion that arbitrary consistency is driven by choices that violate first-order stochastic dominance (FOSD).

While the ironing, salience, and FOSD hypotheses are rejected by the data, we provide evidence for an alternative hypothesis. Specifically, the evidence points to a complexity hypothesis in the spirit of Simon (1956). Under this hypothesis, the added complexity of the kinked choice environment causes rational actors with ordinarily price-responsive decision rules to switch to alternative—but still rational—decision rules. Importantly, this hypothesis would imply that subjects with initially less sophisticated or less price-responsive decision rules do not need to adjust their choices across settings. We offer evidence in favor of this hypothesis by showing that arbitrarily coherent subjects do indeed display more complex and price-responsive decision rules in the simpler linear budget setting than do subjects with fully consistent preferences. This mechanism is consistent with the work of Abeler and Jäger (2015), who find that behavior in more complicated tax incentive schemes fundamentally changes from behavior in less complicated but otherwise equivalent schemes.[3] In contrast to their work, our detailed individual-level data allow us to show that behavior in more complex environments can be rationalized by a utility function and that the behavioral changes brought on by complexity, while large in magnitude, do not systematically diminish responsiveness to complex price incentives.

Our paper is also related to a literature studying behavioral welfare economics. Bernheim and Rangel (2009) develop tools for welfare analysis under the assumption that ax-

[3]Unlike this previous work, our "complex" environment represents a minimal level of complexity (a single kink added to a linear budget set) that is presented in a very easily understood manner, and yet we still observe changes in individual behavior.

ioms of rational behavior may be separately satisfied across alternative decision-making frames. Similarly, Chetty et al. (2009) and Goldin and Homonoff (2013) show how manipulating the salience of sales taxes impacts tax elasticities of demand. In their approaches to welfare analysis, these papers rely on the assumption that individuals display arbitrary consistency across different settings. This paper bolsters the relevance of these approaches by providing individual-level evidence of the existence and prevalence of arbitrarily consistent behavior.

Finally, our findings have important consequences for the economic analysis of behavior under non-linear incentives in a variety of domains where non-linear pricing is common, including labor economics, public finance, health economics, and industrial organization, among others. We discuss some of these implications below, but here we note two important insights. First, our results suggest that caution is warranted in applying elasticity estimates derived from linear settings to non-linear settings (and vice versa). Second, our findings suggest that policy makers should appreciate how the choice to "kinkify" an incentive schedule can fundamentally change preferences, individual price responsiveness, and, therefore, the size of behavioral distortions.

The remainder of the paper is organized as follows. Section 1 outlines the experimental design. Section 2 reports measures of the internal consistency of individual choice data within and across treatments. We present a taxonomy of individual rationality types in Section 3, where we also estimate structural utility models to characterize individual utility parameters and estimates of risk premiums and price elasticities. Section 4 considers alternative explanations for the findings. Section 5 discusses the relevance of the experimental results for economic analysis, and Section 6 concludes.

1 Experimental Design

At a basic level, the laboratory experiment elicits choice data in two settings: a control setting, where incentives are represented by linear budget sets, and a treatment setting, where non-linear incentives are represented by kinked budget sets. We restrict our attention to non-linear sets of this form for several reasons. First, they represent an empirically prevalent case. Second, this case provides the minimal amount of complexity relative to the linear pricing case. Third, the methods for data analysis are familiar to economists. Finally, this type of budget set provides very clear incentives and, consequently, interesting, testable hypotheses.

Each subject in the experiment makes choices from 50 separate budget sets in each of the two treatments. Subjects make choices under risk. Individuals form a portfolio by choosing quantities of two securities that pay experimental tokens (3 tokens = 1 dollar) if the corresponding state of the world occurs and zero otherwise. The budget sets that individuals face in each treatment are randomly generated. At the end of the experiment, the computer randomly selects one of the 50 decision problems for each treatment and, with equal probability, randomizes which state occurs. This determines the payoff to the subject. The experimental design ensures that individuals have incentives compatible with making choices according to their risk preferences. The instructions given to experiment participants are included in Appendix A.

The decision problem is presented to individuals in graphical form, using an experimental interface developed by Choi et al. (2007b). The methods of Choi et al. (2007b) have been applied to different types of decisions (see, e.g., Choi et al., 2007a; Fisman et al., 2007; Ahn et al., 2014; Choi et al., 2014), demonstrating the versatility of the experimental interface. We extend this graphical budget set toolkit by allowing for kinked budget sets. An example of the interface for a kinked budget set case can be seen in Figure 1.[4]

The interface is particularly suited to our design because it allows us to elicit many decisions from each individual relatively quickly. Moreover, the graphical interface justifies the interpretation of the treatment in the experiment as providing full information about the price schedule. Indeed, one of the powerful features of this interface is that subjects have access to a great deal of information. For example, as the subject moves the mouse to select an allocation, the portfolio under consideration is displayed in three different parts of the interface (see Figure 1). Finally, this setting provides context-free decisions

[4]Appendix A provides an example of a linear budget set as displayed by the interface.

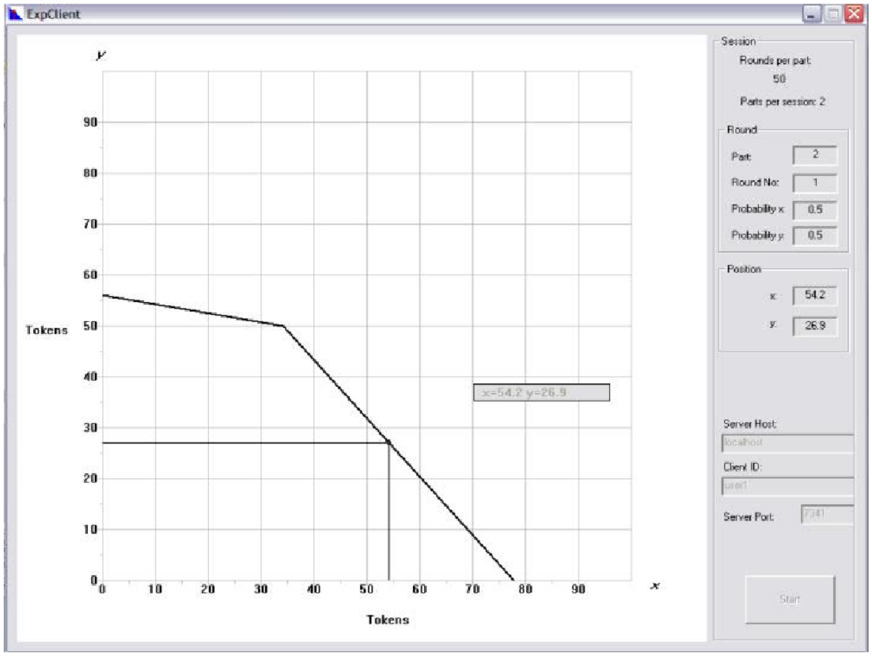
that make inference from the experimental data as free of outside influences as possible. All in all, then, the experimental design allows us to circumvent the usual constraints on empirical work and to attribute any observed differences in behavior across treatments to the form of the incentive structure alone, as we fully control the information sets of subjects in all treatment conditions, ensure that choices are not limited by confounding decision rigidities, and eliminate possible endogeneity from correlations between the form of marginal incentives that individuals face and unobserved characteristics.[5]

A particular decision in the design of the experiment is how to randomize the kinked sets. The linear treatment follows the experiment in Choi et al. (2007a) by selecting linear sets with at least one axis above 50 tokens and with both intercepts below 100 tokens. In the kinked treatment, each budget set is based on one from the linear treatment. Specifically, the interface randomly selects a linear budget from the linear treatment and generates a kinked budget set such that (1) both sets share the smallest intercept, (2) the new budget set is convex, and (3) the new budget set is downward sloping. In practice, the interface randomly chooses a kink point (x^K, y^K) in the linear budget set. Assuming, without loss of generality, that the x-intercept is greater than the y-intercept, the interface then randomly chooses a new intercept $x^{\text{Max Kink}} \in (x^K, x^{\text{Max Linear}})$. This process generates a kinked budget set that is a subset of the original linear budget set. By modifying the largest intercept, the experiment generates kink points that are likely to be located in relevant regions of the budget set for all participants.[6] In addition, in

[5]Even in field experiments that manipulate information sets of decision makers (i.e. "provide information"), researchers usually compare the *status quo* with a fully informed condition (see, e.g., Chetty and Saez, 2013; Jones, 2010). Without controlling the level of information in both treatments, it is difficult to ascribe differences in behavior under linear and non-linear incentives to the structure of the incentives alone and not to differences in the degree of information known about each incentive scheme. Additionally, in empirical settings, individuals might be constrained, for instance, in their labor supply decisions by labor market rigidities (see, e.g., Hoynes, 1996) as well as by complementarities with other decisions, such as in the choice of housing (see, e.g., Chetty and Szeidl, 2007), preventing researchers from studying maximization problems defined over the constraints of interest only. Finally, individuals might face incentives that are a function of characteristics that are unobserved to the econometrician. In the analysis of labor supply and taxation, for example, it is well understood that an individual's ability can potentially determine the tax bracket and marginal tax rate faced by the decision maker (see, e.g., Gruber and Saez, 2002).

[6]For instance, a risk-neutral participant would never choose the bundle represented by the smallest intercept. Altering the smaller intercept would lead to tests of rationality with low statistical power. Our design, by generating budget set differences across treatments in regions of the budget set that are

Figure 1: Experimental Subject Interface (with a Kinked Budget Set)



Notes: The figure presents the interface utilized by experimental subjects in each round of the experiment (with a kinked treatment example in this case).

cases where the choice in the linear set is present in the kinked set, our procedure allows us to test whether individuals choose the kink point solely due to its salience.

The experiment ran for 4 sessions at the UC Berkeley XLab, with a total of 142 subjects recruited. Subjects were a mix of students and staff at UC Berkeley.[7] The order of the treatment and control settings was reversed for two of the sessions.[8] Subjects received a show-up fee of $5 and payouts based on the choices they made, the choice round that was selected, and the realization of the state of the world that occurred. The payment was calibrated so that subjects were compensated at their estimated hourly wage of $15. Each session lasted around 1 hour and 45 minutes, and the average payment to subjects

more likely to contain individual choices, increases the likelihood of detecting differences in choices across treatments.

[7] In accordance with lab policy at the time, demographic characteristics were not collected. While this limits the possibilities of heterogeneity analysis, the within-subjects design employed here makes this immaterial to identification.

[8] The results are not affected by the order of the treatment.

was around $27.

2 Internal Consistency of Choice

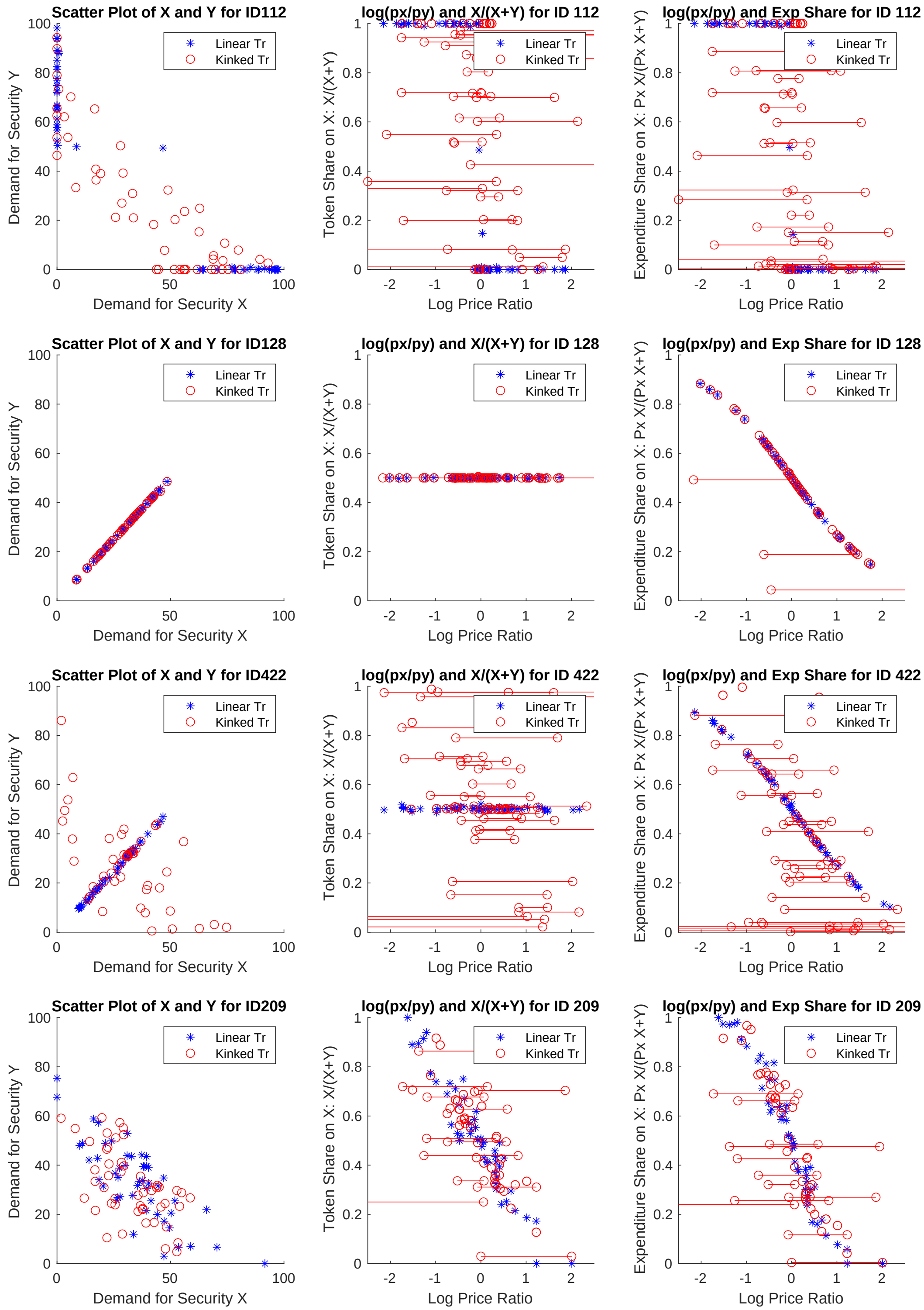
This section describes the data generated by the experiment, tests for internal consistency of the choice data within each treatment, and assesses whether decisions from both treatments can be explained with a common decision rule.

An initial step toward understanding the patterns of behavior elicited in the experiment is to visually analyze scatter plots of subjects' decisions. Figure 2 presents plots for four experimental subjects: IDs 112, 128, 422, and 209.[9] Each row plots the decisions made by each subject in three different ways. The first graph is a scatter of the decisions in each round. In this and all other graphs in the figure, the blue stars represent the choices made in the linear treatment, while the red circles represent the choices made in the kinked treatment. The second graph plots the token share of security X to the log price ratio (throughout the paper we use the natural log). The third graph plots the expenditure share on security X to the log price ratio. Given that the log price ratio is not properly defined at kink points, we plot the range of possible log price ratios at the kink as a line connecting two dots whenever a subject chooses the kink point. The length of the line contains information regarding the angle at the kink point. A longer line represents a smaller (interior) angle and more pronounced non-linearity.

The first two rows in Figure 2, corresponding to IDs 112 and 128, depict rational behavior and show patterns consistent with results from previous graphical budget set experiments. The first row plots the decisions of an individual (ID 112) whose choices are consistent with maximizing expected value in both the linear and the kinked treatment. In the linear treatment, this is evident from the fact that all choices correspond to placing all the tokens on the cheapest security. In the kinked treatment, ID 112 chose a significant number of decisions at the kink point, with virtually all other choices at the corner, as would be expected from an individual maximizing expected value. The second row plots decisions for an individual who is not price responsive at all when deciding the relative

[9] We include choice sets for all subjects in Appendix D of the Supplementary Online Material.

Figure 2: Scatter Plots of Decisions by Selected IDs



Notes: The figure contains three alternative plots of the raw data for four selected experimental subject IDs (equivalent plots for all other experimental subjects can be found in Appendix D of the Supplementary Online Material). Data from the linear treatment are in blue, and data from the kinked treatment are in red. The plot in the first column contains the subject IDs' chosen allocation (for each of the 50 choices) of tokens in the X account and tokens in the Y account. The plot in the second column puts on the y-axis the ratio of the tokens allocated to the X account over the total tokens allocated in each choice and puts on the x-axis the log of the ratio of the X token price to the Y token price. The plot in the third column keeps the same x-axis and replaces the y-axis with the expenditure share on X. Given that the log price ratio is not properly defined at kink points, we plot the range of possible log price ratios at the kink as a line connecting two dots whenever a subject chooses at or very close to (within a 0.5 token epsilon ball of) the kink point (longer lines thus indicate sharper kinks or smaller interior angles).

shares of X and Y to select: ID 128 chose to equate the demand for securities in each round of each treatment. The second column in the second row shows how the token share is constant regardless of price variation or (in the kinked treatment) location of the kink.

