Revealing Risky Mistakes through Revisions *

Zachary Breig †
University of Queensland

Paul Feldman ‡
Johns Hopkins University

First Draft: November 27, 2019

This Version: December 16, 2020

Abstract

We argue that choices that are modified, absent any informational change, can be characterized as mistakes. In an experiment, we allow subjects to choose from budgets over binary lotteries. To identify mistakes, which we interpret as deviations from optimizing behavior, we allow subjects to revise a subset of their initial choices. The set of revised decisions improve under several standard definitions of optimality. These mistakes are prevalent: subjects modify over 75% of their initial choices when given the chance. Subjects make mistakes more often when inexperienced and when choosing over lotteries with small probabilities of winning.

JEL classification: C91, D81, D91

Keywords: mistakes, risk preferences, uncertainty, revealed preference, expected utility, experiment.

*We are very grateful to David Arjona Rojo, Soo Hong Chew, Paul Ferraro, Andrei Gomberg, Kenan Kalayci, Edi Karni, Jim Murphy, John Quah, Karl Schlag, and Tom Wilkening for helpful comments. This research was funded by the National Science Foundation's dissertation grant #1824121. *Mistakes* are ours.

†z.breig@uq.edu.au

‡pjfeldman@jhu.edu

1 Introduction

Mistakes are integral to decision making. Parents tell their children to learn from their mistakes, and political leaders tell their constituents that "mistakes were made." In academic contexts, researchers sometimes refer to failures to optimize some particular objective or adherence to a "biased" decision rule as a mistake. However, this goes against the canonical approaches of revealed preference, and decision makers often may not agree that their choices are mistakes. This, then, raises our research question: how can a researcher identify mistakes when underlying preferences are not known to the researcher *a priori*?

We propose and carry out a methodology to study mistakes, which we interpret as deviations from optimizing behavior. Specifically, we argue that if a choice is revised *without any new information or change in circumstances*, then either the initial choice or the revision must be a mistake. This approach can be used in any choice environment and does not rely on the researcher's evaluation of the correct choice. We use this intuition to study mistakes in a laboratory experiment. We find that when offered the chance to revise earlier choices, subjects overwhelmingly do so. Subjects' revised choices are better according to every normative measure we employ, suggesting that the initial choices are mistakes and stationary models of (random) choice cannot explain their revisions. We then study how the characteristics of decision problems affect the prevalence of mistakes.

In our experiment, 181 undergraduates at the University of Queensland make choices over binary lotteries. Following Andreoni and Harbaugh (2009), subjects trade off the chance of a positive outcome p against the size of that positive outcome $\$x$. Feasible choices satisfy a linear budget constraint of the form $x + \frac{M}{m}p = M$ where M is the maximum outcome, and m is the maximum chance. Our subjects know they will choose over the same twenty-five budget sets twice. Subjects are informed about the complete set of budgets and that any of these fifty tasks can be chosen for payment. After choosing from these fifty budgets, subjects learn that they will revise a random subset of thirty-six of their initial choices. Revision choices feature a 2×2 within-subject treatment that changes the presentation of the tasks. One dimension of treatment adds a reminder of what was initially chosen, while the other dimension allows the subject to revise two choices from the same budget at the same time.

We find that when given a chance, subjects consistently revise their earlier choices. Over 75% of choices are revised, and 176/181 of subjects make at least one revision. Moreover, a majority of these revisions are meaningful: over 40% of revisions shift at least 10% of a subject's budget from one good to the other.

Revisions, when compared to the initial set of choices, improve consistency with a number of normative criteria. First, revisions decrease the number of violations of first-order stochastic dominance (FOSD). Second, revised choices are closer to being rationalized by an increasing utility function and an increasing utility function that satisfies FOSD. Third, this relationship is preserved over the conventional functional families of expected utility and probability weighting. Fourth, revised choices are more likely to be consistent with risk aversion. Finally, making identical choices across repetitions of the same budget increases for revised choices, although this type of stationarity only increases when both choices on the same budget are revised on the same screen. Given that either the original choices or their revisions are mistakes, the fact that revisions are more consistent with optimizing behavior, regardless of how much structure is placed on preferences, suggests that the initial choices are mistakes.[1]

Given that revisions indicate that initial choices contained mistakes, as a proof of concept we show that revisions can be used to study the drivers of mistakes. In particular, we study under what conditions these mistakes are made. First, the type of revision opportunity that subjects face affects revision behavior. We find that giving a subject a reminder about the choice they made earlier decreases the likelihood that they make a revision by 17 percentage points while offering them the chance to revise two choices at once increases the chance of making a revision by just under three percentage points. Second, the effect of decision times on revisions is nuanced. Controlling for subject fixed effects, the amount of time spent making a choice is positively correlated with the size of revisions, but this correlation is driven by the *negative* correlation between experience and time spent. Finally, subjects tend to make more and larger revisions when the budget set contains only lotteries with low probabilities of receiving a monetary prize.

There are several rival explanations for revisions that are unrelated to mistakes. We address them here. First, under a pay-one-choice-at-random mechanism, individuals may want to build a portfolio with their choices. Since revisions *replace* earlier choices, portfolio-building cannot explain any difference between choices and revisions. Second, subjects may be indifferent between both choices and revisions. Because the revised sets have higher normative indices, this seems unlikely. Third, choices and revisions may differ due to randomness from the decision-maker. Some choices may be random; however, the distribution of revisions is distinct from the distribution of initial choices as indicated by the improvement in our normative benchmarks. Hence, choice sets cannot be explained by a *stationary* stochastic choice function. Fourth, subjects may revise because they believe they are expected to. Such experimenter demand effects are improbable because of the neutral framing of revisions. This is in stark contrast with other approaches where subjects are directly confronted with their inconsistencies or arguments about how choices *ought* to be made. Our subjects are simply asked what they would like their revised choice(s) to be, half the time with a reminder of their initial choice(s). Finally, a dual-self model—one "self" makes the original choices and another the revisions—could predict a difference. Temporal contiguity

[1] One may wonder why a violation of these normative measures is not itself an indication of a mistake. While this is likely true for violations of dominance, revisions may reveal mistaken choices even when the option chosen is not dominated. Measures relying on transitivity only reveal that there is a mistake in a *set* of choices and do not show *which* choice is a mistake.

of choices and revisions would rule out most of these models.

What do we think explains these mistakes? Our main focus is to introduce an approach to identify mistakes; distinguishing between specific mechanisms is beyond the scope of this paper. Notwithstanding, we show how our methodology can be applied. For instance, problems that have a higher revision likelihood and magnitude of change are likely more difficult. In this way, we find that subjects struggle more when the probabilities of winning are small.

Revisions can reveal the mistakes subjects make as a result of lack of experience. Subjects may be *learning* about their preferences and our interface after initially having chosen suboptimally. However, unlike standard strategic experiments subjects do not learn the outcome of their choice in the *interim*, but only *ex-post*. Some potential initial confusion about the interface may have lead to a 1.54% of the original choices being dominated. This drops to 0.91% by the revisions stage of the experiment. There are many meaningful contexts, such as investing for retirement or purchasing health insurance, in which this type of unfamiliarity likely contributes to mistakes (Choi, Laibson, and Madrian, 2011; Bhargava, Loewenstein, and Sydnor, 2017).

Mistakes can be a costly part of everyday decision making. A large and growing literature documents ostensible mistakes in the financial domain: Individuals do not efficiently use or pay off their credits cards (Ponce et al., 2017; Gathergood et al., 2019), make sub-optimal mortgage choices (Agarwal et al., 2017), and underreact to taxes that are not salient (Chetty et al., 2009). The existence of mistakes across these domains, where objective decision quality can be assessed, suggests that individuals make mistakes in other consequential domains. Offering a chance to revise a decision may reveal these mistakes even when the researcher has no objective way to evaluate the choice.

The paper proceeds as follows: Section 2 discusses related literature. Section 3 presents the choice environment for binary lotteries. Section 4 describes the experimental procedures. Section 5 features our results contrasting sets of initial choices and sets of revisions using normative benchmarks. Section 6 explores the determinants of mistakes in the experiment. Section 7 features our final remarks.

2 Related Literature

In this section, we discuss adjacent research. We begin with the implications of our paper for behavioral welfare measures. As we focus on risk, we then proceed by considering the empirical literature on random choice and other revealed preference risk approaches. We then present some field evidence on the implications on wealth from mistakes. Finally, we review the experimental literature on failures to maximize when an objective ranking can be ascertained *a priori* and on different revision incentives.

Identifying mistakes and where people make them is a key step in behavioral welfare economics (Bernheim and Taubinsky, 2018). Some have pointed out that with only weak assumptions on preferences, researchers can identify *mistaken beliefs* held by a decision maker (Koszegi and Rabin, 2008). Bernheim and Rangel (2009) and Bernheim (2016) argue that when choices are made under different frames (or *ancillary conditions*) contradict each other, one may be able to use outside information to determine which choice to respect. One may think about our revision decisions as being from a particular frame, and our results show that choices made in that frame are more consistent with a variety of normative benchmarks. More generally, our work is related to a contemporaneous literature that attempts to identify the decision maker's "true" preferences (Allcott and Taubinsky, 2015; Bernheim et al., 2015; Benkert and Netzer, 2018; Goldin and Reck, 2020). We complement this literature with a focus on understanding the mistakes themselves.

We add to the literature on random choice. There is evidence that when making choices from the same choice set multiple times, subjects do not always make the same choice. This occurs both when the decisions are temporally close and when they are distant (Tversky, 1969; Hey and Orme, 1994; Hey, 2001; Agranov and Ortoleva, 2017). In our experiment, all choices are made in a single sitting. Our design features revisions in addition to the more standard repetitions. These revisions *replace* subjects' earlier choices, implying that the difference between revisions and the initial set should not be due to subjects building a portfolio.

The use of revealed preference for the study of risk preferences in experiments is not unique to our study. Choi et al. (2007) uses revealed preferences to study consistency with rationality in a study where subjects choose between arrow securities using budgets. Halevy et al. (2018) employs the same data set and a separate experiment to correlate consistency with rationality to parametric fit using predicted behavior as a benchmark. Our revealed preference approach is closer to Polisson et al. (2020). They provide revealed preference tests for different functional specifications and use them to analyze the Choi et al. (2007) and the Halevy et al. (2018) data sets. We adapt their results to budgets over simple binary lotteries and use their finite-data revealed preferences' measures—adapted to various specifications—to reveal mistakes.

Prior research examines how violating specific norms is correlated with real outcomes and financial decisions. Jacobson and Petrie (2009) shows that subjects who make choices that are inconsistent with a class of theories of choice under risk do not choose optimally over non-experimental financial instruments. Choi et al. (2014) finds that experimental measures of rationality correlate with wealth and education. Rather than using predetermined normative criteria, our measure of a mistake is revealed by the decision makers themselves.

Other studies have considered choice behavior when choices can be objectively ranked, but these rankings must be determined by the decision maker through arithmetic calculation.

Caplin et al. (2011) documents departures from full rationality and towards a satisficing heuristic in search problems. Kalaycı and Serra-Garcia (2016) finds that adding complexity leads to choices that decrease overall payoffs. Gaudeul and Crosetto (2019) finds that adding this sort of complexity can induce the attraction effect in decision makers, but that they eventually make more informed decisions. Martínez-Marquina et al. (2019) finds that adding uncertainty impedes subjects' ability to maximize their payoff. Our identification of mistakes does not rely on there being an optimal choice that the experimenter knows, but the decision maker does not.

Recent work documents how decision makers reconcile potentially inconsistent prior choices. Benjamin et al. (2019) offers subjects hypothetical choices over retirement savings options and confronts them with choices that may be inconsistent. Nielsen and Rehbeck (2019) finds that subjects report a desire for their decisions over lotteries to satisfy several axioms and that a majority of subjects revise their choices if they find that these choices violate the axioms. Yu et al. (2019) finds that a nudge causes subjects to revise their choices in a way that reduces multiple switching in a price list. The majority of the revision opportunities in our experiment did not give any indication to the subject that there were inconsistencies in their choices.

