Varieties of Risk Elicitation *

by Daniel Friedman, Sameh Habib, Duncan James, and Brett Williams†

March 23, 2019

Abstract

We explore risk preference elicitation via direct choice over lotteries. Our choice tasks come in several varieties whose attributes vary incrementally, as the tasks range from choosing between two lotteries to selecting a portfolio from a continuous set of bundled Arrow securities. Each subject completes multiple instances of five different tasks, and responses are summarized in parametric (CRRA) and non-parametric (normalized risk premium) measures of risk preference. The distribution of elicited preferences varies widely across tasks, and within-subject correlation across some pairs of tasks is quite low. Observed variation in elicited preferences across tasks is explained in part by variation in design attributes, such as graphical vs text representation, that have no role in standard decision theory.

Keywords: Risk Aversion, Experiment, Elicitation, Multiple Price List
JEL Classifications: C91, D81, D89

*We are grateful for financial support from the National Science Foundation under grant SES-1357867, and for programming assistance from Matt Jee and also from Logan Collingwood, Emily Hockel and Joshua Pena. We are especially grateful to Sean Crockett for his input on early stages of this project. For helpful comments we thank Gabriele Camera, Dirk Engelmann, Paul Feldman, Mikhail Freer, Frank Heinemann, Eric Kimbrough, Jessie Li, Amma Panin, Stefan Trautmann (and students), Roel van Veldhuizen, Nat Wilcox, and seminar participants at WZB (April 2017), ESA (San Diego, May 2017), ESL Chapman (April 2018) and the Universities of Monash and New South Wales (February 2019). None of these organizations or individuals is responsible for any remaining errors or idiosyncrasies.

†Friedman and Williams: University of California, Santa Cruz, Santa Cruz, CA 95064, dan@ucsc.edu and bwillia4@ucsc.edu; James: Economics Department, Fordham University dujames@fordham.edu; Habib: The Joint Committee on Taxation, sameh.habib@jct.gov

1 Introduction

Over the last seven decades, economists have proposed a wide variety of methods to elicit individual subjects' risk preferences. The methods share the same ultimate scientific goal: to predict out-of-sample risky choice behavior at the population level and, if possible, at the individual level (e.g., Smith, 1989, Friedman, Isaac, James, and Sunder, 2014).

In this paper we pursue a key intermediate goal: to compare subject behavior across elicitation methods, and to seek regularities in terms of the methods' attributes. After all, if one can not predict choices in a different laboratory elicitation task, there is little hope of predicting risky choice behavior elsewhere.

Our focus is direct choice methods. That is, in every elicitation task we consider, individual subjects directly choose a particular lottery from some given feasible set of lotteries; there is no bidding, asking, or strategizing at any point in any procedure. We study versions of many of the best known elicitation tasks, including those introduced by Holt and Laury (2002), Eckel and Grossman (2002), Choi et al. (2007), and Gneezy and Potters (1997). To promote comparability and consistency across elicitation tasks, we reuse particular sets of the key decision-theoretic variables, price (of Arrow securities) and probability. The elicitation tasks, however, differ in their *attributes* — how they present these variables and how subjects enter their responses. For example, some tasks convey price and probability via text, while other tasks convey the same information via a spatial display such as a budget line. To better understand the impact of such design attributes, we vary them incrementally from one task to the next.

Section 2 below surveys some relevant literature and positions our contribution. Section 3 obtains theoretical predictions. Our point of departure, in keeping with much of the risk elicitation literature, is the expected utility hypothesis (EUH): a subject's risk preferences can be represented by a personal Bernoulli function whose expectation a subject always maximizes. We derive testable implications of the EUH, and show that a scalar variable which we call L is a sufficient statistic for the prices and probabilities when feasible lotteries lie on a budget line defined for two Arrow securities. We note that the coefficient γ of relative risk aversion is an index of a subject's risk preferences that permits comparisons across many tasks. We also consider some generalizations of the EUH and define a nonparametric alternative index, RRP, a normalized risk premium.

Section 4 lays out the design of our experiment. It presents our budget line (BL) screen display, as well as alternative non-spatial displays called budget jars (BJ and BJn) that offer precisely the same set of feasible lotteries for given Arrow prices and probabilities. It also shows how the Eckel-

Grossman task (EG below, from Eckel and Grossman (2002, 2008)) can be displayed spatially as a finite discrete subset of a budget line, and how each row from a multiple price list (HL, from Holt and Laury (2002)) can either be displayed as two available points in Arrow-Debreu 2-space or as text with two radio buttons. The section then presents the structure of the 56 trials administered to each subject. Each of our 142 subjects completes 5 of our 6 elicitation tasks with either prices or probabilities varying in each block of trials in either monotone or random order, in a manner balanced across subjects.