The next two subjects in Figure 2 exhibit choices that cannot be fully explained by such simple decision rules. ID 422, like ID 128, made selections at the 45 degree line in the linear treatment, as can be seen in the third row of Figure 2. However, in the kinked treatment, the decision rule that ID 422 used seems to have switched completely, with the subject no longer confining themself to equal shares of the two securities and instead frequently choosing off the 45 degree line. ID 209 (row 4 of Figure 2) made interior choices in all decisions, regardless of treatment. The second column suggests that these choices generated relatively smooth, continuous token share by log price ratio schedules and that this subject is responsive to price changes throughout the price space. While such visual analysis is informative, whether the patterns in each treatment are consistent with a coherent decision rule is not obvious from the choice plots alone (in the case of ID 209, we find that choices are not consistent across settings). Indeed, most subjects' decision rules are not so simple that they can be easily identified visually, and, as in the case of ID 209, they require more robust and exact measurements of both internal consistency and cross-treatment stability of preferences.

An informal definition of consistent preferences is that a subject's choices do not contradict each other. Formally, this is embodied by the Generalized Axiom of Revealed Preference (GARP) of Afriat (1967) or the stronger condition of SARP. Afriat's Theorem states that if choice data satisfy GARP over linear budget sets, then the choice data can be rationalized as the maximization of a continuous, strongly increasing, and concave utility function (and vice versa). Matzkin (1991) and Forges and Minelli (2009) generalize Afriat's Theorem to non-linear budget sets. To ascertain whether the choice behavior elicited in the laboratory is rational, we thus test whether the choice data satisfy SARP. This approach is very robust, as it is purely non-parametric.

Choice data either satisfy or violate SARP. However, as most decision makers exhibit less than perfect rationality when making a large number of choices, several methods have been developed to quantify the degree of deviation from SARP. Afriat (1972) proposed the critical cost efficiency index (CCEI), which measures the amount by which each budget set would have to be relaxed for the data to be consistent. A well-known alternative approach to measuring deviations from rationality is to find the largest set of data that is internally consistent. In this vein, the minimal number of choices that have to be removed for the data to be consistent was first proposed as a measure of the distance from rationality by Houtman and Maks (1985, henceforth HM).

The analyses in this paper focus on the HM measure over the CCEI measure for three reasons. First, the HM measure is conceptually simpler and lends itself to clear interpretation. Second, prior research by Choi et al. (2007b) that we build upon deemphasized the HM measure in part because of computational difficulties. Recent advancements by Dean and Martin (2016), who provided an improved algorithm to find the maximal consistent subset of choice data, have since solved this problem, greatly reducing computation time. The third reason is specific to the particular kind of deviation from rationality that this experiment is interested in analyzing. We set out to compare violations of rationality among choices made from a set of linear budgets, violations from a set of kinked budgets, and violations from across the two sets. The CCEI would be systematically different if we compare violations originating from a set of linear budget set choices to otherwise-equivalent violations originating from a set of both linear and kinked budget set choices (our pooled case), whereas the two violations are treated equivalently under the HM measure. This makes the CCEI an inappropriate and incommensurable measure of rationality when the purpose is to compare levels of rationality for choices made facing linear budget sets only and, alternatively, choices made facing both linear and kinked budget set choices (as is our purpose).[10] The analysis of the experimental data therefore focuses on the HM measure rather than the CCEI score so that violations of rationality within each treatment contribute to overall measures of rationality in the same way that violations across treatments do, allowing for meaningful comparisons. Nonetheless, we report the

[10]Figure 16 in Appendix E of the Supplementary Online Material gives examples of two GARP violations that illustrate this point.

CCEI measures for each subject as well, and we note broad correspondence with the HM measure.[11]

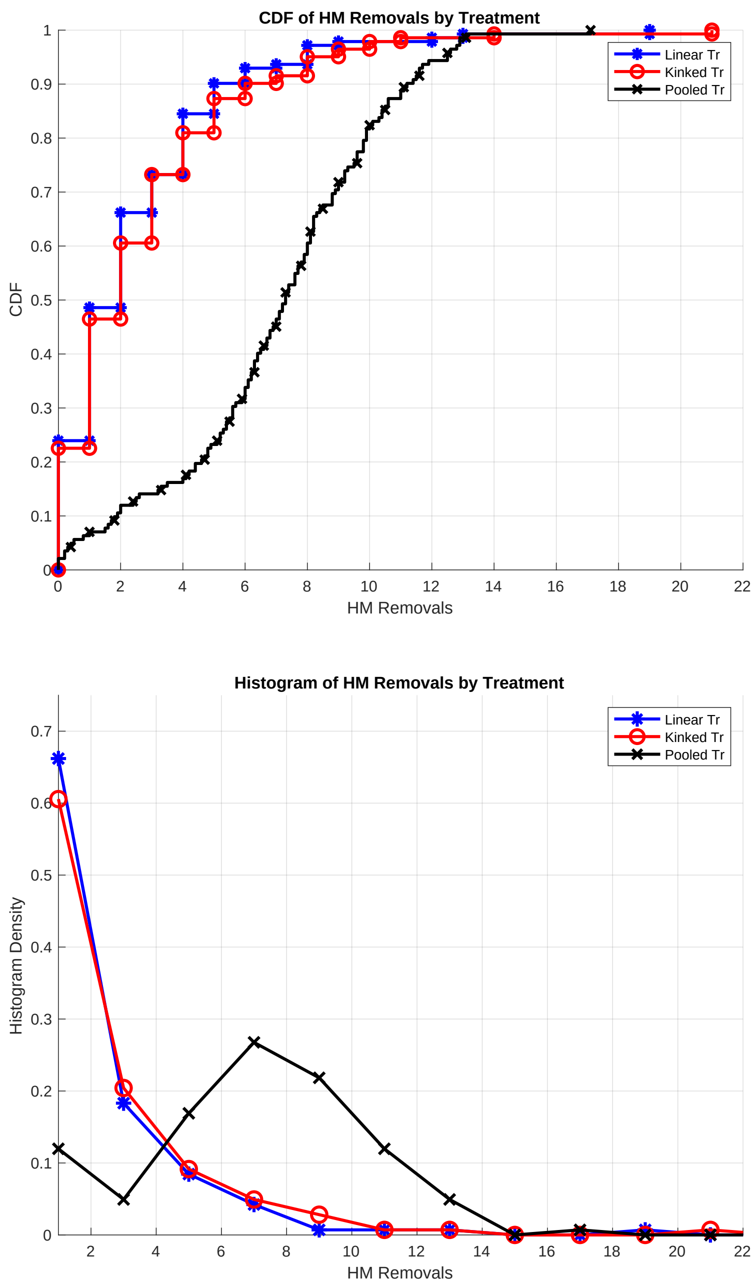
Figure 3 plots the distribution of the HM measure by treatment. This graph answers the first research question by showing that choices in kinked sets follow similar patterns of rationality as those in linear sets. This notion is formalized by testing whether these two distributions (the red and the blue lines) are different in a statistical sense. A Kolmogorov-Smirnoff (KS) test does not reject the hypothesis that these distributions are the same with a p-value of 0.87. In addition, most subjects exhibit behavior that is close to rational, as the number of deviations that have to be removed is very small. For example, in both treatments, removing 6 or fewer choices would leave consistent data for about 90% of subjects. An even more stringent rationality criteria—removing 4 or fewer choices—would leave consistent data for about 75% of subjects.

To analyze whether behavior in both settings can be attributed to a decision rule that aims to satisfy the same set of preferences, we compute the HM measure pooling data from both treatments. As the HM measure depends on the number of decisions, we construct the pooled measure by taking 25 randomly selected observations from each treatment. The measure reported is the mean of 10 repetitions of this process.[12] The HM measures for the pooled data show a significant increase in deviations from SARP. KS tests comparing the distributions of the HM score for the pooled data and the HM score for the data from either treatment reject the null hypothesis of equal distributions at all conventional levels of statistical significance. This evidence suggests that while subjects' decision rules are rationalizable in either treatment alone, these decision rules differ considerably. This is strong evidence that some subjects display preferences that

[11]See Appendix E of the Supplementary Online Material. Notably, the pattern described below—of less consistent behavior across choices pooled from linear and kinked treatments than from either the linear or the kinked treatment choices analyzed alone—is robust to use of the HM measure or either Afriat's or Varian's CCEI measure as the metric for consistency. In the HM case, the pooled distribution of HM scores shifts rightward. In the CCEI case (whether Afriat's or Varian's version), the distribution shifts leftward, indicating that the result of a lesser degree of rationality in the pooled choices is robust across these measures.

[12]The mean of 10 repetitions of the process is chosen to guard against the possibility of randomly drawing a set of particularly "odd" and hard-to-reconcile choices on a given draw. However, the cumulative distribution function (CDF) of HM removals based on any one of these 10 draws in practice is nearly identical to the reported CDF in Figure 3 for the (10 repetition mean) pooled sample.

Figure 3: HM Measure by Treatment



Notes: The figure plots the distribution (as CDF and histogram) of individuals' minimum number of HM removals (out of 50 choices) needed to make their choice data consistent within each treatment (linear in blue, kinked in red) and for the data pooled across treatments (in black).

are arbitrarily consistent.

The results from this section show that most subjects are close to being consistent in their choices in both the linear and kinked treatments. This means that when we analyze the choices made in either treatment, an individual's behavior in that treatment can be rationalized by a utility function (after excluding a small number of deviant choices). Decisions pooled from the linear and kinked treatments, however, are far from being consistent for the typical individual. Taken together, these results provide evidence that decision rules are different in each treatment, with the non-linear nature of kinked budget sets appearing to generate arbitrarily consistent decision making. We now characterize the heterogeneity in the type of behavior observed.

3 Characterizing Rationality, Risk Preferences, and Price Responsiveness

The results from the previous section show that while subjects display mostly internally consistent behavior in each treatment, their choices cannot easily be rationalized across treatments. This section explores heterogeneity in behavior by classifying individuals by rationality type. We then explore how the change in behavior from linear to kinked budget set environments impacts estimates of structural utility parameters, risk premiums, and price elasticities.

3.1 Taxonomy of Rationality Types

Table 1 presents a type taxonomy. Types 1-3 are individuals whose choices are not internally consistent in at least one of the treatments. It follows that for each of these types, the joint set of data between the two treatments would also be inconsistent. Type 4 corresponds to the group of individuals who are arbitrarily consistent (consistent in each treatment but not across treatments). Individuals of Type 5 correspond to the traditional model of rational behavior, being fully rational in each treatment *and* when we combine data from the two.

As there is no natural (non-zero) threshold of HM removals below which we can say

Table 1: Rationality Types

Type	Linear Treatment	Kinked Treatment	Pooled Treatment
1 (Never Consistent)	Not Consistent	Not Consistent	Not Consistent
2 (Linearly Consistent)	Consistent	Not Consistent	Not Consistent
3 (Non-Linearly Consistent)	Not Consistent	Consistent	Not Consistent
4 (Arbitrarily Consistent)	Consistent	Consistent	Not Consistent
5 (Coherently Consistent)	Consistent	Consistent	Consistent

Notes: The table presents the classification of rationality types used throughout the paper. Consistent behavior in each treatment (and across the two treatments) is defined by individual choice data that require 6 or fewer HM removals in the treatment data (or, if across treatments, in the pooled set taking 25 observations from each treatment, as described in the text) for the individual's remaining choices to be internally consistent with maximization of some utility function.

that subjects are close enough to being utility maximizers, we develop a statistical test that compares observed behavior with a benchmark of simulated choices to categorize individuals according to the taxonomy in Table 1.[13] Specifically, we first generate 1300 simulated subjects who maximize a constant relative risk aversion (CRRA) expected utility function subject to logistic taste shocks. We obtain a variety of tests by varying the relative importance of the taste shock, which is determined by the parameter γ. For very large values of γ, choices are close to fully random. In this case, with 50 choices per simulated subject, only about 5% of simulated subjects meet an HM critical value threshold of less than 12. That is, simulated behavior (following 12 or fewer choice removals) would appear consistent in spite of the fact that the choices were completely random. Alternatively put, an actual decision maker could be said to have better than fully random behavior at the 95% confidence level when their HM measure is less than 12.

Since Figure 3 shows that all subjects have behavior that would reject the fully random benchmark, we instead consider a simulated subject who maximizes a CRRA utility function with $\rho = 1/2$ (following Choi et al., 2007b) subject to a more modest logistic taste shock ($\gamma = 10$) that implies a much closer approximation to expected utility max-

[13]Our process is analogous to the benchmarking procedure that Choi et al. (2007b) conduct for the CCEI measure with linear budget sets and generalizes the test of Bronars (1987). We conduct the benchmarking analysis for the HM measure in both the linear and kinked sets in Appendix B.

Table 2: Type Proportion and Average HM Measure by Type

Type	Proportion	HM Measure Linear Tr	Kinked Tr	Pooled Tr
1 (Never Consistent)	13.4%	6.79	8.05	10.48
2 (Linearly Consistent)	13.4%	1.53	5.16	8.35
3 (Non-Linearly Consistent)	13.4%	5.21	0.95	7.11
4 (Arbitrarily Consistent)	45.8%	1.20	1.52	7.43
5 (Coherently Consistent)	14.1%	0.40	0.55	1.39

Notes: The table presents the proportion of individuals of each type in the rationality type taxonomy (see Table 1) and the average HM score across individuals of each type based on decisions in each treatment and, for the pooled observations, on a procedure of randomly sampling 25 observations from each treatment (with an individual's HM score in the pooled case being the average HM score computed across 10 repetitions of this random sampling procedure, as described in the text).

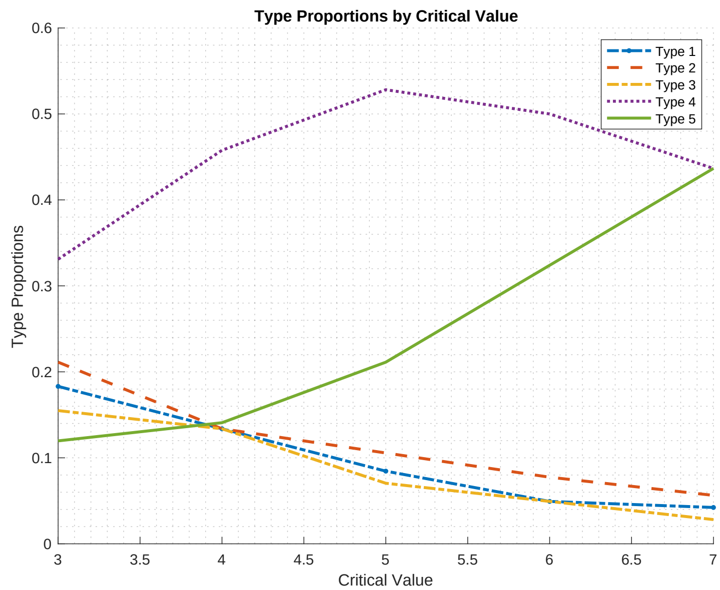
imization and consistent preferences. In this case, the critical value for a test with 95% confidence level is 4. In what follows, we use an HM critical value of 4 as the baseline cutoff to distinguish between *Consistent* and *Not Consistent* data.[14]

The second column of Table 2 shows the type distribution at this critical value. The proportions of Types 1-3 are individually small, as is the proportion of the next largest group, Type 5s. The most striking feature of Table 2 is that almost half the subjects correspond to Type 4, the arbitrarily consistent group, demonstrating rational behavior in both treatments that is nonetheless not stable or consistent across treatments. Table 2 also presents the average HM score by treatment and type. The HM measures by type fit the intuition behind the taxonomy. Members of the arbitrarily consistent (Type 4s) group, for instance, have low HM measures for both treatments and high measures for the pooled data, while those in the coherently consistent (Type 5s) group have low HM measures across the board and thus display rational behavior in the three conditions.

The fact that Type 4 individuals are the most numerous is robust to the use of alternative critical values. Figure 4 shows the proportions across types for different choices of the critical value. This figure shows that the proportion of arbitrarily consistent (Type 4) individuals is approximately a third, and the largest share among all types, at a critical

[14]The full range of results from this benchmarking procedure, for different values of ρ and γ, are presented in Appendix B. Appendix B also reports the type assignment of each individual subject.

Figure 4: Robustness of Type Distribution to Critical Value



Notes: The figure shows the sensitivity of rationality type classification (Types 1- 5, as in Table 1) to variation in the critical value (employed to distinguish consistent from inconsistent behavior) in an individual's HM measure (Houtman and Maks, 1985). The HM measure indicates the minimum number of choices that must be removed from the data for it to be internally consistent, and the HM critical value cutoff indicates the HM measure below which we characterize an individual's choice data as consistent.

value of 3 and then quickly rises to between roughly 45% and 55% of the population when the critical value is between 4 and 7. The proportion of the coherently consistent (Type 5s) group grows as the critical value increases (a mechanical effect of lowering the bar of rationality since if enough mistakes are allowed to be removed, everyone eventually becomes a Type 5), reaching a third of the population at a critical value of 6 and tying with Type 4s once a higher critical value of 7 is chosen. This growth, however, is mostly due to the decrease in the proportions of Types 1-3 rather than re-classification of arbitrarily consistent individuals.