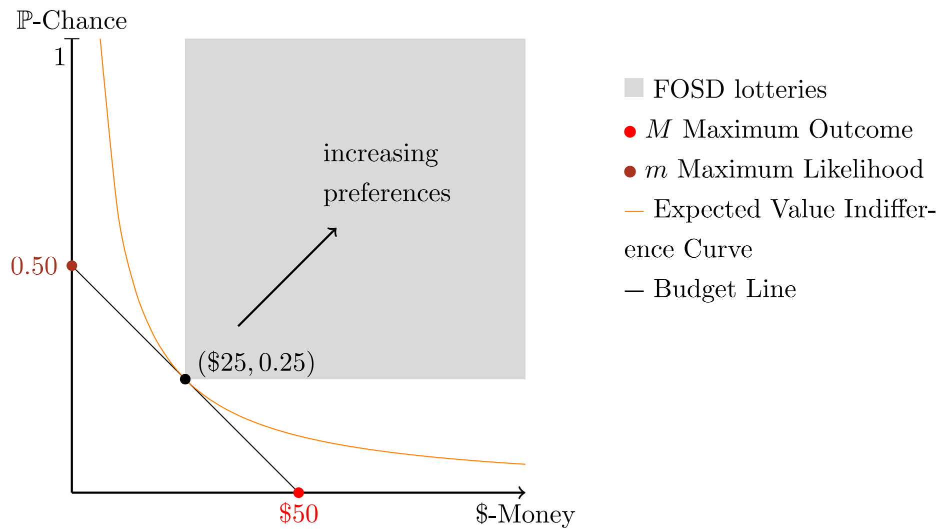
3 Choice Environment

We begin this section by describing our choice environment and some properties of risk preferences. We then show how a decision maker with a canonical form of expected utility preferences makes choices in this environment. We conclude by discussing how we evaluate the concordance of sets of choices with various theories.

Preferences are defined over simple binary lotteries. A simple binary lottery is a lottery that has at most two outcomes, one positive outcome $\$x$ with probability p and $0 with probability $1 - p$. Because one outcome is always $0, we will abuse notation to represent each lottery by the pair $(\$x, p)$.

The choice problem involves a tradeoff between x and p using a linear budget. Each budget can be described by its maximum prize $M \in \mathbb{R}_{++}$ and maximum probability $m \in (0, 1]$. Thus, any choice from the budget must satisfy $x + \frac{M}{m}p = M$, such that $\frac{M}{m}$ is the "price" of increasing the likelihood of receiving the prize. With this construction, corner allocations on a budget line will always yield a certain outcome of $0.

Figure 1: Two-goods Diagram for Binary Lotteries



Notes: The decision maker faces a single budget with endpoints $m = 0.5$ and $M = 50$. An expected value maximizer would choose the option $(\$25, 0.25)$, and the indifference curve that this point is on is given in orange.

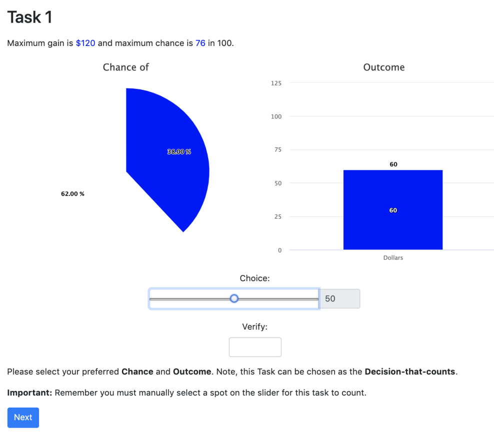
Figure 1 shows how we can plot lotteries, budgets, and increasing preferences using the familiar two-goods diagram. An expected value maximizer would maximize $p \cdot x$, leading to choices $\$x^* = .5M$ and $p^* = .5m$. This highlights two features of expected utility: first, we may restrict attention to (x, p) without loss of generality, and second, any risk-neutral agents devote half their budget to x. Consequently, any risk-averse (risk-tolerant) expected utility maximizer will allocate a budget share of more (less) than one-half to probability.

Now, consider an expected utility maximizer with CRRA preferences given by $u(x) = x^\alpha$. An increasing transformation can be applied to the agent's objective function to obtain $p^{\frac{1}{1+\alpha}} x^{\frac{\alpha}{1+\alpha}}$. Thus, these preferences can be represented by a Cobb-Douglass utility function, and the budget shares the decision maker chooses will be constant across budgets.

In our results, we will opt for non-parametric revealed preferences tests. In particular, we will use Afriat's theorem first to determine whether an increasing, concave, and continuous function can rationalize our data. Second, we will use a generalization of Afriat's theorem (Nishimura et al., 2017; Polisson et al., 2020) that allows us to test for the ability of specific functional forms to rationalize our data and extend a standard measure of rationality. The functional forms we consider are expected utility $(p * u(x))$ and generalized probability weighting $(\pi(p) * v(x))$.

4 Experimental Design

Figure 2: Experimental Task Summary



(a) Sample Task

Maximum Chance	Maximum Outcome
50 in 100 chance	$30
60 in 100 chance	$30
50 in 100 chance	$40
80 in 100 chance	$40
100 in 100 chance	$50
60 in 100 chance	$60
75 in 100 chance	$60
100 in 100 chance	$60
40 in 100 chance	$80
80 in 100 chance	$80
100 in 100 chance	$80
15 in 100 chance	$100
20 in 100 chance	$100
100 in 100 chance	$100
15 in 100 chance	$120
30 in 100 chance	$120
60 in 100 chance	$120
30 in 100 chance	$150
20 in 100 chance	$160
40 in 100 chance	$160
80 in 100 chance	$160
25 in 100 chance	$200
30 in 100 chance	$200
40 in 100 chance	$200
50 in 100 chance	$200

(b) Full Set of Distinct Tasks

Notes: Panel A shows a sample choice task. Panel B summarizes the full set of budgets as it was presented to our subjects.

For each task, we elicit subjects' preferences over the set of binary lotteries—lotteries that give $\$x$ with probability p and \$0 otherwise—in a linear budget with endpoints $\{M, m\}$. The ratio of M to m gives the tradeoff between the size of the outcome and its likelihood. We emphasize three advantages of using this method. First, because budgets are linear in the $(\$x, p)$ plane, most notions of consumer theory can be applied.[2,3] Second, because setting either $\$x$ or p equal to 0 is strictly dominated, choices will typically be interior. This is beneficial because corner choices pose identification issues for budget-based methods. Third, in contrast to other linear budgets over lotteries (for example Feldman and Rehbeck (2019) for probabilities or Choi et al. (2007) for outcomes), this method features variation in both the probabilities and the outcomes simultaneously. A sample task, as subjects saw it, appears in Figure 2a.

[2]Only compactness and downward comprehensiveness are necessary for revealed preference tests, see Nishimura et al. (2017) for a detailed explanation.

[3]This, of course, requires for preferences to be monotonic in money and the probability of receiving money. This is an assumption we maintain throughout the paper.

Subjects select their preferred lottery from each budget using a slider. Before making each choice, no information is displayed on the subject's screen other than the maximum outcome and the maximum chance. Once a subject interacts with the slider, a pie-chart is used to represent probabilities and a bar-chart represents the positive monetary amount. As the subject moves the slider to the right (left), the pie-chart increases (decreases) and the bar decreases (increases). Once a subjects has identified their preferred bundle, they confirm their selection by separately entering it in a box.

Figure 3 summarizes the budget sets used. The fact that the budgets cross allows for analysis of traditional rationality measures. The set also includes parallel budgets and pure price shifts to allow for analysis of income and substitution effects. A pre-analysis plan was submitted to the AER RCT registry (AEARCTR-0004572) prior to the experiment and the visual interface was coded using oTree (Chen et al., 2016).[4]

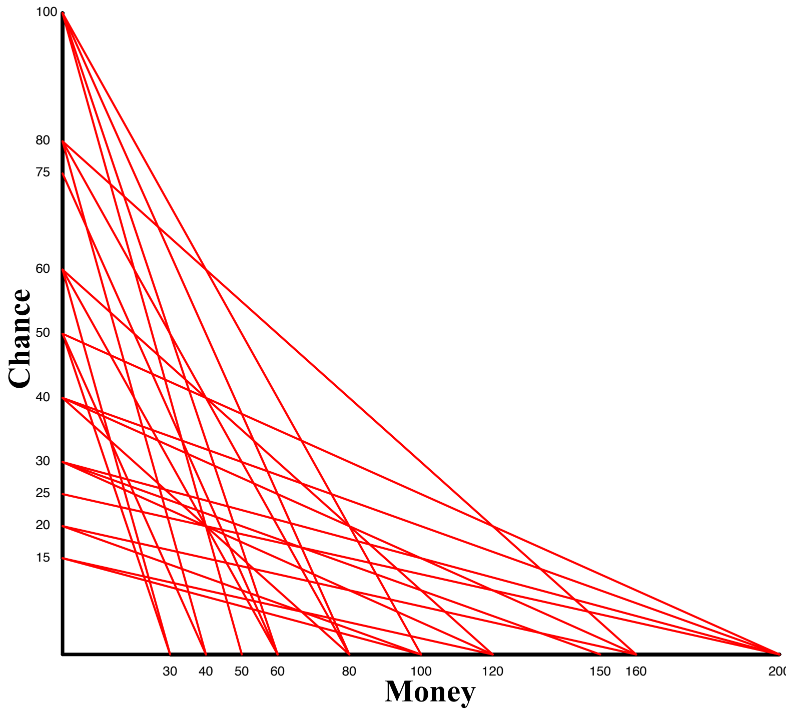
One hundred and eighty-one University of Queensland undergraduates read the instructions on their computer terminal while the experimenter read the instructions aloud. Before starting the main part of the experiment, subjects completed three sample tasks.[5] These examples familiarize the subjects with how the slider affects positive outcomes, chances, and the tradeoff between them. The experiment itself has two parts: repetitions and revisions.

In Part I of the experiment, subjects made choices in 50 tasks. The twenty-five different budgets that were used were described to subjects by presenting them with a list of the pairs of maximum outcomes and chances during the instructions. The information, as subjects saw it, is summarized in Figure 2b. Each subject chose from the twenty-five unique budgets followed by choosing from the same twenty-five budgets for a second time. However, the order across subjects and for each block was random.

[4] A link to the pre-analysis plan and a discussion of changes to our empirical strategy appear in Appendix B.

[5] Sample tasks and the complete instructions appear in Appendix C.

Figure 3: Budgets



Notes: This figure plots the full set of our experimental budgets. This figure was not displayed to subjects.

In Part II of the experiment, subjects revise a subset of the choices they made in these first 50 tasks. These revision tasks feature a 2×2 within-subject treatment that changes the presentation of the tasks (see Table 1). The first change in presentation is the number of revisions they make within a revision task. Each revision task is either a "single" (in which the subject can revise a single earlier choice) or a "double" (in which the subject can revise two earlier identical tasks on a single screen). The second change in presentation is whether or not subjects are given a reminder of the initial choice they made.[6] The subject makes six revisions in each condition, 36 choices, without replacement, being revised. Thus, no single task is revised twice, and at least one task is revised from 24 of the 25 budgets. The order of treatments is randomized at the subject level.

To incentivize choices, one of the fifty choices was chosen at random from the revised set to determine payoffs. Subjects made an average of 9.5 (19.5 s.d.) Australian dollars (AUD) and received a 10 AUD as a participation payment. Each of the experimental parts took

[6]For revisions with reminders, subjects are shown a pie-chart and bar graph that matched their prior choice. The pie-chart and bar graph are replaced with representations of their current choices as soon as they click on the slider. However, a line of text describing their prior choices remains. For all other choices, the initial graph was empty and the additional line of text is not provided.

Table 1: Revisions by Type

	reminders	no reminders
single choice	6	6
double choices	12	12

Notes: Double choices featured the same choice problem twice over the same budget. Appendix C contains samples for each type of revision.

around 30 minutes on average.

Table 2 provides summary statistics. Each of the 181 subjects made 50 choices in the first section of the experiment, for a total of 9050. Each choice is the portion of the budget (out of 100) which is allotted to increasing the probability of receiving the prize. The average choice was to devote just over 54% of their budget towards probability, indicating mild risk aversion. Subjects spent an average of roughly 24 seconds per task on the first fifty tasks.

Table 2: Summary Statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
Original Choice	9050	54.297	20.746	0	100
Seconds on Page	9050	24.024	17.661	3	375
Made Revision	6516	.752	.432	0	1
Revision	6516	.127	19.581	-100	100
Abs. Revision	6516	11.977	15.491	0	100

Each subject faced 36 revisions problems, for a total of 6516. We say that the subject made a revision if their revision choice differs from their initial choice. When given the choice, subjects make revisions roughly 75% of the time. The size of the revision is the difference in the portion of the budget assigned to probability between the initial choice and the revision. These revisions are on average near zero (indicating that revisions are not on average significantly more or less risky than the initial choices). However, the average absolute value of the revision is nearly 12, indicating that subjects are on average shifting more than 10% of their budget from prize to probability (or vice-versa).[7]

5 Do Mistakes Have Normative Content?

This section examines whether the mistakes we identify are "poor" choices. To decide whether choices are indeed worse, we evaluate them according to traditional normative

[7] Camerer (1989) reports the results of an experiment in which subjects were allowed to revise their choices after the decision which counted was selected but before the gamble's outcome was reported. Only 2 of 80 subjects changed their decision in this case. While some of difference between this and our results is likely due to the size of the budget set, the fact that revisions are so prevalent and sizable in our sample may be surprising.

benchmarks. The first benchmark is picking strictly dominated alternatives (violations of monotonicity), the second benchmark is rationalizability by an increasing utility function, the third benchmark is consistency with various functional forms (including expected utility), the fourth benchmark is consistency with risk aversion, and the fifth benchmark is whether behavior across repetitions is stationary (i.e. choices do not vary across the repetitions).[8]

5.1 Monotonicity

We find that 32/181 subjects violate monotonicity by selecting a corner—a certain outcome of zero—on at least one budget for their initial set of choices. In contrast, 17/181 subjects violate monotonicity when we look at their revised sets of choices. For each subject, the initial set consists of their first 50 choices while the revised set consists of 14 of their initial choices and 36 revisions—the revisions that overwrite their initial choices.