Section 5 collects the results. Three preliminary results are of interest in their own right and also set the stage for further analysis: dominated choices are rare; disappointment aversion adds little predictive power; and our variable L indeed does largely capture the impact of varying prices and probabilities, although there is a statistically significant asymmetry in an unexpected direction.

The section then presents five main results. First, the population distributions of elicited risk preferences, whether measured parametrically or non-parametrically, are substantially different across elicitation tasks. Second, at individual subject level, the preference measures γ and RRP are highly correlated with each other within each elicitation task, and each separate measure shows fairly high correlations for repetitions across the same task and across closely related tasks. However, other correlations are much lower, and are low to nil between BL tasks and HL tasks. Third, a simulation exercise suggests that it is difficult or impossible to explain the observed choices as noisy expression of stable underlying preferences. Our last two main results are that a substantial part of the variation in distributions and in correlations across tasks can be explained by task attributes, such as graphical vs text representation, that have no role in standard decision theory.

A concluding discussion summarizes and points to emergent opportunities for broadening theory and empirics. Appendix A proves a result on stochastic dominance relevant to our work, Appendix B reports some empirical robustness checks, and Appendix C is a copy of instructions to subjects.

2 Relevant Literature

A principal goal of our paper is to assess how risk preference estimates are influenced by parameters central to decision theory (such as prices and probabilities) and as well as by design attributes (such as the format for presenting the task) deemed irrelevant by standard decision theory. We now review prior work with that goal in mind.

Auction Bids. The research program initiated by Cox, Roberson, and Smith (1982) and Cox, Smith, and Walker (1988) uses subjects' bids in First Price Sealed Bid (FPSB) auction and

knowledge of their induced values to infer, through the lens of a particular bidding model, the subjects' risk aversion parameters. FPSB data analyzed in this manner tends to suggest a high degree of risk aversion.

Alternatively, one can make inferences about subjects' risk preferences from their behavior in a second price sealed bid auction (SPSB) for lotteries. In particular, the BDM procedure as in Becker, DeGroot, and Marschak (1964) consists of setting a reserve price for selling the lottery in a SPSB with automated truth-telling bidders. It is well known that this procedure tends to generate preference estimates on the risk-seeking side. However, as documented by Kachelmeier and Shehata (1992), a variant called buying-BDM tends to generate estimates on the risk-averse side. Dual-to-selling and dual-to-buying versions can also be constructed; James (2011) finds that dual-to-selling responses tend towards risk aversion, while dual-to-buying responses tend towards risk-seeking.

Direct Choice. Another class of elicitation methods directly asks the subject to choose one lottery from a menu. Leading examples include binary choice (Hey and Orme, 1994) and choice among a handful of lotteries (as in Binswanger, 1980, Eckel and Grossman, 2002, 2008). A more recent example is choice from a continuous budget line consisting of affordable combinations of Arrow securities, as in Choi, Fisman, Gale, and Kariv (2007), or Andreoni and Harbaugh (2009) or Andreoni, Kuhn, and Sprenger (2015). Although it does not explicitly display a budget line, the investment game of Gneezy and Potters (1997) also falls into this category, as noted in Section 4.1 below. Taken one row at a time, the well-known multiple price list of Holt and Laury (2002) is an instance of binary choice, but taken whole — all rows presented simultaneously and payment from a row randomly selected later — it belongs to a different class of tasks, as noted below.

Inconsistency across elicitation tasks. Lichtenstein and Slovic (1971, 1973) is an early comparison of risky choice behavior across different elicitation tasks. They compare (a) direct choice between two lotteries (i.e. binary choice) and (b) numerical valuation of those lotteries by means of selling-BDM, and find that typical subjects exhibit inconsistency and even self-contradiction across the two tasks. A large subsequent literature on "preference reversals" replicates and attempts to resolve those inconsistencies. A recent example, Collins and James (2015), demonstrates that replacing selling-BDM with dual-to-selling-BDM eliminates most reversals, and those that remain are consistent with the noisy response model of Blavatskyy (2014). Isaac and James (2000) provide another striking example of apparently inconsistent individual behavior across elicitation tasks. Subjects in their study who appear risk averse in the FPSB task often seem risk seeking in the selling-BDM task, and vice versa. Many subsequent studies have also found that the ordering of subjects' revealed risk preference is not preserved across tasks (e.g. Berg et al., 2005, Loomes and Pogrebna, 2014, Sprenger, 2015, Pedroni et al., 2017, Deck et al., 2013).