In summary, we categorize individuals by rationality type using a statistical test designed to compare the value of the HM measure to that of a rational expected utility maximizer subject to taste shocks. Using the results of this test to categorize individuals by rationality type, we find that close to half of the subjects in the experiment exhibit choice data that correspond to rational behavior in each treatment but that do not maximize the same utility function across the linear and kinked treatments.

3.2 Risk Preferences

We now characterize how the cross-treatment inconsistency that characterizes our main rationality type—Type 4—impacts measures of economically relevant behavior across treatments. To do so, we estimate individual demand functions via a structural utility model, yielding individual utility parameter estimates and resulting risk premium and price elasticities. We characterize these values separately for the linear and kinked treatments, and we study the change in these values for each individual subject. Across these measures, we find economically and statistically significant changes in behavior relative to a benchmark of the corresponding changes among subjects who act consistently across treatments (Type 5s). As expected, these latter subjects exhibit essentially no changes in utility parameters or in measures of risk or price responsiveness.[15]

3.2.1 Structural Utility Model

Choi et al. (2007a) note that choices in the kind of graphical budget set experiment employed here display heaping at the certain outcome (as is the case in our sample), which is consistent with a model of loss aversion. As a result, they propose a structural model of demand employing a utility function that can accommodate both loss and risk aversion, as first proposed by Gul (1991).[16] This function has the following form:

$$U(X,Y) = \min\{\alpha V(X) + V(Y), V(X) + \alpha V(Y)\}, \tag{1}$$

where $V(X) = X^{1-\rho}/(1-\rho)$ is a constant relative risk aversion (CRRA) utility function. We estimate the parameters in Equation (1) with an additional multiplicative stochastic

[15] Subjects of Types 1-3, in addition to being small in number, also fail to display basic consistency in one or both of the treatments, as we show in Section 3.1, making the structural estimations (which are based on utility-maximizing assumptions) for such subjects uninformative. We thus ignore them for these analyses.

[16] This model is general enough to encompass both expected utility theory (EUT) and other alternatives as specific instantiations.

component using a weighted non-linear least square (NLLS) procedure.[17]

The model in Equation (1) characterizes behavior through the Arrow-Pratt measure or relative risk aversion parameter ρ and a parameter for loss aversion α. Intuitively, a larger value of ρ increases the curvature of a subject's indifference curve. A larger value of α leads to a more pronounced kink in the indifference curve at the certain outcome, which increases the likelihood of observing a choice where $X = Y$. As a way to summarize the overall risk tolerance of a given subject, we also report a calculated risk premium, $r(1)$, that depends on both parameters.[18]

We first consider how these structural parameters differ across subjects categorized as Types 4 and 5.[19] In the linear treatment, the average α value of Type 4 subjects is significantly smaller (p-value of 0.001) than that of Type 5 individuals. The average ρ of Type 4 subjects is also larger, but this difference is only marginally significant (p-value of 0.16).[20] These results suggest some differences in risk preferences across subjects: Type 4 subjects are, on average, less loss averse but more risk averse. When computing the joint

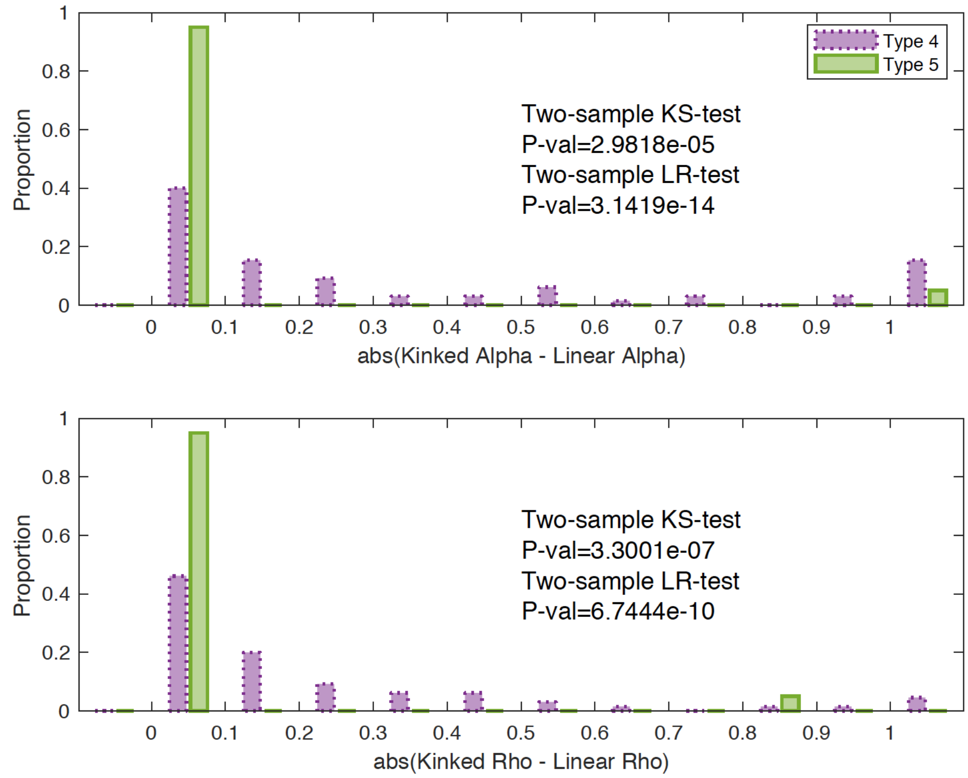
[17] Appendix C contains the full details of the estimation procedure. This appendix also shows that our procedure performs better in recovering the structural parameters of simulated subjects than an alternative maximum likelihood estimation approach. This procedure is also more robust at finding similar estimates in both linear and kinked settings, which is crucial given our objective of measuring changes in structural parameters across settings. Our estimation constrains the value of α based on the range of prices faced by a given subject (for reasons explained in Appendix C). Since the price range is larger in kinked settings, we impose the constraint based on the kinked setting in both the linear and the kinked estimations for each individual. Finally, we heed the warning in Choi et al. (2007a) that structural estimation can be sensitive to outliers by removing a small number of extreme outliers using quartile outlier detection. While Choi et al. (2007a) exclude from their analysis *participants* whose structural parameters are sensitive to outliers, we take a more systematic approach, algorithmically removing *choices*. This procedure affects few choices, with the modal and median number of outliers removed for an individual in the linear and non-linear treatments being zero and the mean being less than 1 in both treatments.

[18] The risk premium $r(h)$ for a 50-50 gamble of winning vs. losing $h*100\%$ of one's wealth (with h being between 0 and 1) is equivalent to the fraction r of current wealth that, if lost for certain, would make the individual indifferent between having the remaining $(1-r)$ of their current wealth and taking the gamble. For the utility function in Equation (1), the second-order approximation of $r(h)$ when $h = 1$ is $r(1) \approx \frac{\alpha-1}{\alpha+1} + \frac{\rho 2\alpha}{(\alpha+1)^2}$. We follow Choi et al. (2007a) by computing this summary measure of risk aversion for each subject at their estimated values of α and ρ. In the case where $\alpha = 1$ (i.e., EUT), $r(1) = \rho/2$.

[19] Appendix G of the Supplementary Online Material contains the full set of results at the individual level, including α, ρ, the risk premium, and elasticities (discussed in Section 3.3) estimated for each subject based on their choices in each of the two treatments (see Tables 19-26).

[20] The same pattern holds in the kinked treatment, with both p-values being below 0.05. Tables 27–28 in Appendix G of the Supplementary Online Material present average values of these parameters and risk premiums by treatment for Types 4 and 5.

Figure 5: Magnitude of Individual Change in Structural Parameters Histogram by Type



Notes: The figure presents the histogram (separately for Types 4 and 5) of the magnitude of change (across the linear and kinked treatments) in individual estimates of α (top panel) and ρ (bottom panel). Extremely large magnitude changes (in excess of 1) are grouped together in the last bin for presentation purposes. KS test p-values are reported for the null hypothesis that the Type 4 and Type 5 distributions in each panel are the same. LR test p-values are reported for the null hypothesis that the Type 4 and Type 5 sample means in each panel are the same.

risk premium measure $r(1)$, however, we find that both types of subjects have similar risk premiums. Estimates using choices from the kinked treatment reveal similar qualitative results for both α and ρ.

3.2.2 Individual-Level Differences

One of the advantages of eliciting multiple decisions for each individual is the ability to measure individual-level changes in estimated utility parameters across the linear and kinked treatments. However, as in known in the literature, (e.g., Schildberg-Hörisch, 2018; Barseghyan et al., 2018), it is not unusual to find some degree of within-subject variation across samples. For this reason, we compare the changes in parameter values for Type 4 subjects to the changes for Type 5 subjects, who are known to be implementing a consistent decision rule across settings. This allows us to separate differences in behavior that may arise from switching between alternative consistent decision rules from those that may be attributable to imprecision in measurement. We therefore test whether the changes in the measured behavioral parameters of Type 4 subjects are significantly greater than whatever shifts may be measured for Type 5 subjects. As these latter shifts—to the extent that they are measured—are solely driven by imperfect measurement, this benchmarking against Type 5s serves as a placebo test of sorts.

Figure 5 presents the distribution of the individual-level differences across linear and kinked settings (in absolute value) of the estimated parameters α and ρ for both Type 4 and Type 5 subjects. For presentation purposes, this and related figures group large magnitude changes (in excess of 1) in the last bin. For both α and ρ, it is clear that the vast majority of Type 5 subjects see little or no change in estimated parameters. In contrast, a large fraction of Type 4 subjects see large changes in both parameters. To formalize this comparison, we conduct KS tests between the parameter distributions of Types 4 and 5. In both cases, we can reject the null hypothesis that the changes for Types 4 and 5 are drawn from the same distribution with p-values smaller than 0.001.[21,22]

We find comparable results when we analyze changes in our risk premium measure, $r(1)$: a KS test indicates that we can reject the null that the magnitude of change in

[21]As an alternative test, we use the fact that these series follow an exponential distribution (visible for Type 5s when zooming in near zero) to conduct a test of sample means between Types 4 and 5. Figure 5 reports the result of these F tests on the null hypothesis that the parameter change magnitudes for Types 4 and 5 are distributed with the same exponential distribution (i.e., that their distribution means are the same). For both α and ρ, we can reject this null with a p-value smaller than 0.001, concluding that the much larger average change in individual-level α and ρ estimates for Type 4 subjects is a statistically significant difference (for reference, the average/median change magnitude in α for Type 4 is 0.746/0.149 vs. 0.064/0.000 for Type 5, and the average/median change magnitude in ρ for Type 4 is 0.351/0.129 vs. 0.053/0.001 for Type 5).

[22]It is worth pointing out that the non-zero changes for Type 5 are driven by subjects who exhibit near-EUT preferences. Specifically, an EUT maximizer would choose corners in the linear setting and choices at the kink in the non-linear setting. Such is the case, for example, of ID 322, a Type 5. However, a small number of deviations from EUT in the non-linear setting dramatically increases the estimated ρ. For this reason, the non-zero changes for Type 5 are mostly driven by imprecision in the measurement of structural parameters.

$r(1)$ across treatments for Type 4 is drawn from the same distribution as the change for Type 5 (p-value of less than 0.001).[23] To appreciate the size of these changes, note that the average magnitude change for Type 4s is 0.197, which is 41.9% of the average in the linear treatment. In contrast, Type 5s see a smaller average change of 0.041, which is a 10.5% difference from their linear-treatment average of 0.39.

In summary, the shift from a linear to a kinked budget environment leads the arbitrarily consistent group to display choices that imply different utility parameters. Importantly, the measured changes are above and beyond those (very small) changes measured in a benchmark population (Type 5s) that makes consistent choices across budget environments. While the small changes observed in the latter group can be attributed to imprecision in measurement, Type 4s exhibit a magnitude of parameter change that is significantly above and beyond what might be ascribed to an artifact of the estimation procedure.

It is also worth noting that while we find robust evidence that Type 4 subjects change their behavior in economically meaningful ways across the linear and kinked settings, there is substantial heterogeneity in the direction of change in risk attitudes. While some Type 4 subjects become more averse to risk, others become less so. On average, there is no precise change in either direction. This result showcases the value of experimental designs that characterize individual-level changes in behavior, as these would otherwise be missed in the aggregate.

3.3 Price Responsiveness

While the changes in structural parameters confirm the non-parametric result that Type 4 subjects significantly change their behavior between linear and kinked settings, the changes do not directly characterize how economic behavior would differ across these two settings or how these changes in behavior depend on price differentials. To further describe the change in behavior of Type 4 subjects, we now consider how the above changes to

[23] An F test of the null hypothesis that the absolute value changes in $r(1)$ are drawn from the same exponential distribution for Types 4 and 5 also indicates a rejection with a p-value smaller than 0.001. See Figure 21 in Appendix G of the Supplementary Online Material.

structural parameters affect the price responsiveness of individual demand functions.

Using the estimated structural parameters, we can recover for each subject and each treatment the log demand ratio $\log(x/y)$ as a function of the log price ratio $\log(p_x/p_y)$. The elasticity of this demand with respect to the relative price ratio then reveals the price elasticity of substitution between the two commodities. As this elasticity may take different values depending on the price ratio, we first present the results of evaluating the elasticity at two salient price ratio values, $\log(p_x/p_y) = 0$ and $\log(p_x/p_y) = 1$, computing the size of the change in elasticity across treatments.

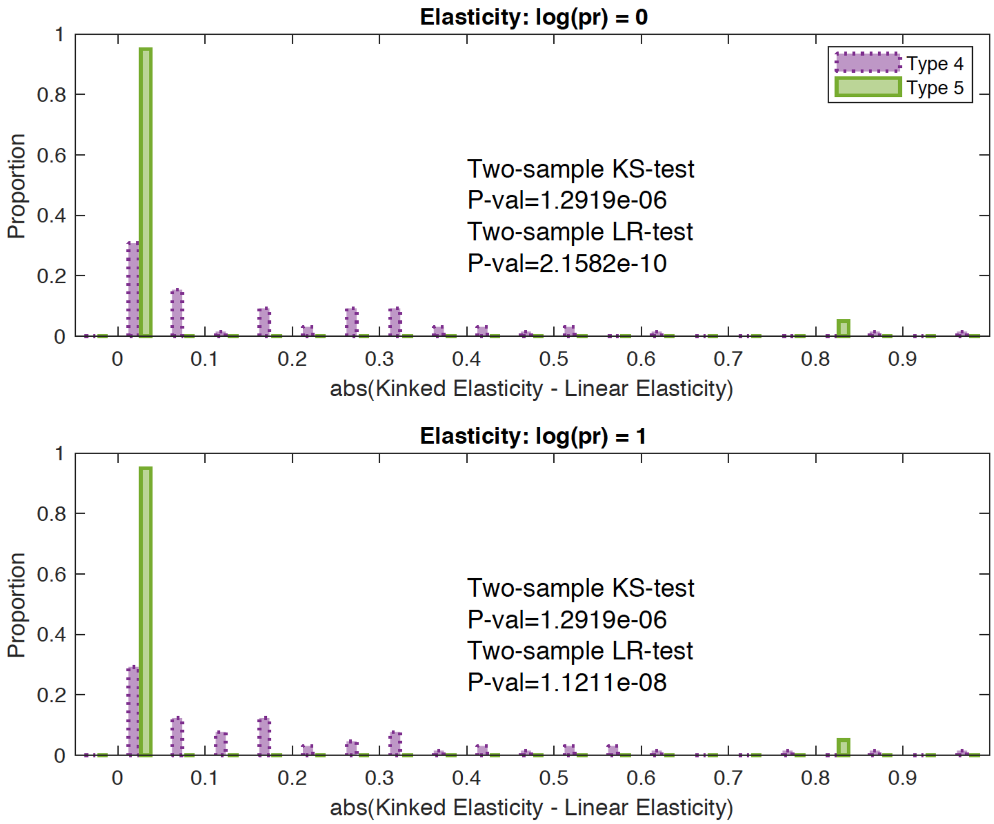
Figure 6 presents the distribution of the elasticity change magnitudes per individual across treatments for both Type 4 and Type 5 subjects, with the elasticity evaluated at $\log(p_x/p_y) = 0$ in the top panel and $\log(p_x/p_y) = 1$ in the bottom panel. At both prices, KS tests indicate that we can reject the null that the magnitude of change in elasticities across treatments for Type 4 is drawn from the same distribution as the change for Type 5 (p-value of less than 0.001 in the top and bottom panel).[24] At both price ratios, the much larger average change in individual-level elasticity estimates that is observed for Type 4 subjects indeed represents a statistically significant difference in price responsiveness brought about by the switch in the budget environment.