The mean number of corners chosen in the initial 50 budget sets is 0.768, while the mean number of corners in the revised set of 50 choices is 0.525. Furthermore, only three subjects increase the number of corners chosen in their revised set, while 29 subjects decrease the number of corners chosen.

5.2 Rationalizability with an Increasing Utility Function

The next benchmark which we use to compare choices to revisions is rationalizability. Following Afriat (1967) and Varian (1982), we define a set of choices to be rationalized if there exists a utility function which the choices maximize. Because every data set can be rationalized by a utility function (e.g. the constant utility function), we further place the restriction that the utility function which is maximized must be increasing.

Because this rationality test has a binary outcome, it is common to use a more continuous measure. The measure of rationalizability we employ is Afriat's index (AI), which is a number e between zero and one (Afriat, 1973). Mathematically, a lower index reduces the number of restrictions that a utility function has to satisfy: rather than requiring the utility from bundle (x_i, p_i) to be higher than the utility from all bundles which satisfy $x + \frac{M_i}{m_i} p \leqslant M_i$, the utility need only be higher than all bundles which satisfy $X + \frac{M_i}{m_i} p \leqslant eM_i$. The AI for a set of choices is the highest e for which the choices are rationalized. This index has become a common measure for how far a set of choices is from being rationalized (Andreoni and Miller, 2002; Choi et al., 2007; Polisson et al., 2020).

In our context, there are two relevant types of monotonicity. The first is monotonicity in the classic sense: the decision maker strictly prefers a bundle which is strictly higher in one dimension and no lower in any other dimension. In this case, we use the Afriat

[8] The primary focus of this section is comparing choices to revisions. Additional empirical results about these benchmarks can be found in Appendix A.2.

Index as it has been classically defined for any collections of choices from linear budget constraints. Our stronger notion of monotonicity is first-order stochastic dominance. This places the same restrictions as standard monotonicity, but also requires that the decision maker never chooses on the endpoints of the budget line (because any interior choice first-order stochastically dominates the endpoints, which guarantee a payoff of zero). When using FOSD as the notion of monotonicity, a set of choices is assigned an index of zero if it includes any choices on the endpoints of the budget line. Otherwise, it is equal to the standard Afriat Index.

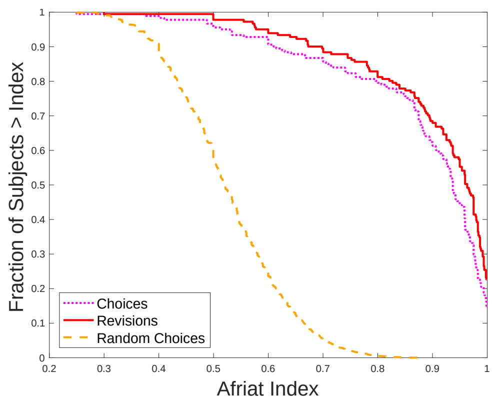
The Afriat indices and Afriat indices under FOSD can be found in Figures 4a and 4b, respectively. The figures also contain the Afriat Index for a uniform random choice rule that measures the power of our design to detect violations of rationality (Bronars, 1987).[9] Clearly, both the Afriat and Afriat FOSD indices of the revised sets of choices first-order stochastically dominate the distributions from the initial sets of choices. Revised decisions are closer to being rationalized by a utility function, indicating that some of the initial decisions may have been of poor quality.

We also report another consistency measure for the maximum acyclic set—the maximum number of choices that could be rationalized by an increasing utility function (Houtman and Maks, 1985; Rehbeck, 2020). This measure appears in Figure 4c and does not alter the result that the consistency of revised choices is always higher for any fraction of subjects.

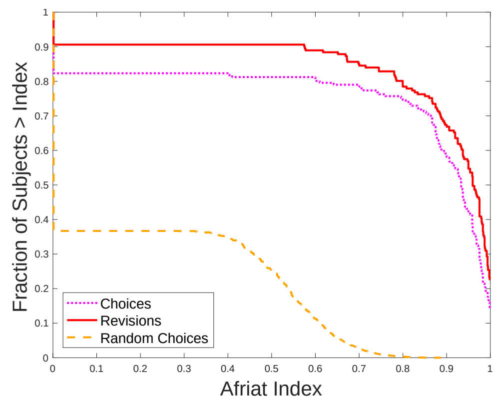
Our general rationalization results are as follows. For their initial choices, 80 subjects have an Afriat Index of at least 95%, 76 subjects have an FOSD consistent Afriat Index of at least 95%, and 95 subjects have their maximum number of consistent choices greater than 47. For their revised choices, the number of consistent subjects increases across all three benchmarks to 100, 99, and 113, respectively. Median consistencies for the initial choices are 94%, 93%, and 47, compared to 96%, 96%, and 48 for the revised choices, across the three benchmarks. A signed rank test rejects ($p<.01$) equality of distributions between initial choices and revised choices for the three benchmarks. Hence, the number of subjects whose choices can be rationalized by some utility function is unambiguously larger for revised choices as implied by these metrics and Figure 4.

[9] Choices on the budgets were discretized to 101 distinct choices that are equidistant on each budget. Our uniform random rule randomizes over the options on a budget subjects could make.

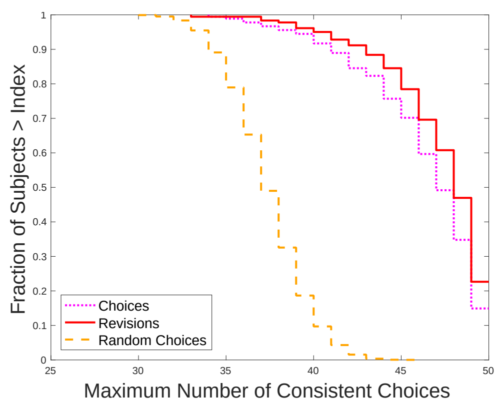
Figure 4: Rationalizibility for Initial Choices and Revised Choices



(a) Afriat Index



(b) Afriat FOSD Index



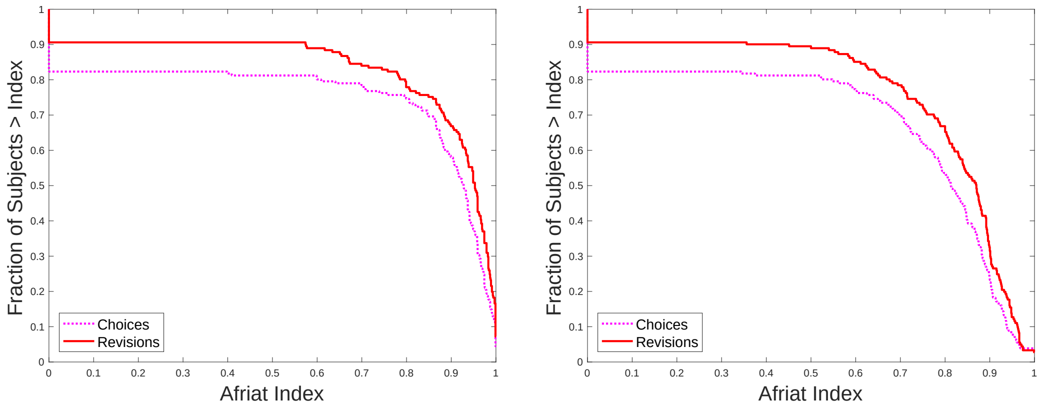
(c) Maximal Acyclic Choices

Notes: These figures contain our main rationality results using Afriat's index (Panel a), Afriat's index under FOSD (Panel b), and maximal transitive relation (Panel c). Each panel contains the fraction of subjects whose rationality index is greater than the x-axis value for their initial choices, their revised choices, and a uniformly random choice rule (n=10,000).

5.3 Consistency with Common Utility Functions

An additional means of evaluating a subject's choices is to establish whether those choices are consistent with a specific normatively appealing utility representation, such as expected utility. Given recent developments in the theory of revealed preferences we can test these specific models of behavior. In particular a corollary of the results in Nishimura et al. (2017) is that any utility functional representation (because it induces a preorder on the set of choices) can be tested by checking for a cyclical mononotonicity condition under that same preorder. We can further adapt results from Polisson et al. (2020) to our context, allowing us to check for these cyclical monotonicity conditions over a finite set of points induced by

Figure 5: Rationalizibility Using Common Utility Functions



(a) Probability Weighting Index

(b) Expected Utility Index

Notes: These figures sum up our main rationality results for probability weighting ($\sum \pi(p_i)u(x_i)$, Panel a) and expected utility ($\sum p_i u(x_i)$, Panel b). Each panel contains the fraction of subjects whose rationality index is greater than x-axis value for the initial choices and the revised choices.

each sequence of choices. Formal details and results are collected in Appendix A.1.

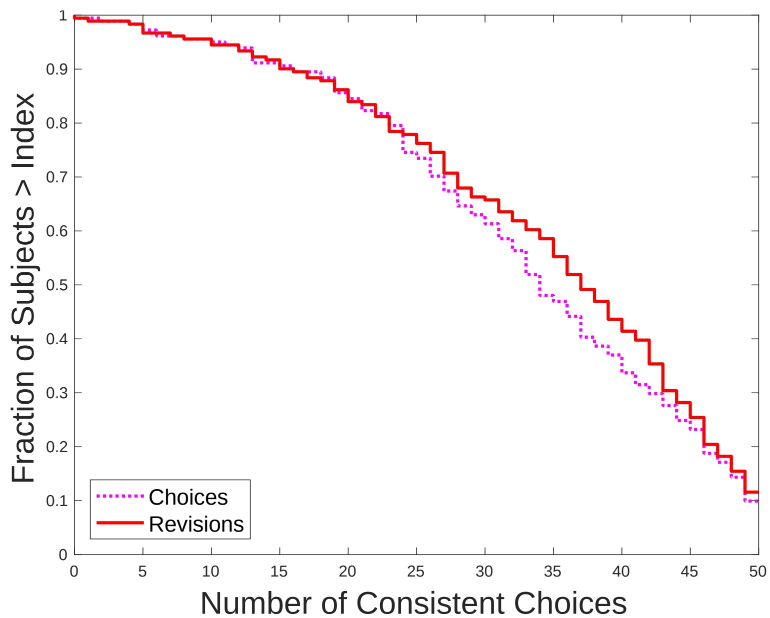
The results of Nishimura et al. (2017) and Polisson et al. (2020) also show that we can use a version of Afriat's index to derive weaker tests of this cyclical monotonicity condition. Essentially, a set of choices will have an index of e if e is the minimum value such that there exists a utility function from the specified family which assigns a utility to each bundle (x_i, p_i) chosen from budget $\{M_i, m_i\}$ that is higher than all bundles that satisfy $X + \frac{M_i}{m_i} p \leqslant eM_i$.

The utility representations we consider are a generalization of Quiggin's (1982) cumulative probability weighting (PW) and expected utility (EU). Because each of these representations places additional restrictions on the previous one and all must satisfy the restrictions from Afriat's theorem, the PW index is lower than the Afriat FOSD index and the EU index is lower than the PW index.

The results for the indices can be found in Figures 5a and 5b. The PW indices of the revised sets of choices first-order stochastically dominate the PW indices of the initial sets of choices. The EU indices of the revised sets of choices *almost* first-order stochastically dominate the EU indices of the initial sets of choices. Thus, when offered the chance, subjects revise their choices in a way that makes them closer to being consistent with commonly used representations.

Our rationality results for the two representations are as follows. For their initial choices, 68 subjects have a PW-consistent Afriat Index of at least 95%, and 14 subjects have an EU-consistent Afriat Index of at least 95%. For their revised choices, the number of consistent subjects increases for both specifications to 92 and 23, respectively. Median consistencies for

Figure 6: Number of Choices that are Consistent with Risk Aversion Across Subjects



Notes: This figure plots the fraction of subjects whose number of consistent risk-averse choices are greater than the x-axis value for both their initial choices and their revised choices.

the initial choices are 93% and 81% compared to 95% and 87% for the revised choices, for the two specifications. A signed rank test rejects (p<.01) equality of distributions between initial choices and revised choices for the two specifications. The number of subjects whose choices can be rationalized by either a probability weighting or an expected utility representation is larger for revised choices as implied by these metrics and Figure 5.