Two recent papers emphasize inconsistencies across elicitation tasks. Crosetto and Filippin (2016) find different population distributions for four different elicitation tasks, and use simulations to confirm that observed differences are not simply due to mechanical differences in the tasks. Since their individual subjects face only one task each, they can't examine inconsistencies or correlations at the individual subject level. Their concluding discussion conjectures that task attributes, such as discrete vs continuous of choice menus or the presence of a riskless choice, might account for the distributional differences. Zhou and Hey (2018) confront each subject with 2D visual representations (prize amount and probability) of four standard elicitation tasks, and find within-subject inconsistencies (e.g., low correlations across tasks in individual revealed risk preferences) as well as different distributions across tasks. They use constant absolute as well as constant relative risk averse functional forms to estimate individual risk preferences, but conclude that functional form matters far less than the choice of elicitation task.

Equivalent Procedures. One can establish that the Holt and Laury (2002) procedure with one row selected randomly for payment is equivalent to the dual-to-selling version of BDM (James, 2011), which is in turn related to Grether (1981), applied to comparison between two lotteries, each with unknown probabilities (as in Crockett and Crockett, 2018). This equivalence was anticipated, though not specifically identified, by Freeman, Halevy, and Kneeland (2016), who pointed out that paying one row of Holt-Laury, randomly selected, rendered Holt-Laury subject to the same non-EUT critique applied to selling-BDM by Karni and Safra (1987). In other words, Holt-Laury implemented with the pay-one-randomly protocol (Cox et al., 2015) is a particular known variant of BDM.

Recently popular balloon (Lejuez et al., 2002) or bomb (Crossetto and Filippin, 2013) elicitation tasks are equivalent to an $N = 2$ FPSB auction. That is, the task of avoiding a zero-payoff absorbing state while being paid in a linearly increasing manner for exploration has the same payoff function as a player in an $N = 2$ FPSB whose opponent submits a bid randomly drawn from the uniform density on $[V_L, V_U]$. (The Cox, Smith, and Walker (1988) robot bidders did just that, for $N > 2$.)

Supposed Irrelevancies. Standard decision theory, for example using the formalism of mechanism design, can be used to classify the different elicitation tasks (e.g., Hurwicz, 1972, Smith, 1976). For example, FPSB has different cost rules from SPSB (and thus BDM), and so induces different strategies. (Of course, given risk preferences are recoverable from either task.) Conversely, two institutions with identical cost and allocation rules may have different practical implementations. For example, FPSB and the Dutch auction have identical cost and allocation rules, but differ in transition rules; empirically they generate different behavior (Cox et al., 1982) despite predicted revenue equivalence.

Yet more subtle differences in implementation may affect behavior, despite theoretical predictions to the contrary. The response mode or the manner in which information is presented may affect subjects' behavior. Indeed, Habib et al. (2017) find that spatial representation of payoff and probability information in the Holt-Laury task supports different (less risk averse) behavior than text representation of the same information; this is despite algebraic and response mode (direct choice via radio button) equivalence across the two versions of the task.

Even the sequencing of task instances might change behavior. Lévy-Garboua et al. (2012) find that the standard Holt-Laury task with rows presented in a monotone order (i.e., $P_h = 0.1$ is followed by $P_h = 0.2, 0.3, ...$) elicits behavior closer to risk neutrality than when the same rows are presented in a random sequence.

Positioning the present paper. Our experiment will revisit these and other effects deemed irrelevant by standard decision theory, and embed them in a unified design. This enables the sharpest assessment to date of the impact of such decision-theoretic irrelevancies and of the connections between them. More specifically, we offer an initial assessment of how a set of design attributes affects the distributions and correlations of revealed risk preferences across elicitation tasks.