The results in Figure 6 further show that Type 4 subjects display economically different behavior in the linear and kinked settings. To obtain a more holistic description of the change in estimated demand functions, we now consider differences in price responsiveness across the full range of prices. Specifically, for each subject i, we calculate elasticities, using the subject's estimated structural parameters, at 100 evenly spaced log price ratios to form vectors of price elasticities for the linear budgets ($\vec{e^i_L} = e^i_L(\log(\vec{p}))$) and non-linear budgets ($\vec{e^i_K} = e^i_K(\log(\vec{p}))$), respectively.[25] Figure 7 presents examples of these elasticity estimates for selected individuals in each of the two budget treatments.

[24] An F test of the null hypothesis that the elasticity change magnitudes for Types 4 and 5 are distributed with the same exponential distribution (i.e., that their distribution means are the same) allows us to reject the null with a p-value smaller than 0.001 in both panels as well (for reference, in the top panel, the average/median change magnitude for Type 4 is 0.381/0.173 vs. 0.054/0.005 for Type 5; in the bottom panel, the average/median change magnitude for Type 4 is 0.298/0.159 vs. 0.054/0.005 for Type 5).

[25] The results are robust to using a higher number of price points.

Figure 6: Magnitude of Individual Change in Elasticity Histogram by Type

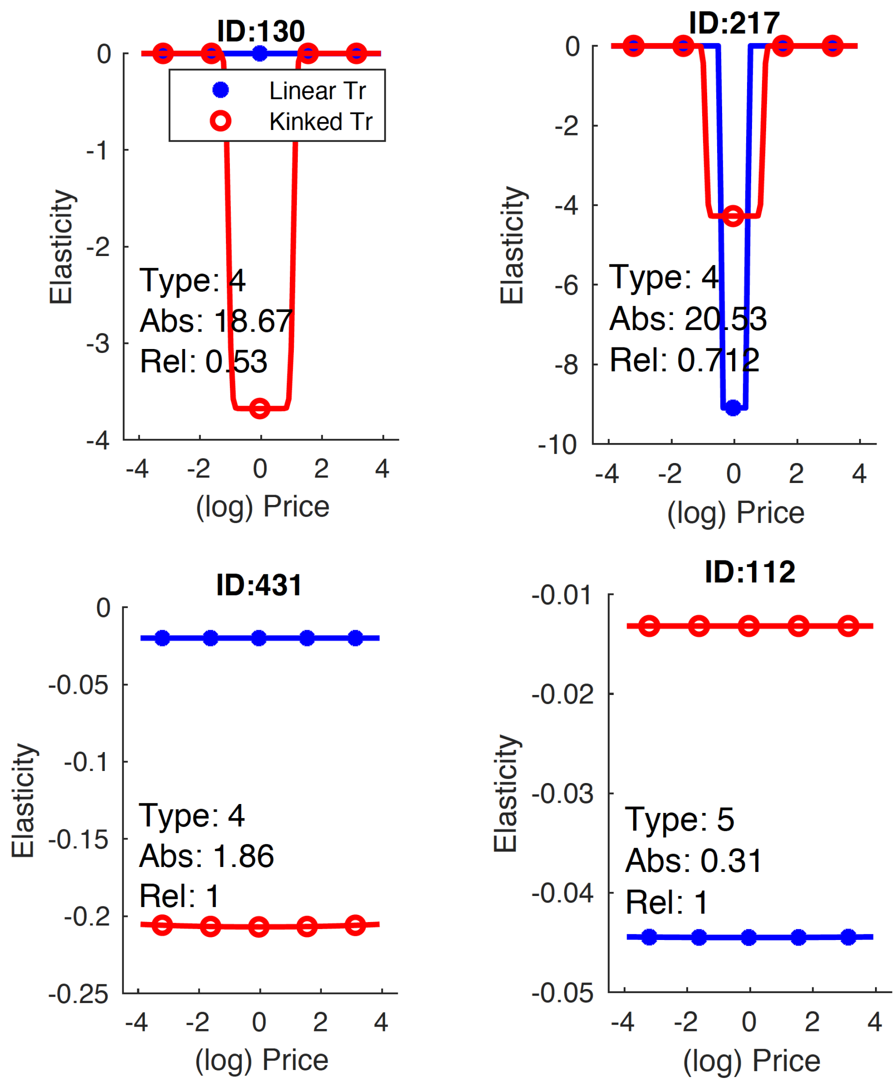


Notes: The figure presents the histogram (separately for Types 4 and 5) of the magnitude of change (across the linear and kinked treatments) in individual elasticity estimates with elasticity evaluated at natural $log(p_x/p_y) = 0$ in the top panel and natural $log(p_x/p_y) = 1$ in the bottom panel. Extremely large magnitude changes (in excess of 1) are grouped together in the last bin for presentation purposes. KS test p-values are reported for the null hypothesis that the Type 4 and Type 5 distributions in each panel are the same. LR test p-values are reported for the null hypothesis that the Type 4 and Type 5 sample means in each panel are the same.

To quantify the change in behavior across settings, we consider two notions of distance: the Euclidean distance and the angular distance. The Euclidean distance $d^i = ||\vec{e^i_L} - \vec{e^i_K}||_2 = \sqrt{\sum_{p=p_{min}}^{p_{max}} (e^i_L(\log(p)) - e^i_K(\log(p)))^2}$ measures the distance in N-th dimensional space between the two price elasticities using the 2-norm.[26] This measure captures the *absolute* change in price elasticities across the spectrum of log prices. In contrast, the

[26]The Euclidean distance can also be thought of as measuring the root mean square error (RMSE) between the elasticities on the linear and non-linear budgets.

Figure 7: Examples of Kinked and Linear Elasticities for Selected Subjects



Notes: For each of the selected subjects, the figure presents the results of the following procedure: using the subject's estimated structural parameters, we calculate demand functions and then elasticity values when the elasticity is evaluated at 100 evenly spaced log price ratios $\log(p_x/p_y)$ within the range of (-log(3.92),log(3.92)), the full range of log price ratios common to all subjects. This is done separately for elasticities derived from the linear budget choices and elasticities derived from the non-linear budget choices. These values are then plotted (and connected by a line) to offer a full description of how the elasticity of the log demand ratio with respect to the log price ratio changes for a subject when moving across budget treatment conditions. The Euclidean distance and angular distance measures of the subject-specific degree of change in elasticity across treatments (averaged across the price range) is also presented (see the text for further details).

angular distance measures the *relative* difference between the price elasticity vectors. For each subject, the angular distance, also known as the cosine similarity, between the two price elasticity vectors is defined as $c_i = \cos\theta_i = \frac{\vec{e_L^i} \cdot \vec{e_K^i}}{||\vec{e_L^i}||_2 ||\vec{e_K^i}||_2}$.[27] For our purposes, this metric measures the change in relative price elasticities, with a value of $c_i = 1$ ($\theta_i = 0$) implying that the price elasticity vectors are parallel and point in the same direction (the relative price elasticities remain unchanged) and a value of $c_i = 0$ ($\theta_i = \pi/2$), implying that they are orthogonal (the relative price elasticities change substantially).

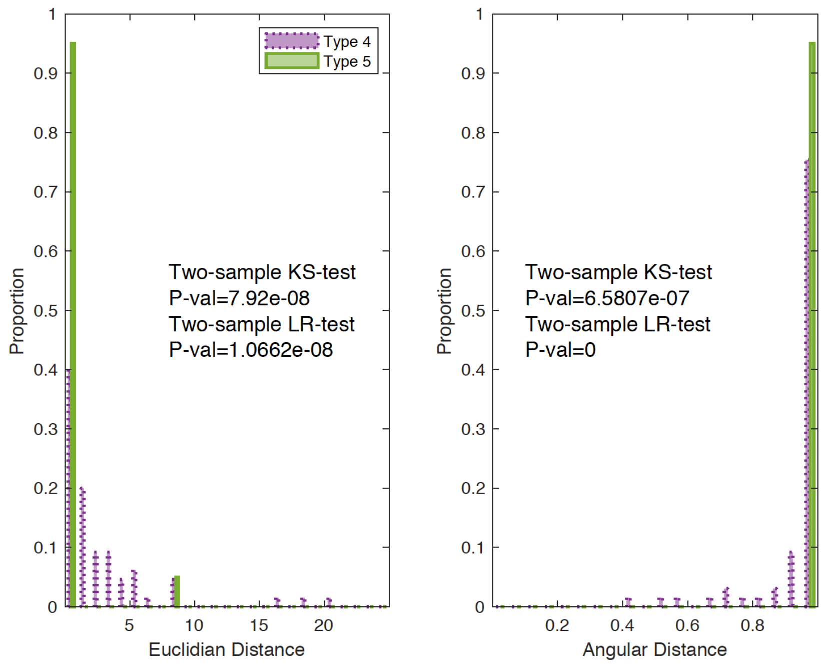
These two metrics of distance capture different, though not mutually exclusive, measures of the change in price elasticities for subjects across the full range of prices. For instance, ID 431 in Figure 7 illustrates a case where the Euclidean distance is non-zero, indicating a difference in the absolute price elasticities between the linear and non-linear budgets, while the angular distance is unity, indicating no change in the relative price elasticities between the linear and non-linear budgets. Comparing ID 130 and ID 217 in the figure, we see that ID 130 has a smaller Euclidean distance and thus a smaller absolute change in price elasticities between the linear and non-linear budgets than does subject ID 217, while ID 130 exhibits larger relative price changes than ID 217, as measured by the angular distance. On the other hand, ID 112 (Type 5) exhibits almost no relative or absolute change.

Figure 8 illustrates both larger absolute changes in price elasticities and larger relative changes in price elasticities for Type 4s in comparison with Type 5s. Type 5s exhibit little absolute change in price elasticities, with some 95% of subjects displaying less than 1 unit of absolute price change, in comparison with less than 40% of Type 4s. In addition, some 95% of Type 5s display no relative price changes, in comparison to less than 75% of Type 4s. For both the absolute change and relative change in price elasticities, the two-sided KS test rejects the null hypothesis that the distribution of these changes for Type 4s and Type 5s are drawn from the same distribution (with a p-value less than 0.001).[28]

[27] This metric, which takes values on $[-1, 1]$, measures the length of the angle between the two vectors originating from the origin in N-th dimensional space and pointing toward the two price elasticity vectors $\vec{e_L^i}, \vec{e_K^i}$.

[28] An F test of the null hypothesis that the elasticity change magnitudes for Types 4 and 5 are distributed with the same exponential distribution (i.e., that their distribution means are the same) indicates a

Figure 8: Magnitude of Normalized Elasticity Change per Individual Histogram by Type



Notes: The figure shows histograms for two alternative measures of the average degree of change in individual price elasticity across the kinked and linear treatments, the Euclidean distance measure (left panel) and the angular distance measure (right panel). KS test p-values are reported for the null hypothesis that the Type 4 and Type 5 distributions in each panel are the same. LR test p-values are reported for the null hypothesis that the Type 4 and Type 5 sample means in each panel are the same.

Figure 8 shows that the results in Figure 6 are generalizable to the elasticity values when evaluated across the full range of prices common to our experimental subjects. Overall, our results indicate that moving to the more complex kinked price environment uniquely affects Type 4s by significantly and sizably changing their demonstrated price elasticity, risk aversion, and, utility parameters.

rejection with a p-value smaller than 0.001 for the absolute change measure and the relative change measure. For the relative change measure, the random variable is scaled by -1 and then added to 1 to generate an exponential distribution.

4 Alternative Explanations of the Findings

Before proceeding to a discussion of how the findings impact economic analysis, we consider potential explanations for our results. If we compare the average elasticities in the linear treatment between the arbitrarily consistent and the coherently consistent groups, it is clear that the former has significantly larger elasticities than the latter.[29] Taking this fact into account, we may offer a conjecture on the source of these results. Coherently consistent subjects may be able to be consistent across treatments because they are, to begin with, implementing relatively simple decision rules that are not very responsive to relative price changes and therefore can be applied more consistently across more complex settings. On the other hand, decision rules where the demand ratios are more sensitive to price changes, such as those already employed in a simple pricing context by the arbitrarily consistent group, may be harder to implement in the face of more complex, non-linear pricing schemes.

Looking more deeply into the patterns of choice behavior for Type 4s and Type 5s, we find evidence consistent with this complexity hypothesis. For instance, in the linear treatment, coherently consistent subjects (Type 5s) exhibit many more decisions where X and Y are chosen in a fixed 1-to-1 proportion than do arbitrarily consistent subjects (Type 4s)—18.2% of choices vs. 3.1%.[30] As choosing a fixed demand ratio requires no attention to relative prices, such a decision rule is quite easy to extend from a context with one set of governing prices to another with two governing price ratios since the price ratios would be irrelevant to the fixed proportion choice in either case (and indeed, Type 5s continue to make choices at the 45 degree line at much higher rates than Type 4s when they move to the kinked environment).

Similarly, a person guided by a decision rule that seeks either to maximize expected

[29] See Table 28 in Appendix G of the Supplementary Online Material.

[30] This is true if we look at the fraction of choices exactly on the 45 degree line or the fraction of choices that are very close to the 45 degree line (to account for decision making that may have been intended to be at the 45 degree line while actual token selection was not perfectly placed in the computer interface). Taking a small circle of ϵ radius around the points on the 45 degree line puts the fraction of near-45 degree line choices at 49% for Type 5s and 29% for Type 4s in the linear treatment for $\epsilon = 3$, with a similar pattern for ϵ in the range of (0,3).

value or to grab the bundle containing the largest state-specific commodity amount (whether X or Y)—yielding mostly corner solutions when facing a linear budget (and, specifically, the larger of the X and Y intercepts for a linear budget)—is also a person whose demand ratio in practice is very insensitive to relative price changes for all but two changes in prices.[31] Such people have another simple and (mostly) price-invariant decision rule—easy to implement both when prices are linear and when they are more complicated and non-linear (since for most price changes, the change is irrelevant to the decision to stay at the corner). In the data, we also see that coherently consistent subjects in the linear treatment choose the most lucrative intercept significantly more often than arbitrarily consistent subjects do (four times as often—37.3% vs. 9.7% of the time).[32]

These patterns point to possible price-insensitive decision rules that are more prevalent among the coherently consistent group than among the arbitrarily consistent group.

While subjects display considerable heterogeneity in their decision rules, even within types, the goal of this discussion is not to fully characterize these decision rules. Instead, we aim to point to examples in the data of choices that would come from some sort of price-insensitive decision rule or heuristic and which would, for coherently consistent subjects, be relatively straightforward to extend in more complicated pricing schemes. Arbitrarily consistent subjects, whose choices we know to be more sensitive to relative price changes in the simplified linear pricing scheme, demonstrate far fewer of these examples and may find it harder in a complicated non-linear price context to maintain the same price-responsive decision rules that they had previously implemented. This interpretation is further supported by the observation that arbitrarily consistent (but not coherently consistent) subjects spend significantly more time per decision in the kinked treatment than in the linear treatment, suggesting the added complexity of the kinked budget sets

[31] Namely, when the price ratio goes from greater than 1 to 1 and when it goes from 1 to less than 1. For any change that leaves the pre and post price ratios above 1 or, alternatively, leaves them both below 1, the demand ratio is invariant to the price change, being effectively 0 or infinity.

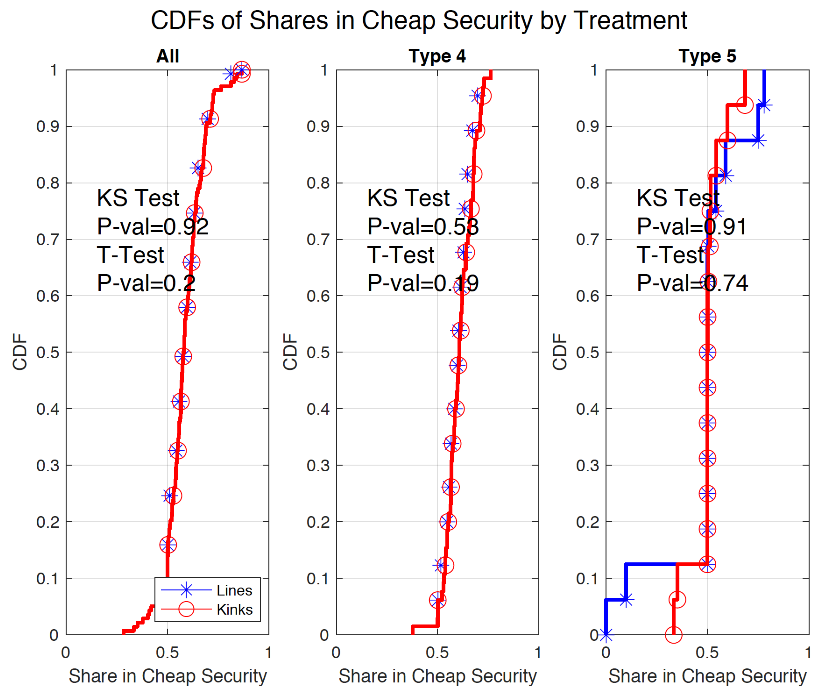
[32] In fact, put more starkly, in the linear treatment, approximately 30% of all Type 5 subjects make choices mostly at the most lucrative intercept (with 20% choosing this 90% of the time or more), while just 1.5% of Type 4 subjects do. Relatedly, regarding our other referenced simple decision rule, in the linear treatment, approximately 15% of Type 5 subjects make choices mostly on the 45 degree line, while 0% of Type 4 subjects do.

indeed requires more processing time for these types in particular, as would be expected if they are struggling (and then failing) to implement a more sophisticated decision rule in this more complex setting.