5.4 Risk Aversion

We also discuss a heuristic benchmark for risk aversion. Note that any allocation where the budget shares favor the outcome (x) over the (p) likelihood will be second order stochastically dominated by equal shares—the optimal allocation for an expected value maximizer. Therefore, any concave EU subject—or any risk-averse subject—can never select an allocation the places a greater budget share on the outcome.[10] Our benchmark counts the number of choices that are consistent with FOSD and that place a greater budget share on the probability. As depicted in Figure 6, this measure provides a benchmark for the maximum number of choices that can be consistent with risk aversion.

[10] Note that for a subject to be risk averse it is not sufficient for U to be concave. For example, $U(x,p) = \log(p) + 2\log(x)$ is concave and it represents the same preferences as $V(x,p) = p * x^2$, a risk tolerant utility function. For probability weighting both u and π must be concave for preferences to be consistent with risk aversion (Hong et al., 1987).

We find that 18/181 subjects do not violate risk aversion—on at least one budget—over their initial choices. Revisions lead to a slight increase in the number of subjects that do not violate risk aversion 21/181 at all. 51 subjects increase the number of violations in their revisions, while 100 subjects decrease the number of violations. A signed rank test rejects the null hypothesis that the number of violations of risk aversion is the same across initial choices and revisions ($p < 0.01$). Whether risk aversion is a normatively compelling criterion is a choice for the reader.

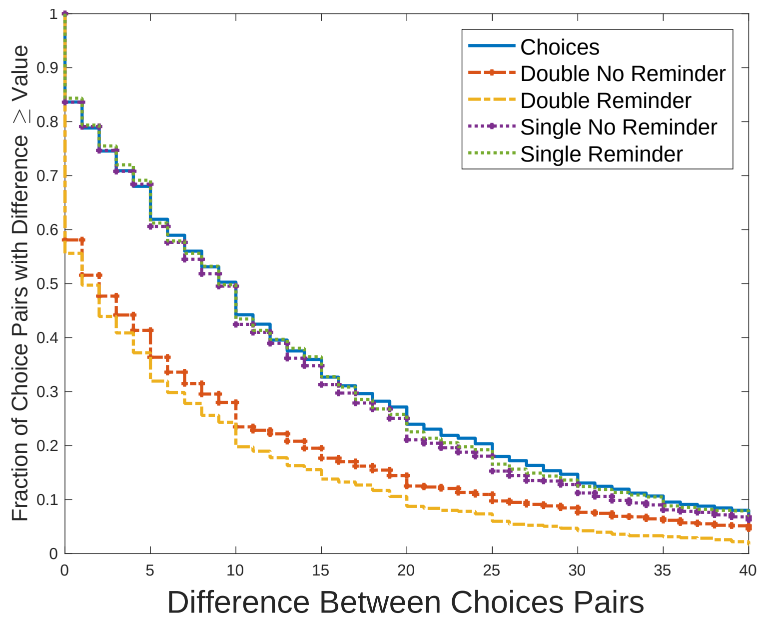
5.5 Stationarity

Only five subjects were stationary across all of their choices.[11] 16.35% of subjects' initial pairs of choices were stationary. When pairing a revised choice in the single revision treatment with its unrevised paired choice, the two are only equal to each other in 16.02% of cases. When two revisions are made at a single moment, they are equal to each other in 43.14% of all cases.

Figure 7 plots the distributions of differences between pairs of decisions in these cases. It is immediate that allowing for a single decision to be revised does not necessarily mean that this revised choice will be any closer to its paired choice than the initial choice was—there is essentially no difference between the CDFs of differences between the initial choices and the single revision problems. On the other hand, there is a clear shift to the left of the distribution of differences when two choices are made at once. Signed-rank tests for equality of distributions of differences between initial sets and revised sets gives a $p = 0.02$ for single revisions and $p < 0.01$ for double revisions.

[11] These five subjects maximized expected value by choosing exactly in the middle of the budget line.

Figure 7: Non-Stationarity in Choice Behavior



Notes: This figure plots the fraction of choice pairs that are inconsistent with stationarity across our experimental treatments. The x-axis captures how far apart choices were across the repetitions in terms of the percentage of the budget allocated towards increasing the prize.

Repeated choices should theoretically match under two criteria. First, preference over single budgets should have a unique maximizer. Second, preference must be dynamically consistent and consistent with consequentialism (Machina, 1989). The latter criteria is satisfied by expected utility. The former is satisfied only if preferences over single outcome lotteries are strictly quasiconcave. For instance, Friedman-Savage expected utility preference can violate stationary.[12] Thus, stationarity is not a property of expected utility. Note further that non-stationarity is implied by *preference for randomization*. Often this behavior has been associated with quasiconcavity in the probabilities, but in our context quasiconvexity in the probabilities can also accommodate it.

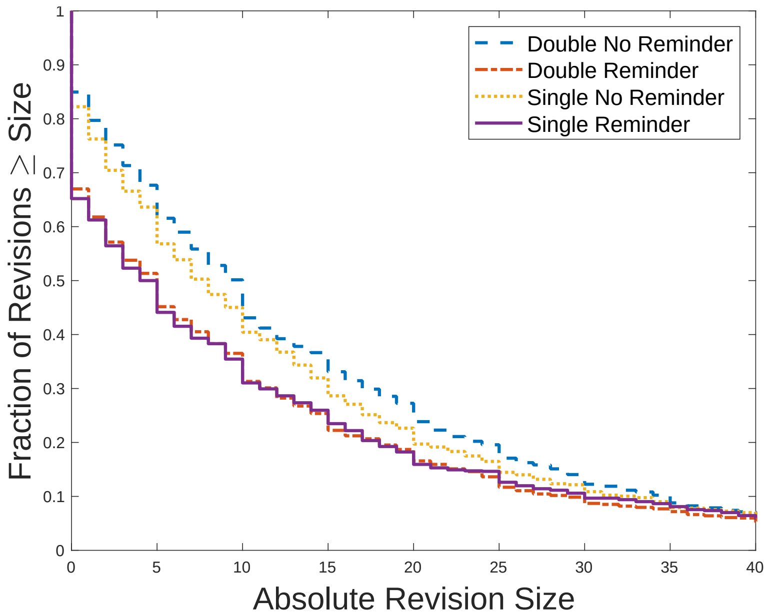
6 Mistakes and Their Determinants

This section discusses the characteristics of the decision problems over which subjects made mistakes. As discussed previously, we label a decision a "mistake" if when given the chance to revise the decision without any new information, the subject decides to make a revision. Subjects were offered the chance to revise 36 of their 50 decisions. Just over 75% of the initial choices were revised when subjects were offered the chance. These revisions could

[12]An example can be provided upon request.

have made the decision less risky (a positive revision) or more risky (a negative revision). Revisions were on average near 0 (mean of 0.127 with clustered standard error 0.603), indicating that subjects did not on average revise their decisions towards probabilities or outcomes.

Figure 8: Absolute Size of Revisions



Notes: This figure showcases the relationship between the initial choices and the revised choices by measuring the distance between them. The curves represent the fraction of choices whose distance was greater than the x-axis value across the experimental treatments. The x-axis is measured in terms of the percentage of the budget allocated towards increasing the prize.

Despite subjects not revising towards one direction or the other on average, the mean absolute value of revisions was 11.977 (clustered s.e. 0.634). This represents over 10% of subjects' budgets. This is not the result of a few outliers: Over 30% of choices had an absolute revision of at least 15.

6.1 Treatments and the Likelihood of Revisions

Figure 8 graphically represents the effects that treatments have on revisions. It shows the distribution of absolute revision size for each of the treatments. Offering subjects a reminder of their previous decision tends to make it less likely that they will revise that decision.

Table 3 shows the effects that treatments have on revisions in regression form. Columns (1) and (2) report how the likelihood of making a revision changes with treatments, while columns (3) and (4) show how the absolute value of revisions change with treatments. The treatment effects are consistent in all cases. Reminding subjects of what they chose previously both makes the subject less likely to revise and makes the average absolute revision smaller. Giving the subject two revisions at once makes subjects slightly more likely to revise and increases the size of revisions. The interaction of these treatments makes revisions less likely and the absolute size of revisions smaller, but only the latter of these effects is significant at the 10% level.

Table 3: Treatment Effects

	(1) Made Revision	(2) Made Revision	(3) Abs. Revision	(4) Abs. Revision
Reminder	-0.17***	-0.17***	-2.27***	-2.21***
	(0.022)	(0.022)	(0.63)	(0.63)
Double	0.027**	0.028**	1.19**	1.24**
	(0.013)	(0.013)	(0.60)	(0.61)
Reminder × Double	-0.0092	-0.0097	-1.30*	-1.41*
	(0.022)	(0.022)	(0.74)	(0.75)
Constant	0.82***	0.83***	12.7***	13.3***
	(0.019)	(0.026)	(0.72)	(0.90)
Subject FE	No	Yes	No	Yes
Task FE	No	Yes	No	Yes
Observations	6516	6516	6516	6516

Notes: Linear regression clustered at the subject level. Each column represents a different regression, with the column head specifying the dependent variable. Significance indicated by: *** p<0.01, ** p<0.05, * p<0.1.

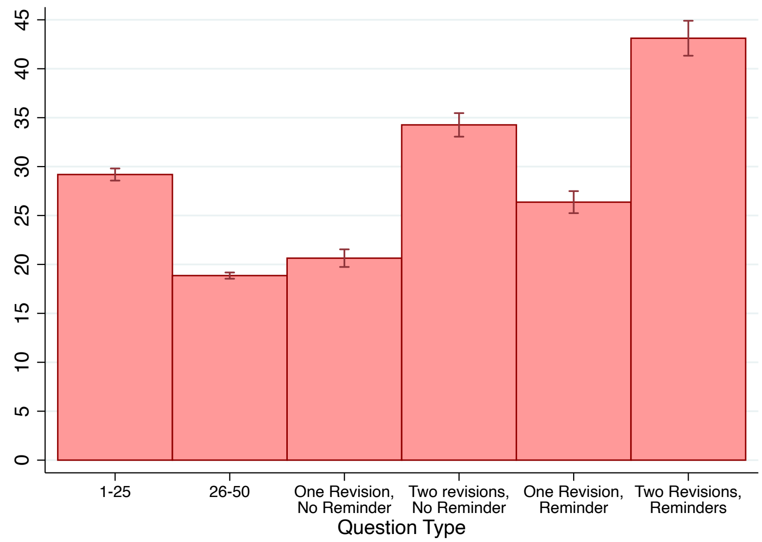
6.2 Decision Times

The amount of time that subjects took to complete each type of problem can be found in Figure 9. Single choices take less time than double choices over the same budget and on the same screen. Earlier choices and choices with reminders also take more time. The average time taken on the first portion of the experiment was just over 24 seconds per task.

The likelihood of revision does vary with the time taken to make the initial decision. This can be seen in Figure 10. The relationship appears to be nonlinear: decisions that are taken very quickly are revised less often, but outside of this range time taken is negatively correlated with revision rates. However, this relationship is not causal. Because subjects are not randomly assigned to time taken, unobservable characteristics of the subject or decision problem may be driving the relationship between decision time and mistake rates.

The relationship between decision time and revisions is further explored in Table 4. The dependent variable in this table is the absolute size of revisions. Column (1) shows that over all observations, the amount of time spent on making a decision is uncorrelated

Figure 9: Time Taken by Decision Type

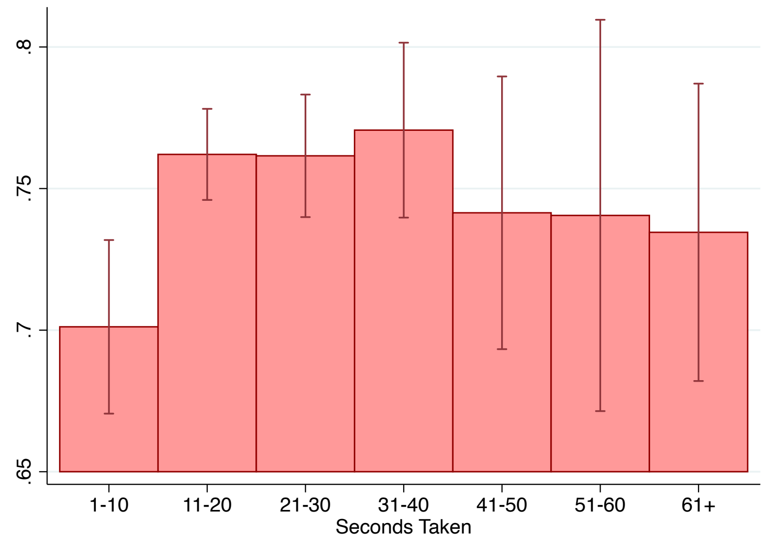


Notes: This figure shows average time spent on a task's page for various decision types. The height of the bar gives the sample mean for each category of decision and the thinner lines give the 95% confidence interval for the mean.

with the amount that this decision is revised. However, Column (3) demonstrates that after controlling for both subject and task (i.e. budget set) fixed effects, there is a positive correlation between between time taken and revision size.[13] This suggests that subjects who make decisions slower make smaller revisions, but that conditional on the subject, spending more time on a decision is associated with larger revisions.