There is a vast literature on testing the implications of particular generalizations of the expected utility hypothesis, and another vast literature on searching for some "best" elicitation method. Our goals in the present paper are quite different from those in either literature. We seek instead (a) to document regularities in how elicited preferences depend on elicitation method, and (b) to connect those regularities to the methods' design attributes. We restrict attention to the simplest possible cost and allocation rule — direct choice from a given set of lotteries — since this rule may give expected utility theory its best shot; see, for example, Harbaugh, Krause, and Vesterlund (2010) and Trautmann and van de Kuilen (2012). As explained in Section 4 below, we connect a broad set of direct choice procedures by incrementally varying the procedures' key attributes.

3 Theoretical Predictions

In all risk preference elicitation tasks that we consider, a subject chooses an allocation (x, y) from a compact feasible set F of Arrow securities. That is, we assume two mutually exclusive possible states, X and Y, of known probabilities $\pi_X > 0$ and $\pi_Y > 0$ with $\pi_X + \pi_Y = 1$; a chosen allocation (x, y) pays x points in state X and y points in state Y. According to standard decision theory, only these opportunities and the subjects' preferences matter; the manner in which F is presented to a decision maker and decisions are recorded are irrelevant, so long as they are clear and unambiguous.

The Expected Utility Hypothesis (EUH) is a leading special case of standard decision theory. It posits that each human subject has her own fixed preferences representable by a Bernoulli function, i.e., a smooth (twice differentiable) and strictly increasing function $u : \mathbb{R} \to \mathbb{R}$, defined up to a positive affine transformation. The EUH states that the subject's choice (x^*, y^*) solves

$$\max_{(x,y)\in F} \pi_X u(x) + \pi_Y u(y). \tag{1}$$

From the EUH perspective, the art of elicitation is for the experimenter to choose a sequence of feasible sets F so that subjects' choices reveal key aspects of their Bernoulli functions u. For some elicitation tasks, the feasible set is a standard budget set: non-negative bundles that are affordable. Since $u' > 0$, there is then no further loss of generality in replacing F by the budget constraint

$$p_x x + p_y y = m, \tag{2}$$

where m is an (implicit or explicit) endowment of cash, and $p_x > 0$ and $p_y > 0$ are the prices of the two Arrow securities. In all elicitation tasks that we study, F is a subset (sometimes a finite subset) of points satisfying (2). We normalize prices so that $p_x + p_y = 1$; this jibes with the convention that a unit of cash is the portfolio $(x, y) = (1, 1)$.

The first order conditions for optimization problem (1)-(2) can be written out in terms of the Lagrange multiplier λ for (2) as

$$\lambda = \frac{\pi_Y}{p_y} u'(y) = \frac{\pi_X}{p_x} u'(x) \tag{3}$$

or as

$$MRS = \frac{u'(x)}{u'(y)} = \frac{\pi_Y}{\pi_X} \frac{p_x}{p_y} \tag{4}$$

or as

$$\ln \frac{u'(x)}{u'(y)} = -[\ln \pi_X - \ln \pi_Y - \ln p_x + \ln p_y] \equiv -L. \tag{5}$$

Thus, for whichever Bernoulli function u a subject may have, the EUH implies that

1. An interior choice (x, y) is determined by ratios of state prices and probabilities.

2. The composite variable $L = \ln \pi_X - \ln \pi_Y - \ln p_x + \ln p_y$ is a sufficient statistic for prices and probabilities. Equation (5) holds at interior solutions, and corner solutions are also defined by L: corner $(\frac{m}{p_x}, 0)$ is chosen if $\ln \frac{u'(\frac{m}{p_x})}{u'(0)} \geq -L$, while corner $(0, \frac{m}{p_y})$ is chosen if $\ln \frac{u'(0)}{u'(\frac{m}{p_y})} \leq -L$.

3. When regressing log marginal rate of substitution on log price ratio and log odds, the coefficients should be equal in magnitude with opposite signs.

We will soon see that some popular generalizations of EUH obey similar rules. But first we note that we can say more in important special cases.

3.1 Special cases

For a **risk neutral** agent we have $u'(x) = u'(y) = constant$, and (3) becomes

$$\frac{\pi_Y}{p_y} = \frac{\pi_X}{p_x}. \tag{6}$$

Equation (6) can only be satisfied if $L = 0$. Otherwise we'll get a corner solution — the risk neutral person spends her entire budget on the asset with higher probability/ price ratio, so $x^* = 0$ if $L < 0$ and $y^* = 0$ if $L > 0$.