While this is our preferred interpretation, it is worth considering alternative explanations of the findings inspired by the literature. One natural explanation of the findings is that Type 4 behavior differs across treatments because subjects choose the kink point due to its salience. Two facts dispel this potential mechanism. First, choice plots such as those in Figure 2 show that subjects do not as a rule switch to choosing the kink point, with many choices made off the kink.[33] Second, the design of the experiment allows us to compare kinked sets that contain the choice the individual made in the linear set. We use these observations to more formally test whether behavior moves toward the kink point from the point of the linear choice. To compare decisions across budgets with different prices, we consider the fraction of the allocation that is allotted to the security with the lower price (i.e., the relatively cheap security). The share of the cheap security at the kink point is always larger than this share at the point chosen by the subject in the linear set since the kinked set was generated by removing a section of the axis with the cheaper security (see Section 1) and we only consider kinked sets that contain the choice the individual made in the linear set. If it were the case that the kink point had some sort of pull on subjects' tastes due to its salience, we would then see this share increase in the kinked case relative to the linear case. Figure 9 presents the distributions of the shares allotted to the cheap security for decisions in which the choice made by the subject in the linear treatment is available in the kinked treatment derived from that linear budget. First, we note that the figure shows that only a very small share of the decisions are under one-half, meaning very few of these choices were stochastically dominated. Moreover, a KS test of the distributions in the kinked and linear case shows that the two distributions are not statistically different for subjects overall or, critically (for the purposes of ruling out this particular alternative explanation), for Type 4 subjects in particular. Individuals do

[33]Choices at the kink may be rational and not just due to salience: for instance, an individual maximizing expected value would be expected to deviate from the most lucrative intercept when facing a kinked budget set where the angle of the kink includes the price ratio of -1.

Figure 9: Allocation Share of Cheaper Security



Notes: The figure presents the distribution of the shares allotted to the cheaper security (as defined by the initial linear treatment's relative prices) out of the total shares allotted to both securities for all decisions in which the choice made by the subject in the linear treatment is available in the kinked treatment (the distribution for these linear treatments is in blue and that for their associated kinked treatments in red). KS test/t test p-values are presented within each panel for the test of the null of no differences between the blue and red distributions/means, respectively, in each panel (with the left panel containing all subjects, the middle panel Type 4 subjects, and the right panel Type 5 subjects).

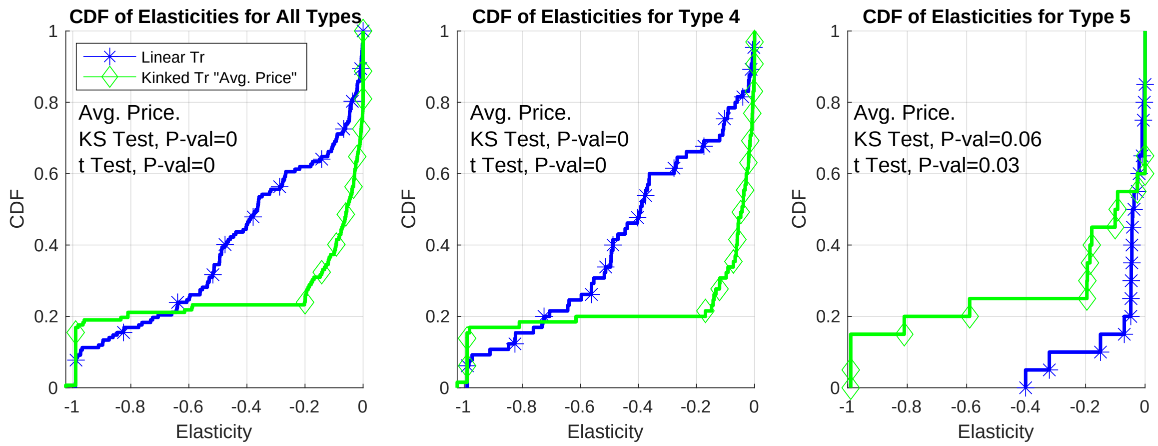
not experience enough gravitational pull toward the kink point to cause them to deviate from a previously preferred choice for some "salience" reward, which constitutes evidence contrary to the hypothesis that choices are more likely to be made at the kink due to its salience.

Readers familiar with the literature on heuristics in the face of marginal tax schedules may wonder if the behavior of arbitrarily consistent subjects can be explained by the "ironing" heuristic. Under this heuristic, posited by Liebman and Zeckhauser (2004), an individual in effect averages a set of marginal price schedules (each applying at different points in the commodity space range) and then acts upon this average price regardless of the marginal price actually at play. Rees-Jones and Taubinsky (2016) show that a large fraction of taxpayers linearize the marginal income tax schedule in this way (see also Gideon (2014) and Ito (2014) for additional empirical evidence of this heuristic). In our case, one possibility is that rather than changing their preferences between the linear and kinked treatments, arbitrarily consistent subjects do indeed have a stable preference relation but employ it over a misperceived "ironed" price schedule in the kinked treatment. This could make choices appear as if they are being determined by different, inconsistent decision rules applied to the actual marginal prices (when in fact the decision maker is applying the same rule to the actual price schedule in the linear treatment and to a misperceived ironed price schedule in the kinked treatment). We can test for supporting evidence for this hypothesis by performing our estimation procedure using the actual kinked choices observed but treating them in the maximization procedure as if they are associated with an ironed, linearized price schedule that averages the actual marginal price schedules defining any given kinked budget set.[34] If we believe that the choice made lies on a budget constraint that takes the ironed price ratio (again, derived by averaging the two actual marginal price ratios forming the kinked budget set) as the perceived price ratio and that preferences are the same as in the linear setting, then we would expect to see similar elasticities with respect to the ironed linearized price for arbitrarily consistent subjects as we do in the linear setting. However, as Figure 10 makes clear, the distribution of elasticities for these Type 4 individuals under this ironing assumption does not at all resemble the elasticity distribution derived from the actual linear setting. We take this evidence as inconsistent with the ironing hypothesis explaining our findings (though it may explain behavior in other kinked budget settings).

A final alternative explanation for the findings that we consider is that Type 4s may be distinguished from Type 5s by a greater propensity to make choices violating first-order stochastic dominance (FOSD). However, the overall pattern of HM violations within and across treatments and the ratio of Type 4 to Type 5 individuals in the population is

[34]Specifically, we compute an average price ratio that is the simple average of P_x^1/P_y^1 for segment 1 of the kinked budget constraint and P_x^1/P_y^1 for segment 2, then run a budget with this averaged slope through the actual choice made and estimate preferences and elasticities with the procedure described in Appendix C on the basis of all paired choice and averaged price combinations in the kinked treatment.

Figure 10: Elasticities with Ironing Assumption



Notes: The figure presents, for the arbitrarily consistent group (Type 4s), the CDF of the per-person mean elasticity derived from structural estimation of the linear treatment (in blue), described in Section 5.1, and a similar structural estimation where the maximum likelihood estimation uses the observed choices in the kinked treatment but treats them as if they are associated with a linearized price schedule that averages the actual marginal price schedules that define any given kinked budget set (in yellow), as a consumer employing an ironing heuristic might do. KS test/t test p-values are presented within each panel for the test of the null of no differences between the linear and ironed distributions/means, respectively, in each panel (with the left panel containing all subjects, the middle panel Type 4 subjects, and the right panel Type 5 subjects).

not meaningfully altered when we drop the FOSD violations for each individual and then determine the HM violations and resulting type assignment on the remaining set. Specifically, the HM violations within each treatment still indicate high levels of internal consistency with markedly different (and less consistent) choices made for the pooled observations across treatments. Similarly, the resulting type distribution is still populated with mostly Type 4s and Type 5s, with the figure showing strictly more of the former for a critical value of 6 or less, as in our baseline analysis. Once we remove the FOSD violations, 72% of the population is categorized as a Type 4 and 23% as a Type 5—with almost no subjects classed as Type 1, Type 2, or Type 3.[35] These results show that FOSD violations do explain part of the type distribution, namely, the part corresponding to subjects who fail to show consistency in at least one of the treatments. Excluding such violations does not change the ratio of the remaining types, with Type 4 subjects

[35] See Figures 17–19 and Table 16 in Appendix F of the Supplementary Online Material.

outnumbering Type 5 subjects by a rate of about 3 to 1, as in our baseline analysis. We thus conclude that this alternative also has little to offer in terms of an explanation for our main findings.

5 Consequences for Economic Analysis

The results of this experiment have several important consequences for economic analysis. First, we show that for the overwhelming majority of individuals, behavior in the kinked treatment is rational and consistent with the maximization of some utility function—an untested assumption in previous studies. This finding tells us that even if individuals in non-linear budget constraint settings act out of step with past behavior in linear settings, we should be hesitant to conclude that they are failing to optimize; rather, they may still be acting in a way that is consistent with some underlying preference relation. Economists should thus show caution in ascribing irrationality to individual behavior in such settings.

Second, the fact that half of our experimental subjects change their maximizing behavior even with full information means that a lack of information is not the only explanation of why behavior in nonlinear settings differs from that in linear settings. Information provision interventions that seek to help individuals understand complex incentive schemes (e.g., Jones, 2010; Chetty and Saez, 2013) are common and powerful. However, our results suggest that they may also have natural limitations and that we should not expect information alone to harmonize behavior across these disparate incentive settings. In fact, our findings indicate a context-dependent switch in decision-making rules and resulting behavior (especially for price-sensitive subjects), suggesting a renewed importance for behavioral models that incorporate features of the environment and, critically, the complexity of the choice setting in determining choice.[36]

In this vein, special discussion is in order regarding related work on the effects of incentive complexity. Abeler and Jäger (2015) show that experimental subjects facing more

[36] Simon (1956) is a notable early proponent of the view that to understand human behavior, one must understand both the environment and the decision maker. Recent related works include Caplin et al. (2011), which employs a satisficing model to explain search behavior, and Kőszegi and Szeidl (2012), which uses the degree of difference between attributes in a choice set to predict behavior, and Reck (2016) which studies the welfare effects of debiasing.

complex lab-designed tax and subsidy schemes under-react to changes in taxes vis-a-vis those facing simpler schemes containing equivalent tax changes. Like these authors, we demonstrate that behavior under more complicated incentive structures fundamentally differs from behavior under less complicated incentive schemes. What is striking is that while in Abeler and Jäger (2015), the complex incentive scheme is very complex (presented in a full page of 22 tax and subsidy rules that change depending on the range of experimental output), in our case, the non-linear incentive treatment is about as simple a deviation from a uniform price as possible (a single kink) and is presented transparently with a graphical interface and its built-in calculators. Still, this very minor addition of complexity to the incentive structure is enough to significantly change the underlying preferences and behavioral responsiveness to prices of half the population. Moreover, while Abeler and Jäger (2015) find that the under-reaction in the complex incentive scheme is sub-optimal relative to what would be payoff-maximizing (and their experiment is not designed with the intent of comparing a payoff-maximizing objective with subjects' actual utility function maximization), we can show that even though added complexity in the kinked incentive scheme does change behavioral price responsiveness, this change can still be characterized as optimal behavior for some utility function. Unlike the previous authors, we also find that the change that the complex incentive scheme brings out in an individual, while large in magnitude, can either be a large decrease *or a large increase* in behavioral responsiveness (in our case measured in terms of price elasticity), depending on the individual.

A third implication of our work relates to this consequence of incentive complexity. Policy makers should be cognizant of the fact that "kink-ifying" incentive schedules—in spite of the rationales for doing so—can fundamentally affect individuals' decision rules, measures of price responsiveness, and therefore the size of the behavioral distortions that an individual exhibits. In our sample, arbitrarily consistent subjects who increase their elasticity in response to a kinked budget set are offset by arbitrarily consistent subjects who decrease their elasticity in the very same circumstance. Even though we do not find a significant change in the average elasticity, the individuals who change their decision rules would experience welfare losses under the preferences revealed in the linear setting. While in our setting there is no average change in elasticity, population average elasticities in other settings may increase or decrease, depending on the context. The choice to "kink-ify" an incentive schedule may then impact not only individual behavioral responsiveness but also aggregate population elasticities by making the population increase or decrease its average price responsiveness, which may impact the efficiency cost of alternative linear and non-linear tax schedules.

Finally, we note that the results in this paper also have consequences for a recent literature that uses marginal responses at points where marginal incentives are discontinuous to estimate features of demand and supply functions. Saez (2010), in his influential work in this area, notes that the fraction of individuals who locate, or bunch, at kink points in the income tax schedule is proportional to the elasticity of reported income with respect to the net of tax rate and proposes an estimator of this elasticity based on this fraction. A natural question to ask, then, is the following: what does arbitrary consistency mean for the bunching estimator-derived elasticities? The implications of our work in this domain are twofold. First, we note that consistent behavior is necessary for this method to be valid, as it depends on maximizing behavior for valid estimates. Our findings thus validate this assumption. Second, arbitrary consistency casts doubt on the comparability of kink-design estimates (bunching estimator derived or otherwise) and estimates derived from choices made under linear budgets, as the two might be estimating features of different demand/supply functions. This point is of particular importance to policy makers seeking to make use of empirically validated elasticities in their policy formation (e.g., in tax simulation models employing measured taxable income elasticities or macroeconomic models utilizing estimated labor supply elasticities). In the context of labor responses, for instance, taking elasticities estimated from environments with linear budget constraints and applying them to non-linear environments (or vice versa) may lead to significant forecasting errors.[37]

[37]For example, estimating labor supply responses using changes in Social Security taxes (which for most workers are in effect a linear tax) or the large experimental lab labor literature (which often models taxes as flat taxes and derives labor supply elasticities with changes to the slope) may lead to inappropriate conclusions about worker responsiveness to a progressive income tax schedule.

6 Conclusion

This paper describes the first experiment to assess the rationality of individuals facing non-linear budgets. In the experiment, we elicit a large number of individual decisions using the graphical budget set toolkit of Choi et al. (2007b) extended to include non-linear budget sets. The analysis reveals several novel results. First, individual measures of rationality under kinked sets are found to display very similar aggregate patterns to those under linear budget sets, with a large majority of participants exhibiting consistent behavior over choices in each setting. Second, half of subjects exhibit behavior that, while rational in both treatments, is not consistent with a common decision rule across treatments, implying that they behave in an arbitrarily consistent manner and maximize different utility functions across treatments. Third, individuals who change their decision rules with the move to the kinked budget set environment are characterized by large shifts in estimated utility parameters, risk premiums, and price elasticities (in comparison to individuals who act consistently across budget environments, whose associated values are, as expected, essentially unchanged). Fourth, we reject prima facie explanations for this finding such as consumer irrationality and the ironing hypothesis, among others. We instead present evidence suggesting that the findings can be explained by rational actors responding to increased price complexity by shifting their decision rules. Interestingly, we find that those who implement less sophisticated, less price-responsive decision rules to begin with are less likely to change decision rules in more complex environments.

Finally, the results have important policy implications. For one, they suggest that caution is warranted in applying individual-level behavioral elasticity estimates derived from linear settings to non-linear settings (and vice versa). Such a finding is of particular note to those attempting to generalize from the large and varied literature on empirical and experimental labor supply estimates, demand estimations, and taxable income elasticity, among others. Additionally, the findings suggest that policy makers should appreciate how the choice to "kink-ify" an incentive schedule can fundamentally change revealed preferences, individual price responsiveness, and therefore the size of resulting behavioral distortions.

While the experimental setting allows control and inference not possible outside the lab, it does so within a particular decision context. It is unknown how the experimental findings might extend to other non-linear settings, such as choices over effort and leisure or between commodity bundles chosen with certainty. Extending the analysis to these additional decision margins while continuing to decouple confounding issues related to information and adjustment costs from incentive complexity in these other contexts is thus an important avenue for future work.

References

Abeler, Johannes and Simon Jäger, "Complex tax incentives," *American Economic Journal: Economic Policy*, 2015, *7* (3), 1–28.

Afriat, S., "The construction of a Utility Function from Demand Data," *International Economic Review*, 1967, *8*, 67–77.

_ , "Efficiency Estimates of Production Functions," *International Economic Review*, 1972, *8*, 568–598.

Ahn, David, Syngjoo Choi, Douglas Gale, and Shachar Kariv, "Estimating ambiguity aversion in a portfolio choice experiment," *Quantitative Economics*, 2014, *5* (2), 195–223.