[13]The difference in coefficients from time taken is due almost entirely to the addition of subject fixed effects rather than task fixed effects.

Figure 10: Revision Rates by Time Taken



Notes: This figure shows how the likelihood of a revision varies with the amount of time spent on the initial choice. The height of the bar gives the sample mean for each time window and the thinner lines give the 95% confidence interval for the mean. Decisions which were made very quickly were less likely to be revised, but outside of that range the time taken on a decision is negatively correlated with revision rates.

Columns (2) and (4) of Table 4 additionally control for the round the decision is made in, which varies between 1 and 50. When controlling for the round, the relationship between time taken and the size of revision is both small and statistically insignificant. After controlling for individual fixed effects, the relationship between time taken and the size of revisions is driven by the fact that subjects both take longer and make more mistakes when they are less experienced.

6.3 Budget Characteristics

Tables 5 and 6 study how the characteristics of the budgets affect revisions. Since a budget is completely characterized by its endpoints, we regress revision rates and revision size on these endpoints.

Table 5 shows the linear relationship between the size of budgets and the size and likelihood of making a revision. The coefficient for both regressions on the maximum prize is near zero. Thus, the potential size of the prize does not affect the likelihood that the decision maker makes a mistake. This contrasts with the coefficient on the maximum likelihood of receiving the prize, which is significantly negative. This implies that subjects have a harder time making choices when the probabilities that they are choosing between are small.

Table 4: Decision Time

	(1) Abs. Revision	(2) Abs. Revision	(3) Abs. Revision	(4) Abs. Revision
Seconds on Page	0.00031	-0.025	0.033***	0.0069
	(0.020)	(0.023)	(0.012)	(0.013)
Round		-0.083***		-0.068***
		(0.018)		(0.017)
Subject FE	No	No	Yes	Yes
Task FE	No	No	Yes	Yes
Observations	6516	6516	6516	6516

Notes: Linear regression clustered at the subject level. Each column represents a different regression, but all columns use the absolute value of the revision as the dependent variable. Significance indicated by: *** p<0.01, ** p<0.05, * p<0.1.

Table 5: Budget Characteristics

	(1) Made Revision	(2) Made Revision	(3) Abs. Revision	(4) Abs. Revision
Max Prize	-0.00012		-0.0023	
	(0.00011)		(0.0039)	
Max Probability	-0.095***		-2.32**	
	(0.025)		(1.00)	
Round	-0.00085**	-0.00087**	-0.071***	-0.071***
	(0.00034)	(0.00034)	(0.015)	(0.015)
Subject FE	Yes	Yes	Yes	Yes
Task FE	No	Yes	No	Yes
Observations	6516	6516	6516	6516

Notes: Linear regression clustered at the subject level. Each column represents a different regression, with the column head specifying the dependent variable. Significance indicated by: *** p<0.01, ** p<0.05, * p<0.1.

Table 6 repeats the analysis of Table 5 in a more flexible way. In particular, it uses dummy variables for each maximum prize when estimating the effect of changing the maximum probability, and it uses dummy variables for each maximum probability when estimating the effect of changing the maximum prize. These results largely confirm the results from Table 5: the maximum size of the prize does not affect the likelihood of revisions, but a smaller maximum probability makes revisions larger and more likely.

7 Conclusion

Do revisions reveal mistakes? We find that indeed revised choices improve welfare according to all our normative benchmarks. Revealed preference analysis suggests further that

Table 6: Robustness of Budget Characteristics

	(1) Made Revision	(2) Made Revision	(3) Abs. Revision	(4) Abs. Revision
Max Prize		-0.00011		-0.0023
		(0.00012)		(0.0040)
Max Probability	-0.093***		-2.01*	
	(0.025)		(1.03)	
Subject FE	Yes	Yes	Yes	Yes
Max Prize FE	Yes	No	Yes	No
Max Probability FE	No	Yes	No	Yes
Observations	6516	6516	6516	6516

Notes: Linear regression clustered at the subject level. Each column represents a different regression, with the column head specifying the dependent variable. Significance indicated by: *** p<0.01, ** p<0.05, * p<0.1.

these revisions are closer to being generated by a strictly increasing utility function. Revised behavior is, therefore, more consistent with models that assume individuals have complete and transitive preferences over all alternatives. Thus, choices that are later revised are likely to be mistakes.

What lessons can we learn from detecting mistakes? One lesson is that mistakes are common, meaningful, and potentiality make it more challenging to observe preferences. Fortunately, adherence to how we believe individuals *ought* to behave improves with a simple prompt to revise. Future applications may use this method to distinguish between biases (preferences) and heuristics (mistakes). For example, present bias may be driven by a preference for the immediate or an inability to plan over a long horizon. A second lesson is that mistakes are made when the outcomes are unlikely and when the environment is unfamiliar. Choosing from sets with these characteristics may be more difficult. A third lesson is that reminders make revisions less likely, highlighting a potential tradeoff between the desire for consistency and choosing what one prefers in the moment. Whether demand effects, status quo bias, or memory is behind this discrepancy remains an open question.

Our results should not be read as a refutation of the core revealed preference hypothesis—that individuals have stable preferences. Mistakes are made, but identifying them is possible. Properly accounting for these inconsistencies improves the ability of utility functions to summarize observed behavior *as if* it is consistent with this hypothesis. Future applications can benefit from detecting and limiting these types of mistakes in order to draw more robust inferences about economic models.

References

Afriat, S. N. (1967). The construction of utility functions from expenditure data. *International economic review 8*(1), 67–77.

Afriat, S. N. (1973). On a system of inequalities in demand analysis: An extension of the classical method. *International economic review*, 460–472.

Agarwal, S., I. Ben-David, and V. Yao (2017). Systematic mistakes in the mortgage market and lack of financial sophistication. *Journal of Financial Economics 123*(1), 42–58.

Agranov, M. and P. Ortoleva (2017). Stochastic choice and preferences for randomization. *Journal of Political Economy 125*(1), 40–68.

Allcott, H. and D. Taubinsky (2015). Evaluating behaviorally motivated policy: Experimental evidence from the lightbulb market. *American Economic Review 105*(8), 2501–38.

Andreoni, J. and W. Harbaugh (2009). Unexpected utility: Five experimental tests of preferences for risk. *Unpublished Manuscript*.

Andreoni, J. and J. Miller (2002). Giving according to garp: An experimental test of the consistency of preferences for altruism. *Econometrica 70*(2), 737–753.

Benjamin, D., M. Fontana, and M. Kimball (2019). Reconsidering risk aversion. *Presentation: Economic Science Association North American Meeting*.

Benkert, J.-M. and N. Netzer (2018). Informational requirements of nudging. *Journal of Political Economy 126*(6), 2323–2355.

Bernheim, B. D. (2016). The good, the bad, and the ugly: a unified approach to behavioral welfare economics. *Journal of Benefit-Cost Analysis 7*(1), 12–68.

Bernheim, B. D., A. Fradkin, and I. Popov (2015). The welfare economics of default options in 401 (k) plans. *American Economic Review 105*(9), 2798–2837.

Bernheim, B. D. and A. Rangel (2009). Beyond revealed preference: choice-theoretic foundations for behavioral welfare economics. *The Quarterly Journal of Economics 124*(1), 51–104.

Bernheim, B. D. and D. Taubinsky (2018). Behavioral public economics. In *Handbook of Behavioral Economics: Applications and Foundations 1*, Volume 1, pp. 381–516. Elsevier.

Bhargava, S., G. Loewenstein, and J. Sydnor (2017). Choose to lose: Health plan choices from a menu with dominated option. *The Quarterly Journal of Economics 132*(3), 1319–1372.

Bronars, S. G. (1987). The power of nonparametric tests of preference maximization. *Econometrica: Journal of the Econometric Society*, 693–698.

Camerer, C. F. (1989). An experimental test of several generalized utility theories. *Journal of Risk and uncertainty 2*(1), 61–104.

Caplin, A., M. Dean, and D. Martin (2011). Search and satisficing. *American Economic Review 101*(7), 2899–2922.

Chen, D. L., M. Schonger, and C. Wickens (2016). otree: an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance 9*(1), 88–97.

Chetty, R., A. Looney, and K. Kroft (2009). Salience and taxation: Theory and evidence. *American economic review 99*(4), 1145–77.

Choi, J. J., D. Laibson, and B. C. Madrian (2011). $100 bills on the sidewalk: Suboptimal investment in 401 (k) plans. *Review of Economics and Statistics 93*(3), 748–763.

Choi, S., R. Fisman, D. Gale, and S. Kariv (2007). Consistency and heterogeneity of individual behavior under uncertainty. *American Economic Review 97*(5), 1921–1938.

Choi, S., S. Kariv, W. Müller, and D. Silverman (2014). Who is (more) rational? *American Economic Review 104*(6), 1518–50.

Feldman, P. and J. Rehbeck (2019). Revealing a preference for mixing: An experimental study of risk. *Unpublished Manuscript*.

Gathergood, J., N. Mahoney, N. Stewart, and J. Weber (2019). How do individuals repay their debt? the balance-matching heuristic. *American Economic Review 109*(3), 844–75.

Gaudeul, A. and P. Crosetto (2019). Fast then slow: A choice process explanation for the attraction effect.

Goldin, J. and D. Reck (2020). Revealed-preference analysis with framing effects. *Journal of Political Economy 128*(7), 2759–2795.

Halevy, Y., D. Persitz, and L. Zrill (2018). Parametric recoverability of preferences. *Journal of Political Economy 126*(4), 1558–1593.

Hey, J. D. (2001). Does repetition improve consistency? *Experimental economics 4*(1), 5–54.

Hey, J. D. and C. Orme (1994). Investigating generalizations of expected utility theory using experimental data. *Econometrica: Journal of the Econometric Society*, 1291–1326.

Hong, C. S., E. Karni, and Z. Safra (1987). Risk aversion in the theory of expected utility with rank dependent probabilities. *Journal of Economic theory 42*(2), 370–381.

Houtman, M. and J. Maks (1985). Determining all maximal data subsets consistent with revealed preference. *Kwantitatieve methoden 19*(1), 89–104.

Jacobson, S. and R. Petrie (2009). Learning from mistakes: What do inconsistent choices over risk tell us? *Journal of risk and uncertainty 38*(2), 143–158.

Kalaycı, K. and M. Serra-Garcia (2016). Complexity and biases. *Experimental Economics 19*(1), 31–50.

Koszegi, B. and M. Rabin (2008). Revealed mistakes and revealed preferences. In A. Caplin and A. Schotter (Eds.), *The Foundations of Positive and Normative Economics: A Handbook*, pp. 193–209. Oxford University Press.

Machina, M. J. (1989). Dynamic consistency and non-expected utility models of choice under uncertainty. *Journal of Economic Literature 27*(4), 1622–1668.

Martínez-Marquina, A., M. Niederle, and E. Vespa (2019). Failures in contingent reasoning: The role of uncertainty. *American Economic Review 109*(10), 3437–74.

Nielsen, K. and J. Rehbeck (2019). When choices are mistakes. *Available at SSRN 3481381*.

Nishimura, H., E. A. Ok, and J. K.-H. Quah (2017). A comprehensive approach to revealed preference theory. *American Economic Review 107*(4), 1239–63.

Polisson, M., J. K.-H. Quah, and L. Renou (2020). Revealed preferences over risk and uncertainty. *American Economic Review 110*(6), 1782–1820.

Ponce, A., E. Seira, and G. Zamarripa (2017). Borrowing on the wrong credit card? evidence from mexico. *American Economic Review 107*(4), 1335–61.

Quiggin, J. (1982). A theory of anticipated utility. *Journal of economic behavior & organization 3*(4), 323–343.

Rehbeck, J. (2020). How to compute the largest number of rationalizable choices. *Available at SSRN 3542493*.

Tversky, A. (1969). Intransitivity of preferences. *Psychological review 76*(1), 31.

Varian, H. R. (1982). The nonparametric approach to demand analysis. *Econometrica: Journal of the Econometric Society*, 945–973.