CRRA, a widely used parametric family of Bernoulli functions, sets $u(c|\gamma) = \frac{c^{1-\gamma}}{1-\gamma}$ where the parameter $\gamma \geq 0$ is the coefficient of relative risk aversion. (For $\gamma = 1$ the Bernoulli function is $\ln c$, as can be seen using L'Hospital's rule.) For this family $u'(c) = c^{-\gamma}$ and MRS $= [\frac{x}{y}]^{-\gamma}$. Inserting this into (4) and taking logs yields

$$\ln \frac{x}{y} = \frac{-1}{\gamma}[\ln \pi_Y - \ln \pi_X - \ln p_y + \ln p_x] = \frac{1}{\gamma} L. \tag{7}$$

That is, regressing log-odds of the chosen allocation on L will directly reveal (as the inverse slope) the subject's coefficient γ of relative risk aversion. Moreover, as separate regressors, all four components of L (log prices and log probabilities) should have exactly the same coefficients, $\pm\gamma^{-1}$. Note that even if the Bernoulli function is not in the CRRA family, the reciprocal of the estimated slope coefficient still can serve as measure of revealed risk aversion, which (at noted following equation (5) above) is equally responsive to log prices and log probabilities.

3.2 Generalizations of EUT

Gul (1991) presents a model with a free parameter $\beta \geq 0$ intended to capture disappointment aversion as a probability distortion in a two state world — people make choices as if maximizing expected utility that assigns extra weight (by a factor of $1+\beta$) to the probability of the less favorable state. In our notation, the unnumbered equations near the top of Gul (1991, p. 678) say that the indifference curve segments have slope

$$-\frac{dy}{dx} = B\frac{\pi_X}{\pi_Y}\frac{u'(x)}{u'(y)} > 0, \tag{8}$$

where $B = (1+\beta)$ if $x < y$ (so X is the less favorable state) and $B = (1+\beta)^{-1}$ if $x > y$ (so Y is less favorable). Thus the indifference curve has a kink on the diagonal $x = y$, with -slope $= \frac{\pi_X}{\pi_Y}(1+\beta)$ on the right and -slope $= \frac{\pi_X}{\pi_Y}(1+\beta)^{-1}$ on the left.

Suppose, as commonly assumed in the subsequent literature, that the underlying Bernoulli function u is CRRA with risk aversion coefficient $\gamma > 0$. The tangency condition $-\frac{dy}{dx} = \frac{p_x}{p_y}$ applies as usual when the optimal choice is interior (not on the diagonal nor at a corner of the budget set). Writing $b = \ln(1+\beta)$, and recalling that in this case $\frac{u'(x)}{u'(y)} = (\frac{x}{y})^{-\gamma}$, we see that disappointment aversion changes equation (7) to

$$\ln \frac{x}{y} = \frac{1}{\gamma}[L - b] \tag{9}$$

when $L > b$ and so $x > y$. By symmetry, when $L < -b$, we again have a tangency but with $x < y$ and with $-b$ replaced by $+b$ in (9). For values of $L \in [-b, b]$, a DA agent will choose at the kink in the indifference curve where the diagonal $x = y$ intersects the budget line. Of course, for extreme values of L and sufficiently small parameters b and γ, corner solutions (on the x or y axis) are also possible. Estimating equations in Appendix A spell this out explicitly.

An older generalization of expected utility theory allows for diminishing sensitivity to probability.[1] Diminishing sensitivity is commonly represented (e.g., Kahneman and Tversky, 1979, Tversky and Kahneman, 1992, Camerer and Ho, 1994) with an "inverse-S shaped" probability weighting function $w(\pi_X)$. That is, individuals are said to act as if maximizing $w(\pi_X)u(x) + w(\pi_Y)u(y)$, where over the middle range of probabilities $\pi_X, \pi_Y \in [0.3, 0.8]$ that we consider in our experiment, $w(\pi) > \pi$ for $\pi = \pi_X$ or π_Y below a crossover point of approximately 0.5, and $w(\pi) < \pi$ for π above the crossover point. Thus in our experiment the log weighted odds ratio $[\ln w(\pi_X) - \ln w(\pi_Y)]$ is smaller in absolute value than the unweighted log odds ratio $[\ln \pi_X - \ln \pi_Y]$ over the relevant range. Consequently diminishing sensitivity modifies EUH implication 3 above to state that the coefficient on objective log odds will have smaller magnitude than the coefficient on log price ratio.