Ariely, Dan, George Loewenstein, and Drazen Prelec, "Coherent Arbitrariness: Stable Demand Curves without Stable Preferences," *The Quarterly Journal of Economics*, 2003, *118*, 73–105.

Barseghyan, Levon, Francesca Molinari, Ted O'Donoghue, and Joshua C Teitelbaum, "Estimating risk preferences in the field," *Journal of Economic Literature*, 2018, *56* (2), 501–64.

Bernheim, B. Douglas and Antonio Rangel, "Beyond Revealed Preference: Choice-Theoretic Foundations for Behavioral Welfare Economics," *The Quarterly Journal of Economics*, 2009, *124* (1), 51–104.

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

Caplin, Andrew, Mark Dean, and Daniel Martin, "Search and satisficing," *American Economic Review*, 2011, *101* (7), 2899–2922.

Chetty, Raj, Adam Looney, and Kory Kroft, "Salience and Taxation: Theory and Evidence," *American Economic Review*, 2009.

— **and Adam Szeidl**, "Consumption Commitments and Risk Preferences," *Quarterly Journal of Economics*, 2007, *122* (2), 831–877.

— **and Emmanuel Saez**, "Teaching the tax code: Earnings responses to an experiment with EITC recipients," *American Economic Journal: Applied Economics*, 2013, *5* (1), 1–31.

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

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

Choi, Syngjoo, Shachar Kariv, Wieland Müller, and Dan Silverman, "Who is (More) Rational?," *American Economic Review*, 2014, *104* (6), 1518–50.

Dean, Mark and Daniel Martin, "Measuring rationality with the minimum cost of revealed preference violations," *Review of Economics and Statistics*, 2016, *98* (3), 524–534.

Feldman, Naomi E., Peter Katuk, and Laura Kawano, "Taxpayer Confusion: Evidence from the Child Tax Credit," *American Economic Review*, March 2016, *106* (3), 807–35.

Fisman, Ray, Shachar Kariv, and Daniel Markovits, "Individual Preferences for Giving," *American Economic Review*, 2007, *97* (2), 153–158.

Forges, Francoise and Enrico Minelli, "Afriat's theorem for general budget sets," *Journal of Economic Theory*, January 2009, *144* (1), 135–145.

Gideon, Michael S, "Uncovering Heterogeneity in Income Tax Perceptions," Working Paper 2014.

Goldin, Jacob and Tatiana Homonoff, "Smoke gets in your eyes: cigarette tax salience and regressivity," *American Economic Journal: Economic Policy*, 2013, *5* (1), 302–36.

Gruber, Jon and Emmanuel Saez, "The elasticity of taxable income: evidence and implications," *Journal of Public Economics*, 2002, *84*, 1–32.

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

Hodge, Victoria and Jim Austin, "A survey of outlier detection methodologies," *Artificial intelligence review*, 2004, *22* (2), 85–126.

Holland, Paul W and Roy E Welsch, "Robust regression using iteratively reweighted least-squares," *Communications in Statistics-theory and Methods*, 1977, *6* (9), 813–827.

Houtman, M. and J. Maks, "Determining all Maximal Data Subsets Consistent with Revealed Preference," *Kwantitatieve Methoden*, 1985, *19*, 89–104.

Hoynes, Hillary, "Welfare Transfers in Two-Parent Families: Labor Supply and Welfare Participation Under the AFDC-UP Program," *Econometrica*, March 1996, *64* (2), 295–332.

Huber, Peter J, *Robust statistics*, Vol. 523, John Wiley & Sons, 2004.

Ito, Koichiro, "Do consumers respond to marginal or average price? Evidence from nonlinear electricity pricing," *American Economic Review*, 2014, *104* (2), 537–63.

Jones, Damon, "Information, Preferences, and Public Benefit Participation: Experimental Evidence from the Advance EITC and 401(k) Savings," *American Economic Journal: Applied Economics*, April 2010, *2*, 147–163.

Kőszegi, Botond and Adam Szeidl, "A model of focusing in economic choice," *The Quarterly journal of economics*, 2012, *128* (1), 53–104.

Liebman, Jeffrey and Richard Zeckhauser, "Schmeduling," Working Paper, Harvard University, Kennedy School of Government 2004.

Matzkin, Rosa L, "Axioms of revealed preference for nonlinear choice sets," *Econometrica: Journal of the Econometric Society*, 1991, pp. 1779–1786.

Moffitt, Robert, "The econometrics of kinked budget constraints," *Journal of Economic Perspectives*, 1990, *4* (2), 119–139.

Reck, Daniel, "Taxes and Mistakes: What's in a Sufficient Statistic?," Working Paper, Available at SSRN: https://ssrn.com/abstract=2268617 2016.

Rees-Jones, Alex and Dmitry Taubinsky, "Heuristic Perceptions of the Income Tax: Evidence and Implications for Debiasing," NBER Working Papers 22884, National Bureau of Economic Research, Inc December 2016.

Rousseeuw, Peter J and Mia Hubert, "Robust statistics for outlier detection," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2011, *1* (1), 73–79.

Saez, Emmanuel, "Do Taxpayers Bunch at Kink Points?," *American Economic Journal: Economic Policy*, 2010, *2* (3), 180–212.

Schildberg-Hörisch, Hannah, "Are Risk Preferences Stable?," *Journal of Economic Perspectives*, 2018, *32* (2), 135–54.

Simon, H. A., "Rational Choice and the Structure of the Environment," *Psychological Review*, 1956, *63*, 129–138.

Online Appendix: Not For Publication

This online appendix includes the experiment instructions in Appendix A, details on the procedure to benchmark our test of rationality in Appendix B, and details on the structural estimation in Appendix C. Our online supplementary appendix includes additional material and be found here www.jcsuarez.com/Files/Kinks_Supplementary_Appendix.pdf.

A Experiment Instructions

Sample instructions

Introduction

This is an experiment in decision-making. Research foundations have provided funds for conducting this research. Your payoffs will depend only on your decisions and on chance. It will not depend on the decisions of the other participants in the experiments. Please pay careful attention to the instructions as a considerable amount of money is at stake. After you read this part of the instructions, it will also be read aloud by the instructor, and you may also ask any questions.

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

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

3 Tokens = 1 Dollar

Your participation in the experiment and any information about your payoffs will be kept strictly confidential. Each participant will be assigned a participant ID number. This number will be used to record all data. Only the Xlab administrator but not the experimenter will have both the list of participant ID numbers and names.

Please do not talk with anyone during the experiment. In order to keep your decisions private, please do not show your choices to any other participant. We also ask everyone to remain silent until the end of the experiment. At the end of the experiment you will be paid privately according to your participant ID number.

This experiment consists of two parts. At the end of Part I you will be given the instructions for Part II.

Part I

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

In each decision problem you will be asked to allocate tokens between two accounts, labeled *x* and *y*. The *x* account corresponds to the *x*-axis and the *y* account corresponds to the *y*-axis in a two-dimensional graph. Each choice will involve choosing a point on a line representing possible token allocations. Examples of lines that you might face appear in Attachment 1.

In each choice, you may choose any *x* and *y* pair that is on the line. For example, as illustrated in Attachment 2, choice *A* represents a decision to allocate *q* tokens in the *x* account and *r* tokens in the *y* account. Another possible allocation is *B*, in which you allocate *w* tokens in the *x* account and *z* tokens in the *y* account.

Each decision problem will start by having the computer select such a line randomly from the set of lines that intersect with at least one of the axes at 50 or more tokens but with no axis exceeding 100 tokens. The lines selected for you in different decision problems are independent of each other and of the lines selected for any of the other participants in their decision problems.

To choose an allocation, use the mouse to move the pointer on the computer screen to the allocation that you desire. When you are ready to make your decision, left-click to enter your chosen allocation. After that, confirm your decision by clicking on the Submit button. Note that you can choose only *x* and *y* combinations that are on the line. To move on to the next round, press the OK button. The computer program dialog window is shown in Attachment 3.

Your payoff at each decision round is determined by the number of tokens in your *x* account and the number of tokens in your *y* account. At the end of the round, the computer will randomly select one of the accounts, *x* or *y*. It is equally likely that account *x* or account *y* will be chosen. You will only receive the number of tokens you allocated to the account that was chosen.

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

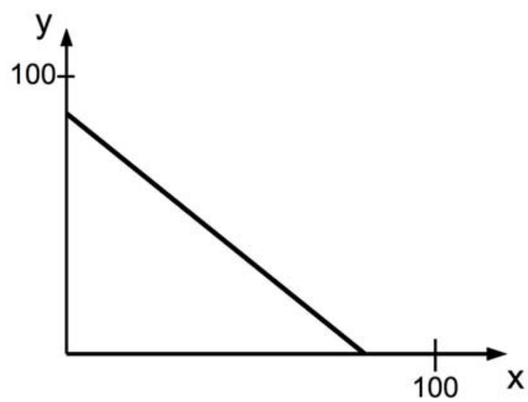
Your earnings for this part of the experiment will be determined as follows. At the end of the experiment, the computer will randomly select one decision round to carry out (that is, 1 out of 50) for payoffs. The round selected depends solely upon chance. For each participant, it is equally likely that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

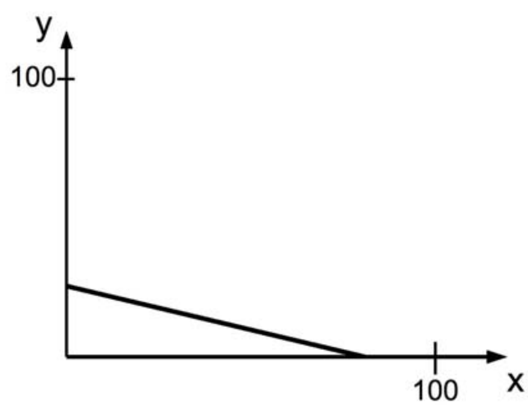
For example, suppose that in the round the computer chose to carry out for payoffs, you chose allocation *A*, as illustrated in Attachment 2, and that the computer chose account *y* for you in that round. In that case you would receive *r* tokens in total. Similarly, if the computer chose account *x* for you in that round then you would receive *q* tokens in total. If you chose allocation *B* and the computer chose account *y* you would receive *z* tokens in total, and if the computer chose account *x* then you would receive *w* tokens in total.

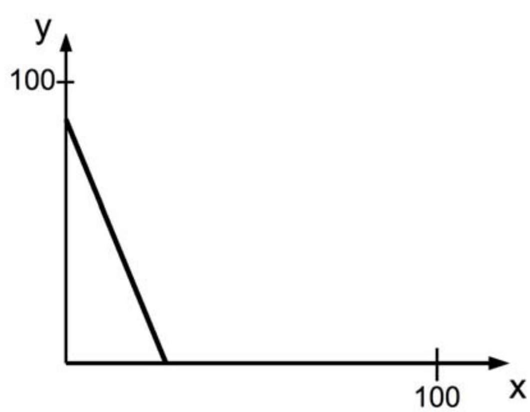
At the end of the experiment, the tokens will be converted into money. Each token will be worth 0.33 Dollars. You will receive your payment as you leave the experiment.

If there are no further questions, you are ready to start. At the end of this part of the experiment, you will receive further instructions.

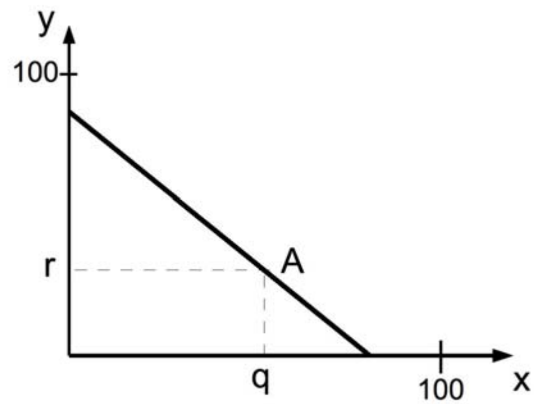
Attachment 1

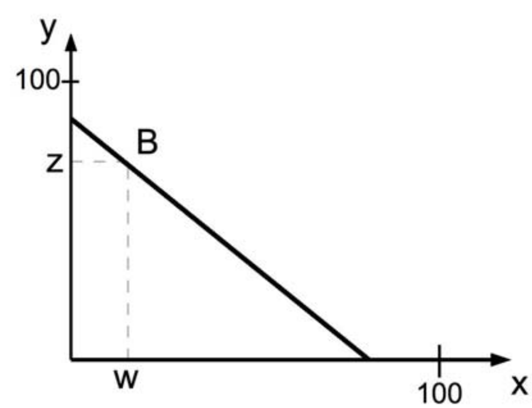




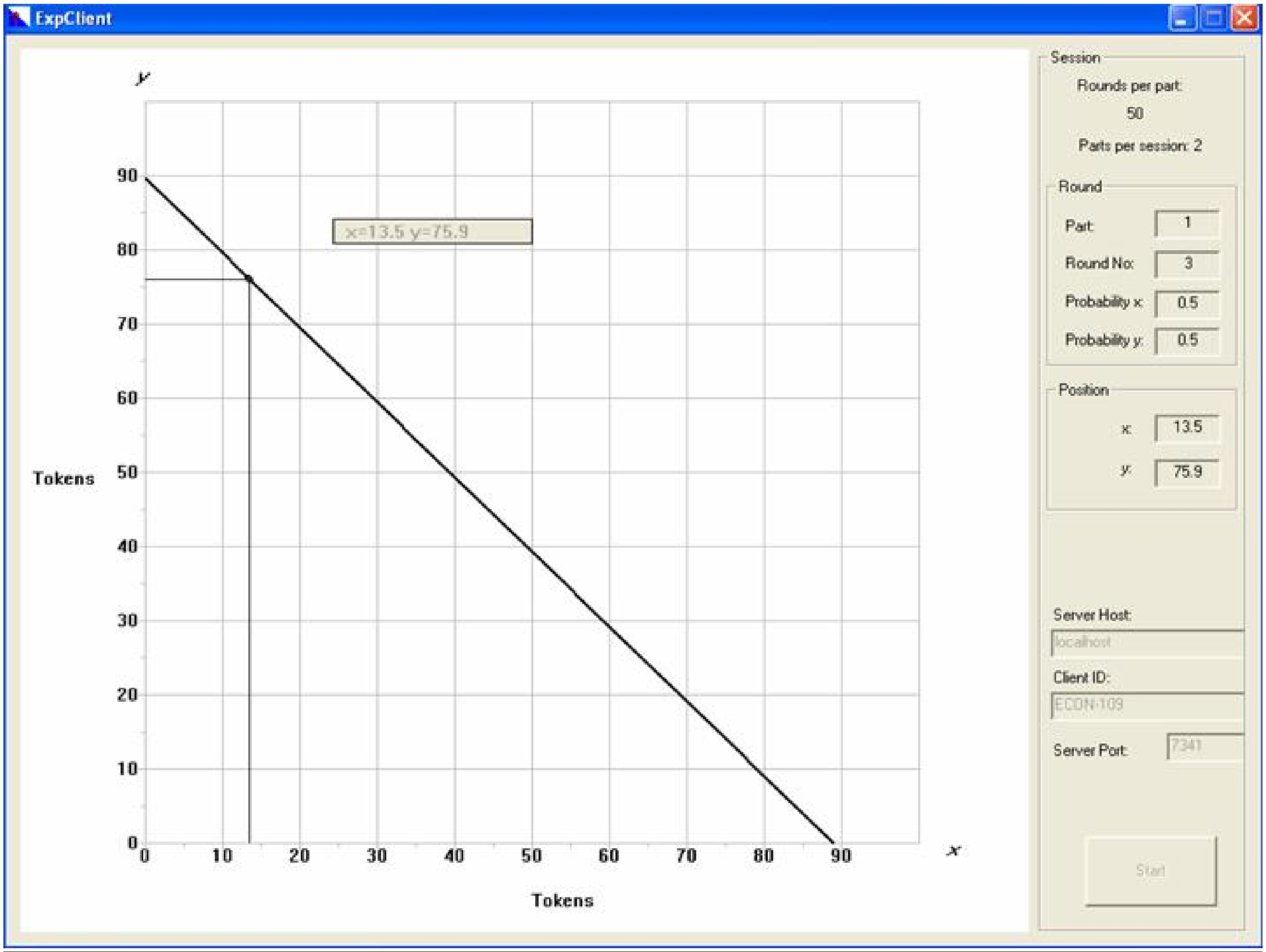


Attachment 2





Attachment 3



Part II

This part of the experiment employs the same experimental computer program. In this part of the experiment, you will also participate repeatedly in 50 independent decision problems that share a common form. This section describes in detail the differences between the two parts of the experiment. After you read this part of the instructions, it will also be read aloud by the instructor, and you may also ask any questions.

In each decision problem you will again be asked to allocate tokens between two accounts, labeled *x* and *y*. The *x* account corresponds to the *x*-axis and the *y* account corresponds to the *y*-axis in a two-dimensional graph. Once again, each choice will involve choosing a point representing possible token allocations.