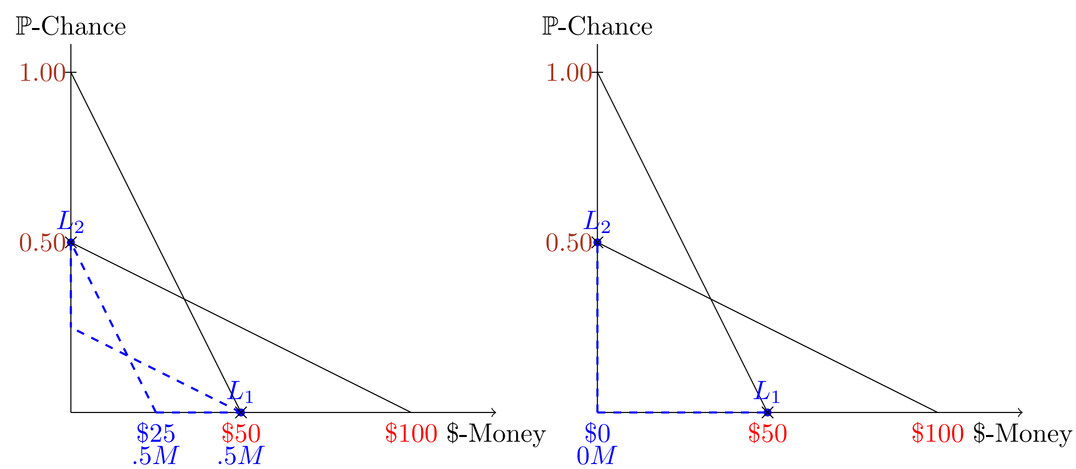
Yu, C. W., Y. J. Zhang, and S. X. Zuo (2019). Multiple switching and data quality in the multiple price list. *Review of Economics and Statistics*, 1–45.

A Revealed Preference Results

A.1 Probability Weighting and Expected Utility Index Computation

In this section, we show why our revealed preference tests are valid and how we compute them. Our results use the basic intuition from Polisson et al. (2020). However, our results are different because our choice environment is different. In particular, subjects in our experiment choose both p and x, which differs from the choice environment they consider, outcomes x_1 and x_2 with fixed likelihoods. In their environment, both consumption goods are measured in the same units, and both enter the Bernoulli utility function.

For the general existence of a utility function and its Afriat Index, we implement the tests as described in Nishimura et al. (2017). Although Afriat's theorem only assumes local non-satiation, our results extend trivially to first-order stochastic dominance in our environment. First, if all choices are at the interior of a budget, then our strictly revealed preference relation is the same as in Afriat's original theorem. In the case of a corner choice, the subject is encoded as having an Afriat Index of 0. Figure 11 illustrates the distinction between AIs using a WARP violation.



(a) Afriat Index

(b) Afriat Index under FOSD

Figure 11: Violations of WARP and their Afriat Indices

To simplify exposition we first describe the validity of the proofs in terms of two abstract commodities x_1 and x_2 and for some utility specification $U(x_1, x_2) = u_1(x_1) + u_2(x_2)$. All of the utility specifications we test are of the form $U(p, x) = \pi(p) * u(x)$, which is ordinally equivalent to $U(p, x) = log(\pi(p)) + log(u(x))$. Hence, $x_1 = p$, $x_2 = x$, $u_1 = log(\pi)$ and $u_2 = log(u)$. Our results are organized by decreasing generality: we first discuss probability weighting and then expected utility, which imposes further restrictions on π being an identity function.

Let the $\mathcal{X} \subseteq \mathbb{R}^2_+$ be the consumption space. Define the set of observation, choices and budgets, to be $\mathcal{O} = \{x^t, B^t\}_{t=1}^T$. Now define the downward closure of a budget set by $\underline{\mathrm{B}}^t = \{y \in \mathbb{R}^2_+ : y \leqslant x \text{ for some } x \in B^t\}$. Also let $x_i(x_j, B^t) : \mathbb{R}_+ \times \mathcal{X} \to \mathbb{R}_+$ be the i coordinate of element x_j on budget B^t or the origin if $(x_j, 0)$ is not in $\underline{\mathrm{B}}^t$.

Our generalized restriction of infinite domains (GRID) consists of

$$\left\{x^t \cup \bigcup_{s=1}^T \left(x_1^t, x_2(x_1^t, B^s)\right) \cup \bigcup_{s=1}^T \left(x_1(x_2^t, B^s), x_2^t\right)\right\}_{t=1}^T \cup \{0\}.$$

Theorem A.1. ***(Sufficiency of $\mathcal{G}$)*** *There exist a strictly increasing and continuous function that rationalizes $U(a,b) = u_1(a) + u_2(b)$ that rationalizes $\mathcal{O}$ on $\mathcal{X}$ if and only if there exists $\bar{U}(a,b) = \bar{u}_1(a) + \bar{u}_2(b)$ increasing function that rationalizes $\mathcal{O}$ on $\mathcal{X} \cap \mathcal{G}$.*

Proof. Clearly, if U rationalizes $\mathcal{O}$ on $\mathcal{X}$ and is strictly increasing, then it also rationalizes $\mathcal{O}$ on $\mathcal{X} \cap \mathcal{G}$.

For the converse, let be strictly increasing functions $\bar{u}_1$ and $\bar{u}_2$ that rationalize $\mathcal{O}$ on $\mathcal{X} \cap \mathcal{G}$. Suppose that x_1' and x_1'' are both numbers such that elements of the grid have these numbers as their first dimension, $x_1' < x_1''$ and no element of the grid has first dimension which is between these numbers. We define $\hat{u}_1$ as an extension of $\bar{u}_1$ such that for ε near zero,

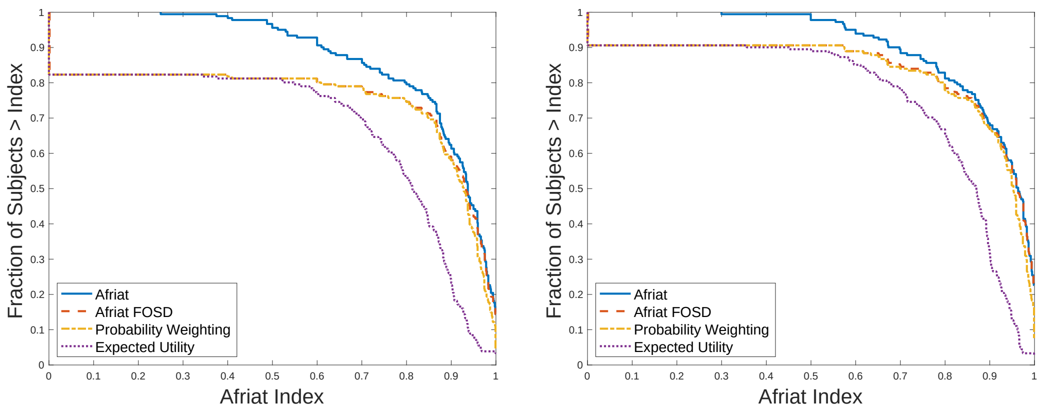
$$\hat{u}_1(x) = \begin{cases} \bar{u}(x_1') + \varepsilon(x - x_1') \text{ for } x \in [x_1', x_1'' - \varepsilon] \\ \bar{u}(x_1'') + \left(\frac{\bar{u}(x_1'') - \bar{u}(x_1') - \varepsilon(x_1'' - \varepsilon - x_1')}{\varepsilon}\right)(x - x_1'') \end{cases}.$$

$\hat{u}_1$ is a continuous and increasing piecewise linear extension of $\bar{u}_1$ which approaches the "step-function" extension of $\bar{u}_1$ as $\varepsilon \to 0$. Define $\hat{u}_2$ as a similar piecewise linear extension of $\bar{u}_2$, and $\hat{U}(x) = \hat{u}_1(x_1) + \hat{u}_2(x_2)$. For ε small enough, $\hat{U}$ rationalizes $\mathcal{O}$ on $\mathcal{X}$. To see this, note that at all points which are on the budget line but not in the grid, the marginal rate of substitution approaches either zero or infinity as $\varepsilon \to 0$. Thus, there will always be a point on the grid which is preferred to a point which is not on the grid. Since $\hat{U}$ extends $\bar{U}$ (which rationalizes $\mathcal{O}$ on the grid), $\hat{U}$ must rationalize $\mathcal{O}$ on $\mathcal{X}$. $\square$

We use two additional observations for the results given in the paper. First, it is straightforward to extend these results to budgets "scaled" by the index e. In this case, the generalized restriction of infinite domains (GRID) consists of

$$\left\{x^t \cup \bigcup_{s=1}^T \left(x_1^t, x_2(x_1^t, e \times B^s)\right) \cup \bigcup_{s=1}^T \left(x_1(x_2^t, e \times B^s), x_2^t\right)\right\}_{t=1}^T \cup \{0\}.$$

Figure 12: Combined Rationalizability Results



(a) Choices Indices

(b) Revisions Indices

Notes: These figures reproduce the results from Figures 4 and 5 to allow for comparison of the distributions of indices within data sets. By construction, each distribution must dominate the next.

Second, to test the expected utility model rather than the probability weighting model, it is sufficient to restrict $\pi(p) = p$ (and thus $u_1(p) = log(p)$).

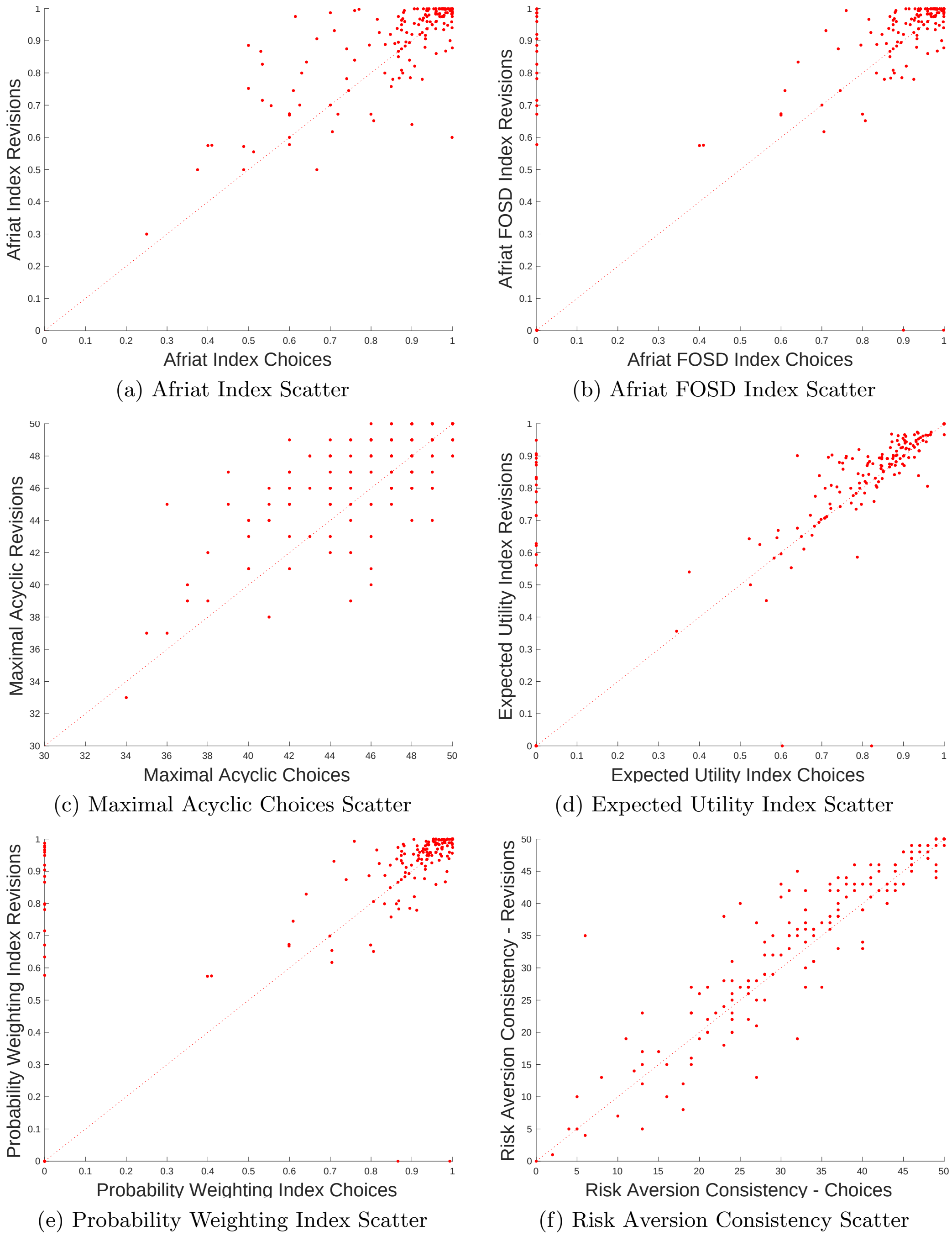
A.2 Additional Revealed Preference Results

To facilitate comparison with previous research, in this section, we report additional empirical results about our revealed preference measures.