3.3 Non-parametric summary statistic

We have seen that we can recover an estimate of a subject's coefficient of relative risk aversion from her responses to a budget line elicitation task. For non-CRRA preferences (and even for heuristics that are not consistent with a preference relation), the estimated γ can still be regarded as a way to summarize a subject's risk attitude, but it no longer has a precise interpretation. A researcher may

[1] Prospect theory (e.g., Kahneman and Tversky, 1979) is the best known instance. Besides capturing diminishing sensitivity via a probability weighting function, Prospect Theory also uses a Value function V, whose shape may change on either side of a reference point. A natural reference point in our context is zero. In that case, V is equivalent to a Bernoulli function.

prefer some sort of nonparametric summary statistic, but we are not aware of any such statistic that is defined and comparable across all of our elicitation tasks. We considered several possibilities[2] and eventually settled on a normalized risk premium, defined as follows.

Let $M = \max_{(x,y)\in F} \pi_X x + \pi_Y y$ be the maximum feasible expected payoff in an elicitation task. When $L \neq 0$, there is a unique point (x_M, y_M) that achieves that maximum and would be selected by a risk neutral agent. As usual, define $\mu_M = \pi_X x_M + \pi_Y y_M$ and $\sigma_M^2 = \pi_X(x_M - \mu_M)^2 + \pi_Y(y_M - \mu_M)^2$; note that $\sigma_M > 0$ in all our elicitation tasks. Let $C = \pi_X x_C + \pi_Y y_C$ be the expected payoff of the subject's actual choice $(x_C, y_C) \in F$. Then the revealed Relative Risk Premium is

$$RRP = \frac{M - C}{\sigma_M}. \tag{10}$$

if $L \neq 0$ and otherwise is 0. Thus RRP resembles a coefficient of variation or a Sharpe ratio, and captures the agent's willingness to forego expected payoff in order to reduce dispersion.

3.4 Stochastic Dominance

Suppose that $\pi_X = \pi_Y = 0.5$ and $p_x = 0.4$ while $p_y = 0.6$. No matter what her risk preferences, an agent facing these prices and probabilities should never choose a point on the budget line with $x < y$. For example, suppose she considered choosing $(x, y) = (7.5, 15)$, exhausting her budget $m = 12$. Since the states are equally likely, she'd be just as happy with $(15, 7.5)$, no matter what her Bernoulli function is. But the portfolio $(15, 7.5)$ costs only 10.5, so she could afford to spend 1.5 more on either Arrow security and be strictly better off than at $(x, y) = (7.5, 15)$.

The general result is expressed in terms of first order stochastic dominance (FOSD). Recall that lottery A (strictly) FOSDs lottery B iff $F_A(x) \leq F_B(x)$ for all x, with strict inequality for some x. The definition refers to the cumulative distribution function $F_Z(x)$, the probability that the realized payoff in lottery Z is no greater than x. Recall also (e.g., Mas-Colell, Whinston, and Green (1995), p. 195) that every expected utility maximizing agent prefers lottery A to B iff A FOSDs B.

Proposition 1 *A choice (x, y) on the budget line (2) is strictly first order stochastically dominated by another choice on the same budget line iff*

a. one Arrow state (e.g., X) is more likely and its security is less expensive (e.g., $\pi_X \geq \pi_Y$ and $p_x \leq p_y$), with at least one of these comparisons strict; and

[2] E.g., Heufer (2014) offers a non-parametric indicator derived from revealed preference considerations. It is not immediately obvious how to modify Heufer's approach to deal with varying probabilities as in half of our sessions. Also, since we vary prices while holding constant the x-endowment, our design is not conducive to revealed preference analysis, which relies on budget lines that cross each other.

b. the choice includes strictly less of the less-expensive-more-likely security (e.g., $x < y$).

See Appendix A for a proof, which can be generalized in a straightforward manner to cover Prospect Theory with symmetric probability weighting as well as Disappointment Aversion and some other generalizations of expected utility theory.

The Proposition tells us that every choice on the budget line can be rationalized by some Bernoulli function if the more likely state has a higher price, or if $L = 0$. But some choices will be dominated when prices are equal and probabilities differ, or the reverse, and when the more likely state has a lower price. In those cases, we can test for the rationality of subjects without committing to a functional form. For example, in Figure 1 below, the budget line crosses the diagonal at (400, 400)/9; any choice on the budget line with $x > 400/9$ is strictly dominated by an interval of choices with $x < 400/9$.