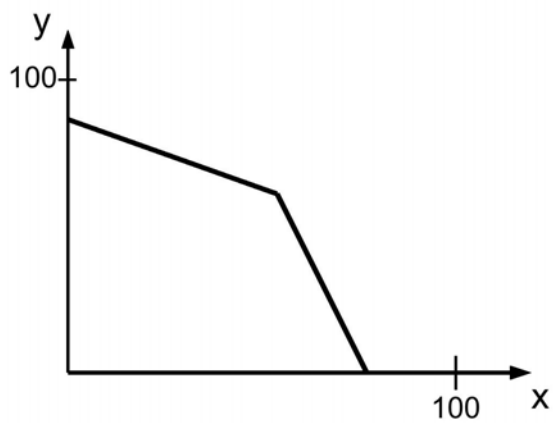
Again, each choice will involve choosing a point on a graph representing possible token allocations. The *x*-axis and *y*-axis are again scaled from 0 to 100 tokens. In each choice, you may choose any allocation that is on the kinked-shaped lines. Examples of lines that you might face appear in Attachment 4.

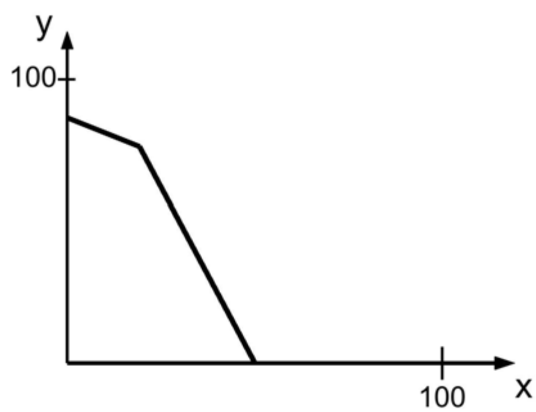
Each decision problem will start by having the computer select such a kinked-shaped line randomly. That is, the lines selected depend solely upon chance and it is equally likely that you will face any kinked-shaped line. The lines selected for you in different decision problems are independent of each other and of the lines selected for any of the other participants in their decision problems.

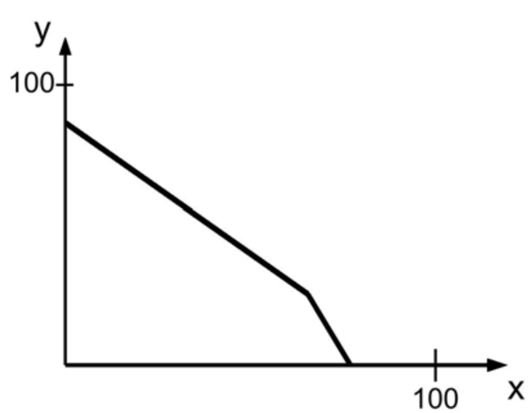
Recall that to choose an allocation, use the mouse to move the pointer on the computer screen to the allocation that you desire and click on your chosen allocation. Examples of possible choices appear in Attachment 5. For example, suppose that in the round the computer chose to carry out for payoffs, you chose allocation *A*, as illustrated in Attachment 5, and that the computer chose account *y* for you in that round. In that case you would receive *q* tokens in total. Similarly, if the computer chose account *x* for you in that round then you would receive *r* tokens in total. If you chose allocation *B* and the computer chose account *y* you would receive *z* tokens in total, and if the computer chose account *x* then you would receive *w* tokens in total.

In this part of the experiment, the method of determining payment is the same as in the previous part. Recall that in each round it is equally likely that account *x* or account *y* will be chosen. Once again, at the end of this part of the experiment, the computer will randomly select one of the fifty decision rounds from each participant to carry out for payoffs. You will receive your payment for this part of the experiment, together with your payment for the previous part, and the $5 participation fee, as you leave the experiment.

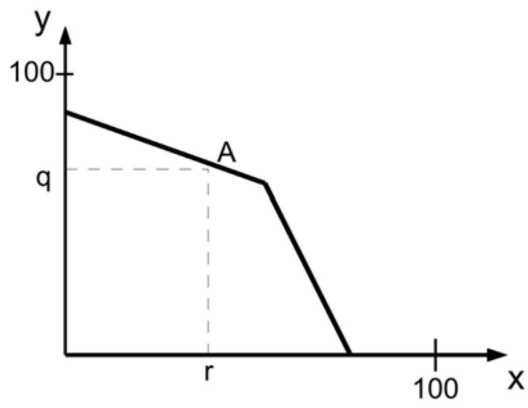
Attachment 4

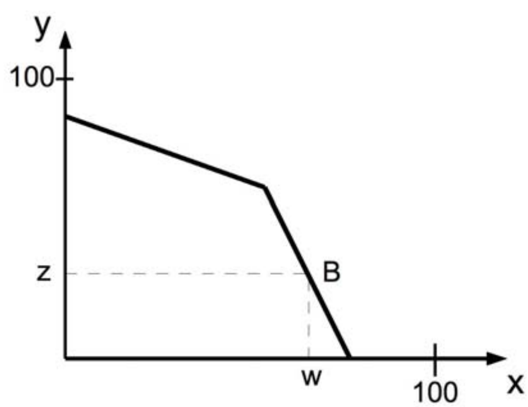






Attachment 5





B Benchmark Randomization and Type Taxonomy

This appendix presents the results of the benchmarking procedure used to determine the critical values for categorizing rationality types. We closely follow the methods of Choi et al. (2007a). We start with an individual whose behavior is determined by maximization of a CRRA (ρ set to 0.5, following Choi et al. (2007a)) expected utility function subject to logistic taste shocks, where the relative importance of the taste shock is determined by parameter γ (if $\gamma = 0$, the decisions are purely random, while when $\gamma \to \infty$, the decisions are purely the result of expected utility maximization by a rational agent). For each panel presented in the the first four figures in this appendix, we simulate 1300 subjects with these tastes for each combination of γ and n (the number of choice observations per subject) that we use.[38] The figures then plot the resulting distributions of the HM measure for each case, with Figure 11 showing the HM distributions derived from maximization subject to linear constraint sets and Figure 12 showing the HM distributions derived from kinked constraint sets.

In the first panel of Figures 11 and 12, we consider decision makers whose choices are completely random ($\gamma = 0$) and assess how the resulting distributions of the HM measure change as the number of observations increases. The results indicate that a larger number of decisions increases the statistical power to detect whether a set of data was generated by pure randomization, as is to be expected. Moreover, 50 decisions ($n = 50$) give us significant power against the weak hypothesis of pure randomization. Note that draws of 50 observations are just as powerful for the case of linear budget sets as for the case of kinked budget sets.

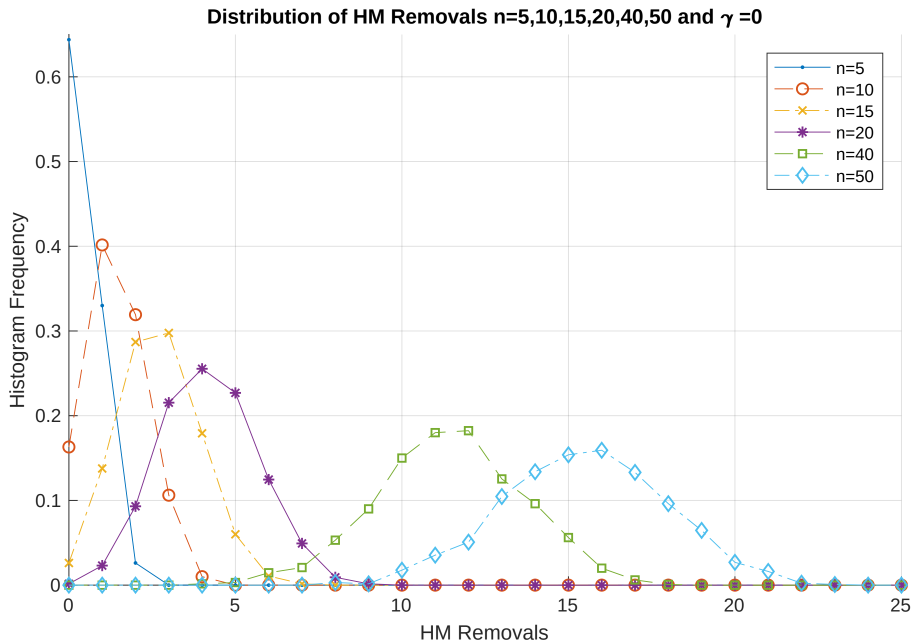
Stronger hypotheses can be formulated by comparing the experimental data to expected utility maximization subject to a random taste shock determined by non-zero γ values. For each combination of n and γ, one can test the null hypothesis that the observed behavior and resulting HM measure come from a distribution of individuals whose maximizing behavior is modulated with a degree of random shocks defined by level γ.

[38] Each of the choice observations from a simulated subject come from a scenario with a randomly generated budget set that is generated in the same way as those faced by the actual decision makers in our experiment.

The bottom panels of Figures 11-12 show the distributions (for $n = 50$ observations) in each budget context for differing degrees of γ. For instance, if one wanted to test the hypothesis of consistency equivalent to that from a distribution that has $\gamma = 5$ at the 95% level, one would compare the value of the HM score with the critical value 6. An HM score *greater* than 6 leads one to conclude that $\gamma < 5$ at the 95% confidence level or, equivalently, that the observed behavior is less rational, or consistent, than would be the case for 95% of a population that maximizes utility with only a moderately sized random taste shock. An individual with an HM greater than 6 would thus be said to be less consistent, with a 95% confidence level, than would be expected from individuals who only occasionally allow deviations from an otherwise rational CRRA utility maximization due to a moderate ($\gamma = 5$) random logistic taste shock. Similarly, for a somewhat higher standard of rationality ($\gamma = 10$), 4 HM removals becomes the critical value (i.e., an individual with an HM greater than 4 would thus be said to be less consistent, with a 95% confidence level, than would be expected from individuals who only occasionally allow deviations from an otherwise rational CRRA utility maximization due to a random logistic taste shock with parameter $\gamma = 10$).

Table 2 (in the main text) shows the type distribution using the critical value of 4. Individual type assignments using this critical value can be seen in Table 3. Figure 4 (in the main text) explores the robustness of the type distribution to the choice of critical value. Importantly, the proportion of Type 4s remains quite large as the critical value deviates from 4. The proportion of Type 5s increases as the critical value increases. This increase is due mostly to the decrease in Type 1s, Type 2s, and Type 3s. This is purely a mechanical effect of lowering the bar of rationality.

Figure 11: HM Randomization Results (Linear)



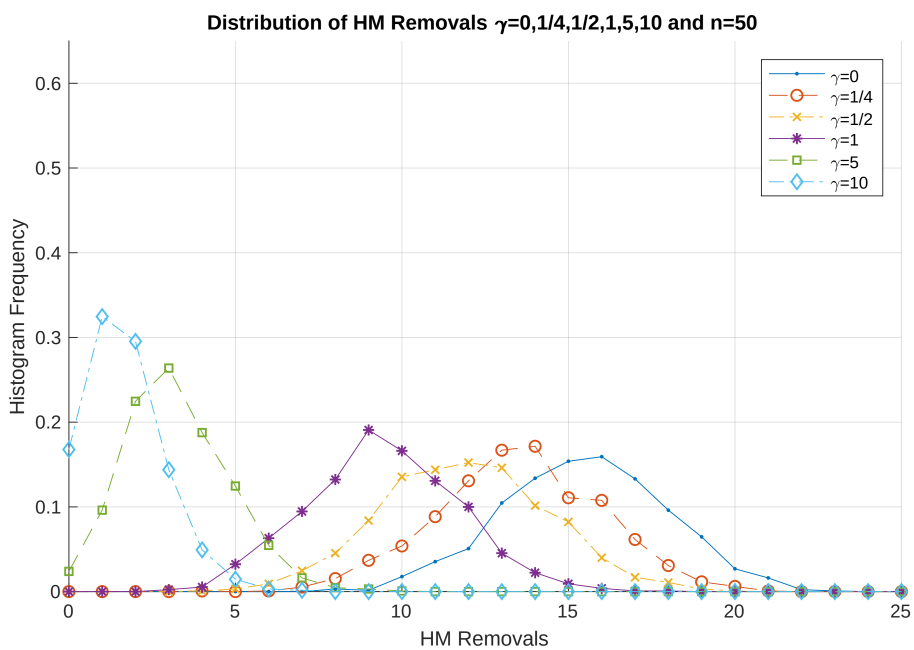


Figure 12: HM Randomization Results (Kinked)

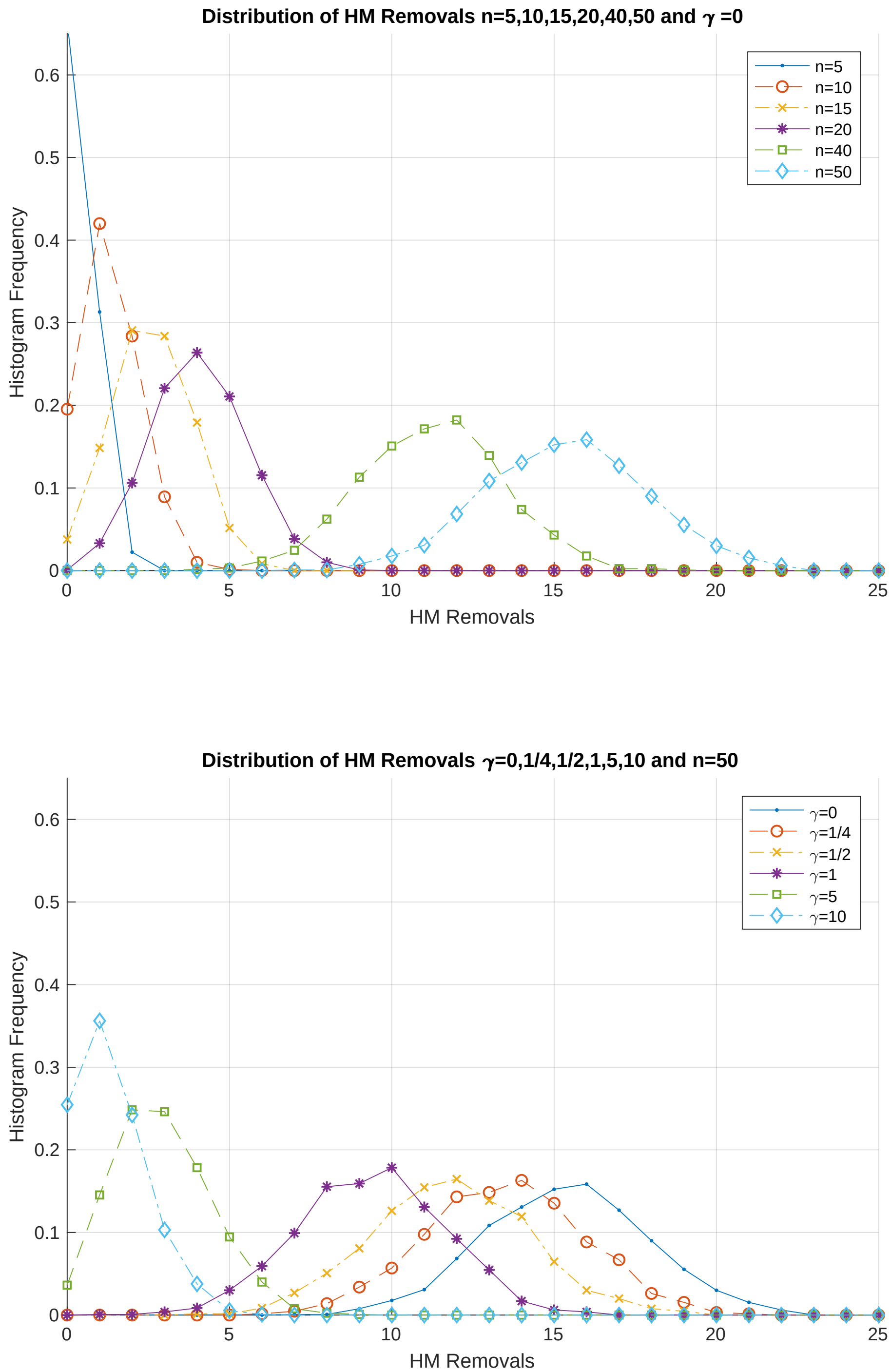


Table 3: Individual Type Listing

Session 1	Type	Session 2	Type	Session 3	Type	Session 4	Type
101	4	201	3	301	4	401	5
102	1	202	2	302	4	402	4
103	4	203	4	303	1	403	1
104	3	204	3	304	5	404	4
105	4	205	4	305	2	405	4
106	3	206	3	306	4	406	2
107	4	207	4	307	2	407	1
108	2	208	3	308	4	408	4
109	4	209	4	309	3	409	4
110	5	210	1	310	4	410	2
111	4	211	4	311	5	411	5
112	5	212	3	312	2	412	2
113	5	213	4	313	2	413	2
114	4	214	3	314	4	414	1
115	4	215	4	315	2	415	4
116	3	216	5	316	5	416	1
117	2	217	4	317	4	417	3
118	5	218	3	318	3	418	4
119	4	219	4	319	4	419	3
120	4	220	4	320	4	420	2
121	5	221	3	321	1	421	1
122	2	222	4	322	5	422	4
123	1	223	1	323	4	423	4
124	4	224	5	324	1	424	5
125	4	225	4	325	4	425	5
126	4	226	1	326	4	426	1
127	4	227	4	327	5	427	1
128	5	228	4	328	4	428	1
129	4	229	4	329	3	429	4
130	4	230	4	330	4	430	2
131	3	231	4	331	5	431	4
132	1	232	1	332	4	432	4
133	5	233	3	333	2	433	2
134	4	234	4	334	4	434	1
		235	2	335	2	435	5
		236	3	336	4	436	4

C Parametric Estimation

Moffitt (1990) discusses difficulties in estimating models with kinked budget sets. An advantage of our setting is that the experiment generates random variation in the size and location of the kink. Our parametric estimation in the linear and non-linear settings builds upon Choi et al. (2007a). Like them, we estimate the following disappointment aversion utility function from Gul (1991) with the inclusion of a multiplicative stochastic component:

$$\min\left\{\alpha u(x)e^{\varepsilon_1} + u(y)e^{\varepsilon_2}, u(x)e^{\varepsilon_1} + \alpha u(y)e^{\varepsilon_2}\right\} \tag{2}$$

where the function $u(x)$ has the CRRA form $u(x) = \frac{x^{1-\rho}}{1-\rho}$. We specify in this Appendix the NLLS problem—which forms the basis for the structural estimates used throughout the paper—for the estimation in the kinked treatment case (as the estimation of the linear treatment is included as a special case). Furthermore, for exposition purposes, we illustrate the case where the kink point is to the left of the 45 degree line. The opposite case requires a trivial symmetric modification.