Figure 12 organizes the results from Figures 4 and 5 by initial choices and revisions. Because the models being compared are nested, each index is dominated by the next. These results show that there are two primary model-based restrictions on the data that subjects violate. First, a sizable number of subjects violate FOSD by choosing at least one corner allocation, causing a difference between the distributions of Afriat and Afriat FOSD indices. Second, subjects' choices violate the assumptions of expected utility much more than they violate the assumptions of probability weighting, as evidenced by the differences between the yellow and purple lines.

Figure 13 disaggregates the results from Figures 4, 5, and 6 and instead reports the scores in scatter plots. Each observation in the scatter plot reports the indices of a single subject. These figures show that while revisions improve the overall distribution of indices, some subjects' indices fall while others increase.

Figure 13: Index Scatter Plots



(a) Afriat Index Scatter

(b) Afriat FOSD Index Scatter

(c) Maximal Acyclic Choices Scatter

(d) Expected Utility Index Scatter

(e) Probability Weighting Index Scatter

(f) Risk Aversion Consistency Scatter

Notes: These figures reproduce the results from Figures 4, 5, and 6 as scatter plots. Observations which are above the 45 degree line indicate an increase in a subject's index between the initial choice set and the revised choice set.

B Pre-Analysis Plan

Our pre-analysis plan was submitted to the AER RCT registry (AEARCTR-0004572).[14] The pre-analysis plan accurately reflects our experimental design, and our total number of subjects (181) was within the range of subjects we aimed to recruit (160-200).

The analysis included in Sections 6.1 and 6.3 are the main regression analyses that were discussed in the pre-analysis plan. The regressions in Table 3 correspond to columns 1 and 2 of Table 1 in the pre-analysis plan and the regressions in Table 5 correspond to columns 3 and 4 of Table 1 in the pre-analysis plan.

B.1 Analysis Omitted

The pre-analysis plan reported power calculations to identify a difference in the distributions of Afriat scores using a paired-sample t-test. These analyses were omitted in favor of non-parametric signed-rank tests. The null hypotheses of equality of means between initial choices and revisions for the Afriat Index, Afriat FOSD Index, HMI, Probability Weighting Index, and Expected Utility Index are all rejected with p-values of less than 0.001.

The pre-analysis plan specified that "we will parametrically estimate the one parameter CRRA model of risk preferences using both the initial and revised sets of decisions. With these two parametric estimates, we will compare the implied utility level (as a fraction of the maximum possible utility) of both the initial and revised decisions." This was omitted in favor of the non-parametric analysis in Section 5. We complete and report the analysis here. We assume that the von Neumann-Morgenstern utility function is $u(c;\rho) = \frac{1}{1-\rho}c^{1-\rho}$, so the decision maker solves

$$\max_{(x,p)} \frac{p}{1-\rho} x^{1-\rho}$$

subject to $x + \frac{M}{m}p = M$. The optimal prize choice is then $x^*(M,m) = \frac{1-\rho}{2-\rho}M$ Thus, we estimate the CRRA curvature parameter for each subject using nonlinear least squares on *budget shares*, solving the problem

$$\min_{\rho} \sum_i \left(\frac{x_i}{M_i} - \frac{1-\rho}{2-\rho} \right)^2 .$$

We complete this estimation exercise twice for each subject: once for the initial 50 choices giving $\hat{\rho}_C$ and once for the 50 choices in which choices are revised giving $\hat{\rho}_R$. We then calculate the *proportional utility improvement* for each budget which could be revised, which

[14] This can be downloaded at https://www.socialscienceregistry.org/versions/72424/docs/version/document.

is

$$\Delta u_i(\rho) = \frac{u(x_{i,R};\rho) - u(x_{i,C};\rho)}{u(x_i^*;\rho)},$$

where $x_{i,C}$ is the initial choice, $x_{i,R}$ is the revised choice, and x_i^* is the utility maximizing choice given ρ and the budget constraints. $\Delta u_i(\rho)$ can be thought of as the change in utility (as a fraction of maximal utility) the decision maker receives from revising their choice. If this value is positive, then revising the decision increases utility and if it is negative, revising the decision decreases utility.

The results can be found in Table 7. Column one focuses on $\Delta u_i(\hat{\rho}_C)$. The coefficient on the constant indicates that subjects gain roughly 1.5% of their maximal utility by revising their decisions. This welfare increase can be interpreted as a lower bound on the utility gains because parameters estimated from the initial choice set will tend to favor those initial choices. If we instead estimate the utility function based on the revised choice set, the estimates of utility gains are higher than 3%. Columns 2 and 4 of Table 7 differentiate the utility gains based on the *type* of revision it was.

Table 7: Treatment Effects on Utility

	(1) $\Delta u_{\mathrm{i}}(\hat{\rho}_{\mathrm{C}})$	(2) $\Delta u_{\mathrm{i}}(\hat{\rho}_{\mathrm{C}})$	(3) $\Delta u_{\mathrm{i}}(\hat{\rho}_{\mathrm{R}})$	(4) $\Delta u_{\mathrm{i}}(\hat{\rho}_{\mathrm{R}})$
Reminder		0.012		0.015*
		(0.0080)		(0.0080)
Double		0.0062		0.0094
		(0.0084)		(0.0086)
Reminder × Double		0.00088		-0.0031
		(0.011)		(0.011)
Constant	0.015***	0.0052	0.033***	0.020***
	(0.0052)	(0.0083)	(0.0039)	(0.0068)
Observations	6516	6516	6516	6516

Notes: Linear regression clustered at the subject level. Each column represents a different regression, with the column head specifying the dependent variable. Significance indicated by: *** p<0.01, ** p<0.05, * p<0.1.

C Experimental Instructions

The full set of instructions appears below.

Instructions:

PLEASE READ CAREFULLY AND DO NOT PRESS NEXT UNTIL INSTRUCTED TO DO SO.

This is a study about your own preferences. There are no right or wrong answers. This study has two parts. Part I has 50 tasks and Part II has 24 tasks. For Part I, you will see 25 different tasks first and then you will the same tasks again in a potentially different order. Once you have finished, we will pick a task at random as the **Decision-that-counts**. Since all decisions are equally likely to be chosen, you should approach each task as if it is the **Decision-that-counts**. Part II will be explained once you complete Part I.

For Part I your objective in each of the 50 tasks is to pick the **Choice** that you like the most. Every task has an interactive visual aid to assist with picking your preferred **Choice**.

In every task, you must choose among different **Choices**. Your **Choice** will determine a monetary prize and its chance. All your **Choices** will involve some chance of a monetary prize and $0. As you move the slider to the right, the monetary prize will decrease but the chance you will receive the prize will increase. For each task you will have to determine your preferred combination of a positive amount and its chance. The size of the potential outcome and how its chance changes with the slider will be different for different tasks.

The next few pages contains 3 examples to familiarize you with "How this works." The examples have outcomes that are different from the main tasks. Take your time and make sure you understand "How it works." We will not begin until everybody completes these examples and payments are explained. After the examples, there will be a detailed explanation of how payments will be determined.

Most **Choices** involve some risk. For example, a **Choice** could be a 25 in 100 chance of $30, and the corresponding 75 in 100 chance of $0. To aid with your choice, there will be a changing display for every possible **Choice**. Therefore, for any **Choice** you will always be able to see the chance of receiving a positive amount.

Important: You must move the slider around and then verify your answer next to it. If the chosen **Choice** on the slider does not match the **Verified Choice** next to it, or if you do not move the slider around, you will not be allowed to proceed to the next task. Once you have picked a **Choice** for a given task and verified it, you will no longer be able to change it.

Very Important: For each task, the slider is a tool to help you decide the choice you like the best. Therefore, it is in your best interest to move it around to help you determine which **Choice** you like better.

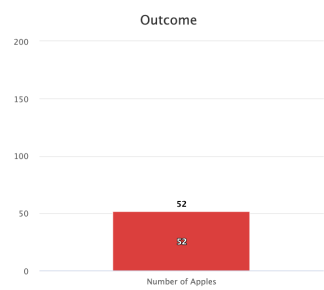
Next

Figure 14: General Instructions

Example 1

The maximum number of apples is 200.

As you move the slider to the right, the number of apples will decrease.



Choice ex1:

74

Verify ex1:

74

Please select your preferred **Chance** and **Outcome**. Note, this Task can be chosen as the **Decision-that-counts**.

Important: Remember you must manually select a spot on the slider for this task to count.

Next

Figure 15: First Example

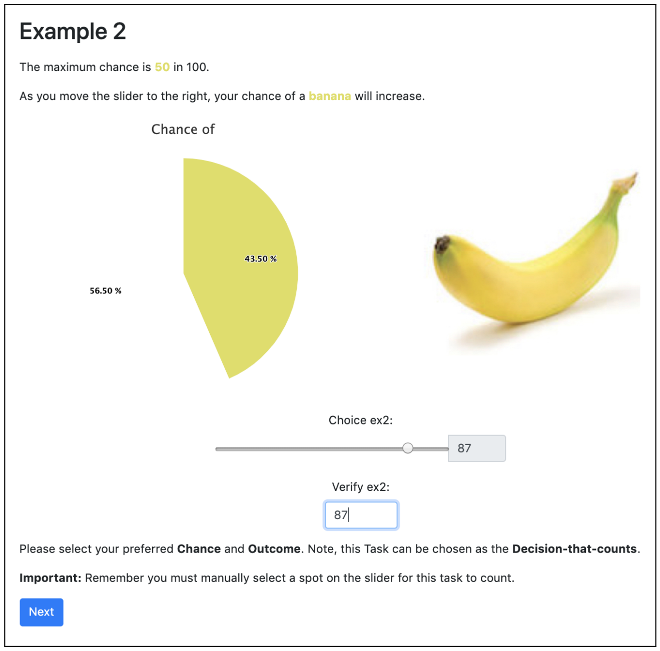


Figure 16: Second Example

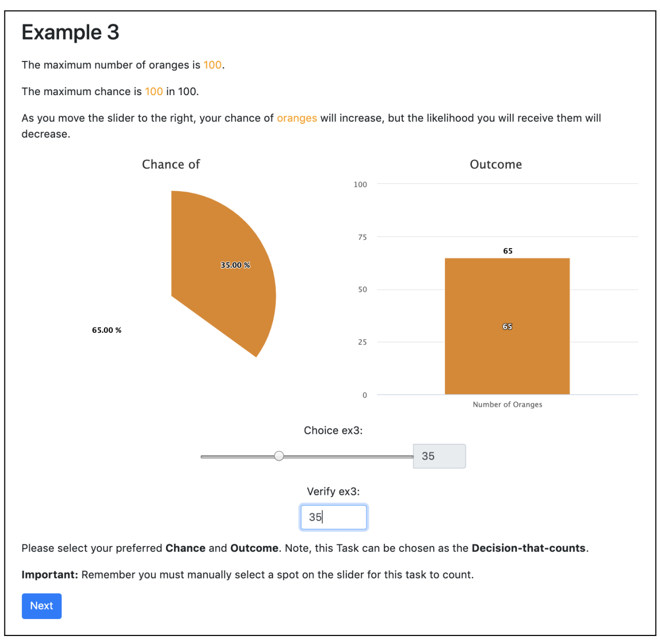


Figure 17: Third Example

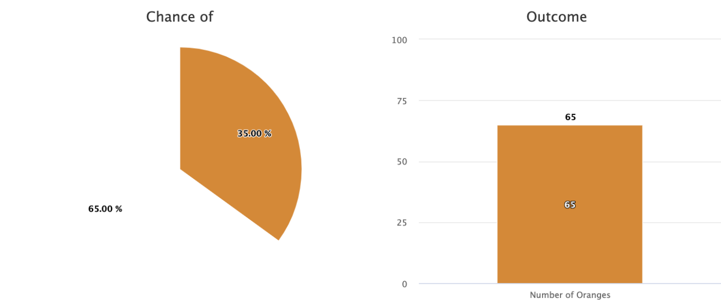
Earning Money:

PLEASE READ CAREFULLY AND DO NOT PRESS NEXT UNTIL INSTRUCTED TO DO SO.

After Parts I and II are completed, chance will determine your payment.

First, we will roll two ten-sided dice to determine the **Decision-that-counts**. One for the ten's digit and another for the one's digit. Any number up to 50, the number of tasks, will count. Therefore any ten's die that is more than five will have to be rerolled. A double zero will count as a hundred so this will also trigger a re-roll. In both cases, both dice will be rerolled.

Second, we will determine the chance of your payoff according to your selected **Choice**:



Therefore:

- You get 65.0 Oranges with a 35.0 in 100 chance.

- You get NOTHING with a 65.0 in 100 chance.

Next

Figure 18: Earnings

Things to Remember:

PLEASE READ CAREFULLY AND DO NOT PRESS NEXT UNTIL INSTRUCTED TO DO SO.

- You may adjust the size of your screen at any moment. Pressing CTRL and + together will zoom in and CTRL and - will zoom out.

- You will receive a $10.00 participation payment.

- You will complete 50 tasks for Part I. Part II has 24 tasks and will be explained after Part I.

- For Part I you will face the same 25 tasks twice in a row, in a potentially different order.

- Different **Choices** can have a different chances of a different prize. All you have to do is pick the **Choice** you like the best.

- There is no right or wrong answer for any of these questions.

- Once all of your decisions have been made, we will choose one task and one decision as the **Decision-that-counts** and will implement your preferred **Choice**.

- Every decision is equally likely to be the **Decision-that-counts**. So, it is in your interest to treat each **Choice** as if it could be the one that determines your payoffs.

- For each task, you must move the slider and verify your preferred **Choice**. Failure to move the slider or not matching it will prevent you from moving to the next task.

- Once you have selected your preferred **Choice** and verified it, you may not be able to change it.

- The slider is a tool to help you determine your preferred **Choice**. Therefore, it is in your best interest to use it to evaluate all potential alternatives.

Next

Figure 19: Reminders

Set of Unique Tasks:

PLEASE READ CAREFULLY AND DO NOT PRESS NEXT UNTIL INSTRUCTED TO DO SO.

- The set of tasks you will complete appears below.

- The column on the left gives the maximum amount of money available for each task

- The column on the right gives the maximum amount of chance available for each task

- Your **Choice** in each task will be some combination of this maximum outcome and chance

Maximum Chance	Maximum Outcome
8 in 100 chance	$140
8 in 100 chance	$160
8 in 100 chance	$180
8 in 100 chance	$200
24 in 100 chance	$120
24 in 100 chance	$140
24 in 100 chance	$160
24 in 100 chance	$180
24 in 100 chance	$200
50 in 100 chance	$80
50 in 100 chance	$100
50 in 100 chance	$120
50 in 100 chance	$140
50 in 100 chance	$160
50 in 100 chance	$180
50 in 100 chance	$200
76 in 100 chance	$80
76 in 100 chance	$100
76 in 100 chance	$120
76 in 100 chance	$140
76 in 100 chance	$160
92 in 100 chance	$80
92 in 100 chance	$100
92 in 100 chance	$120
92 in 100 chance	$140

Next

Figure 20: Full Set of Budgets

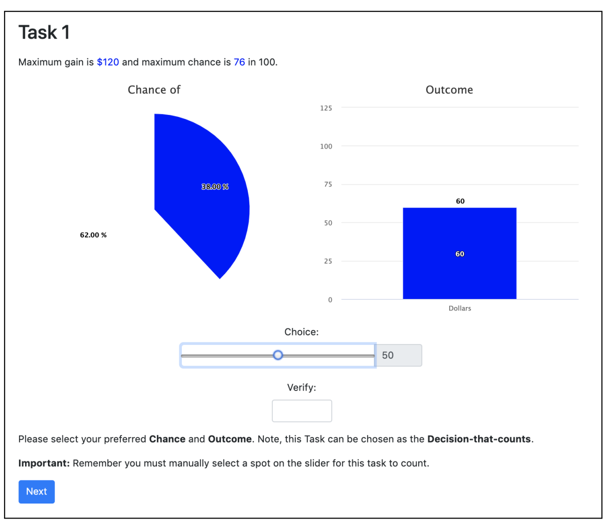
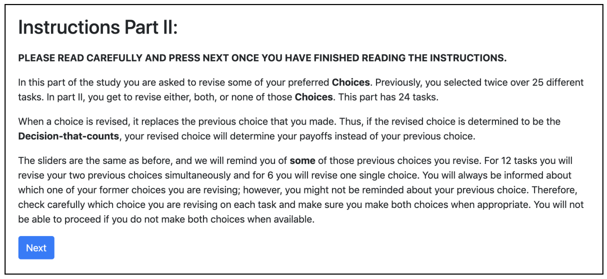


Figure 21: Sample Task


Instructions Part II:

PLEASE READ CAREFULLY AND PRESS NEXT ONCE YOU HAVE FINISHED READING THE INSTRUCTIONS.

In this part of the study you are asked to revise some of your preferred **Choices**. Previously, you selected twice over 25 different tasks. In part II, you get to revise either, both, or none of those **Choices**. This part has 24 tasks.

When a choice is revised, it replaces the previous choice that you made. Thus, if the revised choice is determined to be the **Decision-that-counts**, your revised choice will determine your payoffs instead of your previous choice.

The sliders are the same as before, and we will remind you of **some** of those previous choices you revise. For 12 tasks you will revise your two previous choices simultaneously and for 6 you will revise one single choice. You will always be informed about which one of your former choices you are revising; however, you might not be reminded about your previous choice. Therefore, check carefully which choice you are revising on each task and make sure you make both choices when appropriate. You will not be able to proceed if you do not make both choices when available.

Next

Figure 22: Instructions Part 2

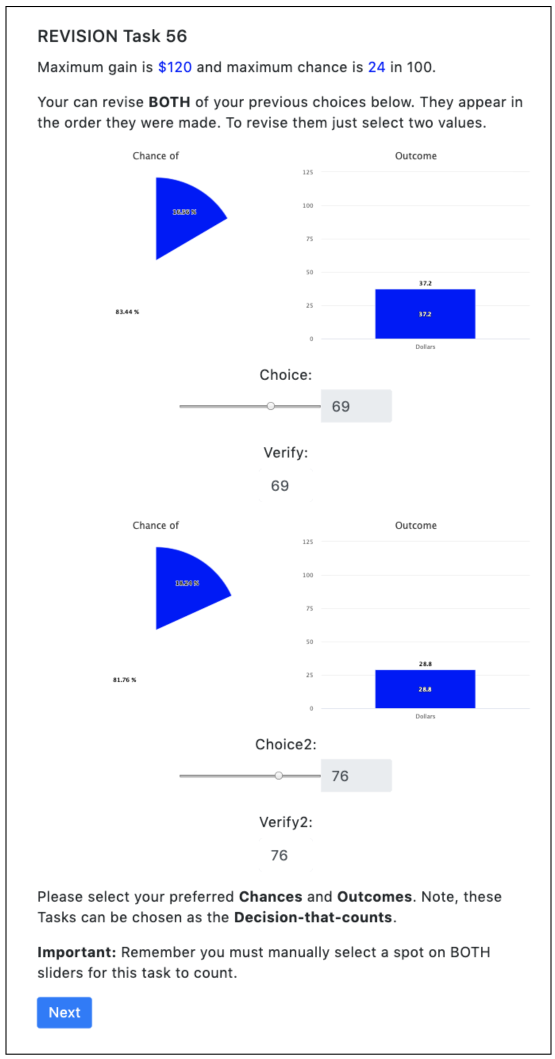


Figure 23: Revisions without Reminders

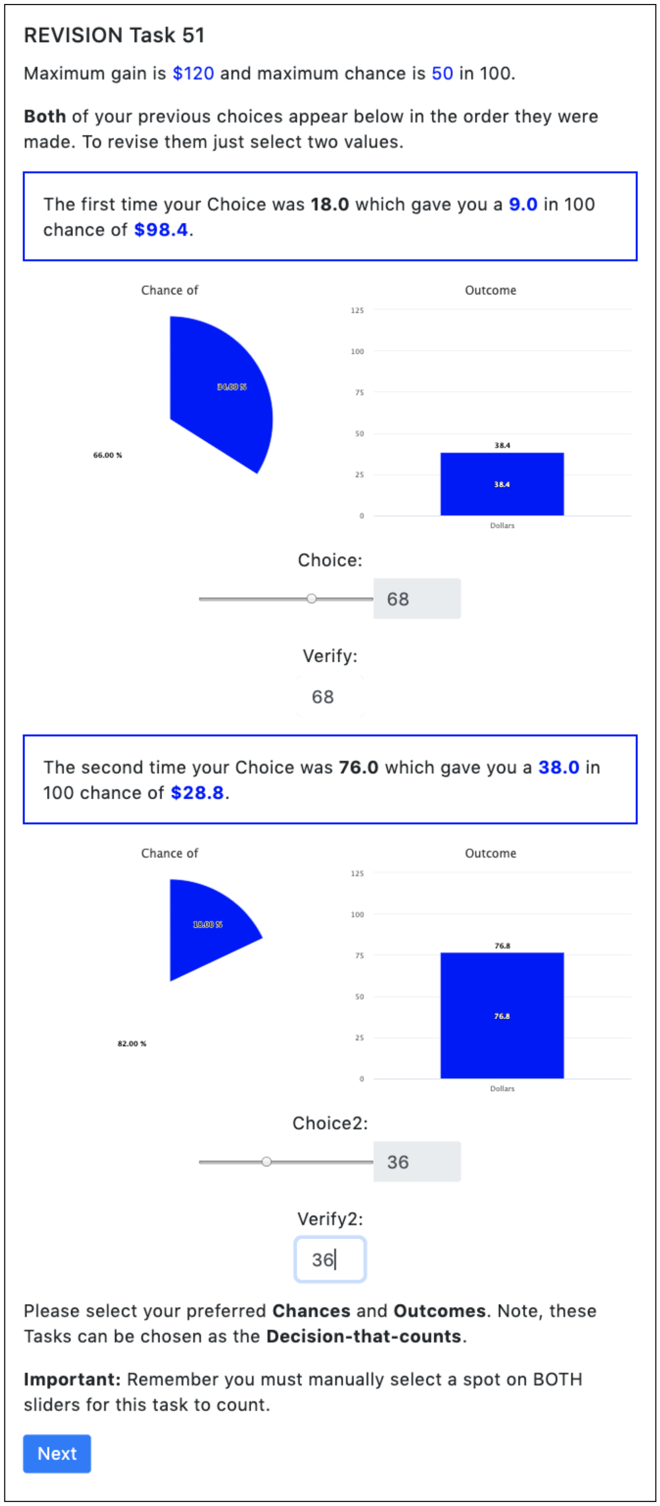


Figure 24: Revisions with Reminders

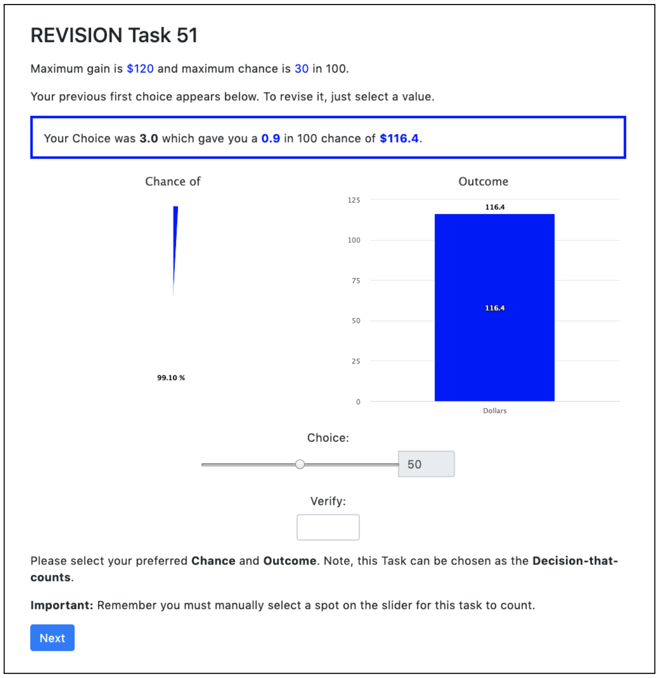


Figure 25: One Revision with Reminders

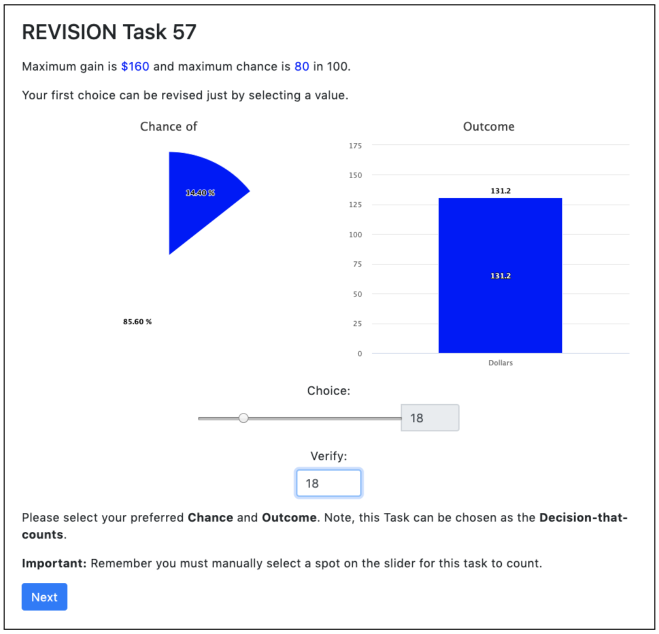


Figure 26: One Revision without Reminders