To start, denote the price ratio p_x^i/p_y^i for the lines to the left and right of the kink point by p_1 and p_2, the kink point by $\log(y_k^i/x_k^i) = K^i$, and $\varepsilon = \varepsilon_1 - \varepsilon_2$. As boundary observations are not well defined with the power function, we incorporate them (following Choi et al. (2007a)) by replacing the zero component of a boundary choice with a very small consumption level such that the ratio of choices (x^i/y^i) is truncated to be between ω and $1/\omega$, where $\omega = 0.001$. A decision maker with preferences determined by equation (1) who faces a kinked budget set (defined by prices (p_1^i, p_2^i) and kink point K^i) will have their optimal choice (x^{i*}, y^{i*}) determined by the following log demand ratio schedule[39]:

[39] Accounting for the multiplicative stochastic component in (2) results in an additive error term ε in the log demand ratio expressions and underlies the criterion function minimized in (3) via NLLS.

$$\log(x^{i*}/y^{i*}) = f(\log(p_1^i), \log(p_2^i), K^i; \alpha, \rho) =$$

$$\begin{cases} \log(\omega) & \text{for } -\rho\log(\omega) \leq \log(p_1^i) - \log(\alpha) \\ -\dfrac{\log(p_1^i) - \log(\alpha)}{\rho} & \text{for } \rho K^i < \log(p_1^i) - \log(\alpha) < -\rho\log(\omega) \\ -\log(y_k^i/x_k^I) & \text{for } \log(p_1^i) - \log(\alpha) < \rho K^i < \log(p_2^i) - \log(\alpha) \\ -\dfrac{\log(p_2^i) - \log(\alpha)}{\rho} & \text{for } 0 < \log(p_2^i) - \log(\alpha) < \rho K^i \\ 0 & \text{for } \log(p_2^i) - \log(\alpha) < 0 < \log(p_2^i) + \log(\alpha) \\ -\dfrac{\log(p_2^i) + \log(\alpha)}{\rho} & \text{for } \rho\log(\omega) < \log(p_2^i) + \log(\alpha) < 0 \\ -\log(\omega) & \text{for } \log(p_2^i) + \log(\alpha) \leq \rho\log(\omega) \end{cases}$$

Then, for each subject, we choose the parameters α, ρ to minimize:

$$\sum_{i=1}^{50} (\log(x^i/y^i) - f(\log(p_1^i), \log(p_2^i), K^i; \alpha, \rho))^2 \tag{3}$$

using NLLS with standard errors computed using a robust variance estimator. Tables 19-26 in Appendix G of the Supplementary Online Material present the estimates $(\hat{\alpha}, \hat{\rho})$ that result for each individual in each of the linear and kinked treatments. In Section E.2 of this Appendix, we go into greater detail about some of the finer points and particulars of the estimation procedure that generates these tables, but before that, in Section E.1, we discuss the rationale for our use of the NLLS estimation strategy.

C.1 Choice of Estimation Strategy

Non-linear least squares estimation was chosen over maximum likelihood estimation (MLE) for two reasons.[40] First, NLLS more accurately recovers structural parameters for simu-

[40] The MLE estimator comes about from the maximization problem for an individual with utility defined by equation (2), which yields the following conditions used to define the log-likelihood function (with the

lated subjects, especially as the variance of the stochastic component increases. We generated the simulated subjects using the utility function in (2) with $\varepsilon \sim N(0, \sigma_\varepsilon)$ for a suite of structural parameter combinations (α, ρ) (which span the plurality of the structural parameters estimated in the data). We randomly choose 50 linear and 50 corresponding kinked budget sets from the actual budget sets offered to subjects in the study and then simulated the choices that 100 simulated subjects would make in each budget set as dictated by their assumed parameter values and the specified error-generating process. Using these choices, we could then reverse engineer (with either NLLS or MLE) a recovery of the implied structural parameters that would be expected to generate the choices of each of the simulated subjects, calculate the average of these estimated structural parameters, and then compare them to the average of the true values that generated the choices. For small values of σ_ε (0.01), both MLE and NLLS perform similarly well, judged by a

log price ratio defined as p and other terms defined as above):

$$\begin{aligned}
p &\geq \log\alpha + \rho\log(1/\omega) + \varepsilon \;\text{ for }\; x/y = \omega \\
p &= \log\alpha + \rho\log(y/x) + \varepsilon \;\text{ for }\; \omega < x/y < 1 \\
-\log\alpha + \rho\log(y/x) + \varepsilon &\leq p \leq \log\alpha + \rho\log(y/x) + \varepsilon \;\text{ for }\; x = y \\
p &= -\log\alpha + \rho\log(y/x) + \varepsilon \;\text{ for }\; 1 < x/y < 1/\omega \\
p &\leq -\log\alpha + \rho\log(\omega) + \varepsilon \;\text{ for }\; x/y = 1/\omega
\end{aligned}$$

For the equality conditions, the probability that ε satisfies that relation is well defined. The inequality conditions characterize an interval of values that ε can take to satisfy that relation. Denoting by ϕ and Φ the normal PDF and CDF, respectively, of the error term's distribution (mean zero with variance σ), the log-likelihood function is given by:

$$\begin{aligned}
\mathscr{L} = & \prod_{\text{for } x^i/y^i=\omega} \Phi[p - \log\alpha - \rho\log(1/\omega)] \\
\times & \prod_{\text{for } \omega<x^i/y^i<1} \phi[p - \log\alpha - \rho\log(y^i/x^i)] \\
\times & \prod_{\text{for } x^i=y^i} [\Phi[p + \log\alpha] - \Phi[p - \log\alpha]] \\
\times & \prod_{\text{for } 1<x^i/y^i<1/\omega} \phi[p + \log\alpha - \rho\log(y^i/x^i)] \\
\times & \prod_{\text{for } x^i/y^i=1/\omega} [1 - \Phi[p + \log\alpha - \rho\log(\omega)]]
\end{aligned}$$

The maximum likelihood estimators $(\hat{\alpha}_{MLE}, \hat{\rho}_{MLE}, \hat{\sigma}_{MLE})$ are then the parameters that maximize this log-likelihood function.

comparison of the average recovered structural parameter value and the average of the actual structural parameters used. However, increasing σ_ε even slightly (e.g., to 0.1 or higher; for reference, the average estimated σ found in the data is significantly greater than that for both MLE and NLLS) makes the NLLS estimation perform markedly better than MLE in recovering parameters. In particular, NLLS is better able to recover more extreme structural parameters for both the linear and non-linear budgets. Tables 17 and 18 in Appendix G of the Supplementary Online Material contain the tests for $\sigma_\varepsilon = 0.1$. Moreover, the ability of NLLS to recover these parameters is not significantly impacted by the linearity of the budget, dovetailing with our second rationale for preferring NLLS detailed below.

The second reason NLLS estimation is preferred over MLE estimation is that the former appears to be less susceptible to bias in the parameter estimates that comes about strictly from switching from estimating choices over linear budget sets to estimating choices over non-linear budget sets. Structural parameters are notoriously difficult to estimate using choices on non-linear budgets, and as we are particularly interested in identifying actual changes in structural parameters that result from a shift from linear to non-linear budget settings, we want to guard against differences in parameter estimates coming about solely because the estimation method itself leads to changes when applied to linear vis-a-vis non-linear budget settings (without there being any actual change in the structural parameters). To further assess whether NLLS/MLE estimation on non-linear budgets accurately recovers parameters without a bias in comparison to the parameters recovered for linear budgets, we translate subjects' linear choices on linear budgets into choices on pseudo-non-linear budgets with an imposed pseudo-kink and then use NLLS/MLE to recover the parameters in both scenarios. Specifically, we take participants' budgets and choices from the linear treatment and add the pseudo-kink by randomly choosing a point between the riskless point (the 45 degree line intersection with the budget set) and the corner of the cheaper asset and classify it as the kink point K^i above (though it does not in fact represent any actual convexity in the budget line). We then apply the non-linear budget set variation of NLLS and MLE to this budget with "kink" K^i to ensure that the same parameters are recovered with this method as with the linear budget set variation of NLLS and MLE (as they should be given that there is no actual change in the data-generating process). NLLS succeeds in this regard, while MLE fails. In particular, MLE biases the ρ values toward zero when non-linearizing linear budgets and choices.

C.2 Computational Particulars of the Estimation Strategy

We now address several particulars of our NLLS estimation strategy. First, to avoid arriving at a local minimum, we use a variety of initial parameter conditions to seed our NLLS optimization algorithm. In particular, for linear budget sets, three α values are used as initial guesses: the largest and smallest values of the relative prices offered to an individual and the average of the two.[41] Additionally, two ρ parameters are used as initial guesses, 0.5 and 1.5. Together, this creates a 3x2 matrix with six potential initial conditions. For the non-linear budgets, these six initial guesses are used and then the estimated structural parameters from the linear budget sets of each participant are used as a seventh initial guess. The program iterates through the initial guesses, performing a kind of branch-and-bound algorithm, keeping the estimated structural parameters that minimize the loss function.

Second, due to data limitations, for each participant, the estimation of α is capped at the largest value among the set of offered price ratios (p_x^i/p_y^i) and their inverse. To illustrate why this is necessary, suppose that a participant's true α is 4.5 but that the largest value of the price ratio (p_x^i/p_y^i) or its inverse that they is ever offered is 4. This participant would choose the riskless asset for every budget he or she is offered (see the fifth line of the log demand function $\log(x^{i*}/y^{i*})$ defined above to see this), and thus any value of $\alpha \geq 4$ would be consistent with their choices, but estimating $\alpha > 4.5$ would be inconsistent with the true α. To discipline the model, we minimize the loss function under the constraint that α cannot exceed the maximum of (p_x^i/p_y^i) or (p_y^i/p_x^i)

[41] This maximum and minimum of the relative prices offered are the maximum and minimum values (across all 50 budget sets) of the ratio of the commodity with the higher price (for a given budget set) over the commodity with the lower price (for the same budget set). See the second point in this subsection for why this range bounds alpha.

offered to the participant (as would be expected given the randomization procedure for assigning budget constraints, there is no significant difference in the distribution of these maximum prices across types). In essence, data limitations restrict what can be known about subjects' structural parameters. Importantly, the maximum prices in the linear and non-linear budgets differ, with the non-linear budgets having larger maximum prices by construction. Thus, *after* estimating α for the non-linear budget, we restrict it to not exceed the α constraint from the linear budgets; otherwise, the α values would be biased when we compare the structural parameters across the linear and non-linear treatments.

Third, choices at or near the riskless point on the budget are reclassified. The structural estimation is sensitive to small deviations around the riskless point on the budget curves. For example, suppose a participant chooses the point $(x^i, y^i) = (33.5, 33)$ for a linear budget $(x^{\text{Max}}, y^{\text{Max}}) = (50, 100)$. While this choice is close to the riskless point of the budget, the participant has chosen a higher portion of the expensive good, resulting in an upward-sloping demand curve. Moreover, suppose a participant chooses the point $(x^i, y^i) = (33.3, 33.4)$ for the same linear budget $(x^{\text{Max}}, y^{\text{Max}}) = (50, 100)$. While this choice is consistent with a downward-sloping demand curve, the deviation from the pure riskless choice of $(x^i, y^i) = (33.33, 33.33)$ could be due to price sensitivity or computational rounding of choices to the nearest tenth place when subjects are offered choices. Additionally, subjects may be myopic when making decisions around the riskless point, perceiving $(x^i, y^i) = (33.3, 33.4)$ and $(x^i, y^i) = (33.33, 33.33)$ as negligibly different choices (on the visual interface, they appear very much the same). The structural estimation procedure, on the other hand, treats these alternative choices as fundamentally different, with the potential to markedly change the estimated parameters. Thus, to be conservative about subjects' price sensitivity, we reclassify all choices such that $|x^i - y^i| \leq 1$ as if they are at the riskless point, $x^i = y^i$, on the linear and non-linear budgets.

Fourth, extreme outliers are removed from the structural estimation using quartile outlier detection. As noted in Choi et al. (2007a), the structural estimation is sensitive to outliers. This sensitivity is especially pronounced for participants with a majority of points around the riskless point of each budget. To make our estimation robust to extreme outliers in a formal and non-ad hoc manner, we employ quartile outlier detection, an outlier detection method used for choices that are not distributed normally (Hodge and Austin, 2004; Rousseeuw and Hubert, 2011). Specifically, we calculate the interquartile region of the absolute value of the log choice ratio $iqr = |\log(x^i/y^i)|_{.75} - |\log(x^i/y^i)|_{.25}$ for each subject and remove all choices that exceed $20iqr$ below the lower quartile or $20iqr$ above the upper quartile of choices from the structural estimation. This process is performed on both the linear and non-linear treatments separately. This large band was chosen to identify choices that seem to be outliers (e.g., see ID 220) but to not remove choices for other participants whose extreme choices may result from dramatic changes in prices. For instance, if a participant's true structural parameters were a large value of α and a zero value of ρ, a utility-maximizing subject would pick the riskless point of the budget for prices below a certain α and the corners for prices above α. The outlier detection outlined here would not remove the corner choices of this individual, as shown in Figure 20 in Appendix G of the Supplementary Online Material, which presents the choice plots for two individuals: one whose outlier choice is removed by the procedure and one whose choice is not. In any case, the modal and median number of outliers removed for an individual in the linear and non-linear treatments is zero, and the mean is less than 1 in both treatments.

Fifth, to make the estimation robust to smaller outliers and larger error processes, σ_ε, we use weighted non-linear least squares on subjects' choices. Specifically, we employ bi-square weighting with the common tuning parameter $\kappa = 4.685, w(r, \kappa) = 1(|r < \kappa|)(1 - (r/\kappa)^2)^2$, where r is the estimated residual (Huber (2004)). As is common with weighted non-linear least squares, this weighting process is evaluated iteratively as in Holland and Welsch (1977). We use the residuals from the last iteration of non-linear least squares estimation process to weight the loss function in the current iteration. This iteration continues until the loss function is below a function tolerance or the normed difference between the last iteration's estimated structural parameters and the current iteration's estimated structural parameters are below a tolerance band.

Thus, for each subject, we choose the parameters α, ρ to minimize:

$$\sum_{i \notin \mathbb{O}}^{50} w(r_{k-1}, \kappa)_i (\log(y^i/x^i) - f(\log(p_1^i), \log(p_2^i), K^i; \alpha_k, \rho_k, \omega))^2$$

$$\text{s.t. } \alpha_k \leq \max \{p_1^i, p_2^i, (p_1^i)^{-1}, (p_2^i)^{-1}\}_{i=1 \notin \mathbb{O}}^{50}$$

Where $\mathbb{O}$ is the set of outliers for each subject and $w(r_{k-1}, \kappa)$ is the bi-square weighting vector, $r_{k-1} = \log(y^i/x^i) - f(\log(p_1^i), \log(p_2^i), K^i; \alpha_{k-1}, \rho_{k-1}, \omega)$ is the vector of residuals from the previous iteration of non-linear least squares, and k denotes the iteration of the weighted non-linear least squares algorithm. As noted, Tables 19-26 in Appendix G of the Supplementary Online Material present the results of this estimation, $(\hat{\alpha}, \hat{\rho})$, for each individual in each of the linear and kinked treatments.