4 Laboratory Procedures

There are many ways to display information and to collect subjects' responses even in direct choice elicitation tasks. In order to facilitate eventual analysis of how such attributes affect elicited preferences, we implement standard tasks so that adjacent tasks differ by only a minimal change in attributes. Section 4.1 lays out our implementations task-by-task, while Sections 4.2 and 4.3 offer details of parameterization and session design.

4.1 Elicitation Tasks

Budget line (BL). One task is to choose a lottery from a simple budget line in the tradition of Choi et al. (2007), as in Figure 1. The tentatively chosen lottery, a portfolio of Arrow securities, appears as a large dot, with coordinates (state-contingent payments) shown in a text box. State probabilities are shown in text, while the state prices and the cash endowment are implicit in the slope and intercepts of the displayed budget line. In different trials, the price ratio varies from 0.23 to 1.23, and the the X state probability varies from 0.3 to 0.8.

Budget Jars (BJ and BJn). Figure 2 shows the user interface for an alternative way to present the same feasible set as in BL. We refer to the task as BJn (Budget Jars with no cash retention). Subjects start with an explicit cash endowment (shown in green in the wide jar) and use sliders on the other two jars to buy the two Arrow securities. The level in the cash jar decreases (resp. increases) as the subject drags up (resp. down) the level in the red (security X) or blue (security Y) jar, at a rate proportional to the price of that security. The text below the jars spells out the

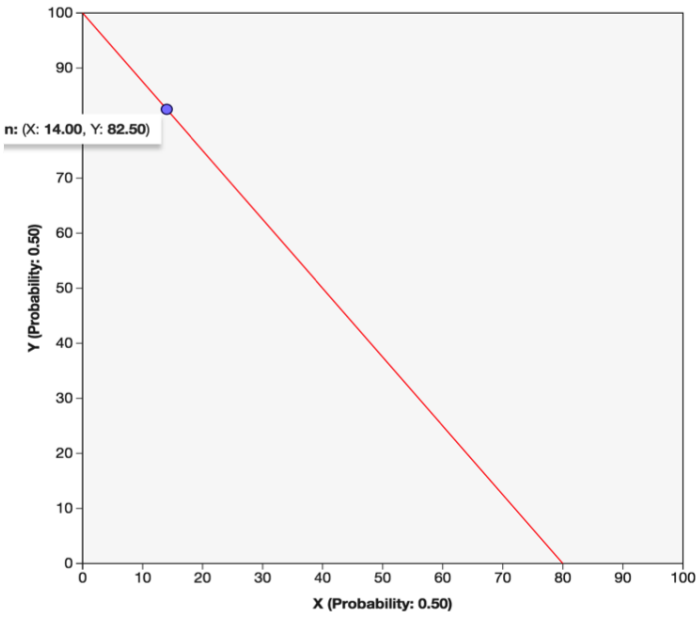

Figure 1: In treatment BL, the subject chooses an portfolio of Arrow securities by clicking any point on a given budget line, then clicking Confirm bar (not shown). Text box shows values (x, y) at clicked point, here $(14, 82.5)$. Axis labels note π_X and π_Y; here, each is 0.5.

state contingent payoffs (and state probabilities) at the current allocation. The subject clicks the Submit bar to finalize the current allocation. The Submit bar is grayed out (not clickable) until the cash jar is empty in treatment BJn.

One advantage of BJn, not exploited here, is that it can easily accommodate three or more Arrow securities. With two securities, the final allocations clearly map 1:1 onto the entire budget line, but that spatial representation is not present in the BJn task.

The task BJ is the same as BJn except that cash retention is allowed. Feasible levels of cash and security jars again map onto the entire budget line (though no longer 1:1). Proposition 1 implies that all choices consistent with FOSD can be achieved without purchasing any of a strictly more expensive security when state probabilities are equal. Suppressing the more expensive jar to prevent such dominated choices would render an instance of the BJ task equivalent (according to standard decision theory) to the Investment Game of Gneezy and Potters (1997); their parameters correspond to endowed cash of 4.0, x-intercept of 10, and price ratio $4/(10-4) \approx 0.667$. The comparison between BJ and BJn isolates the impact of allowing cash retention, while the comparison between BL and either BJ isolates the impact of text vs. spatial representation of lotteries.

Budget Dots: Eckel and Grossman (BDEG). Eckel and Grossman (2002, 2008) ask subjects

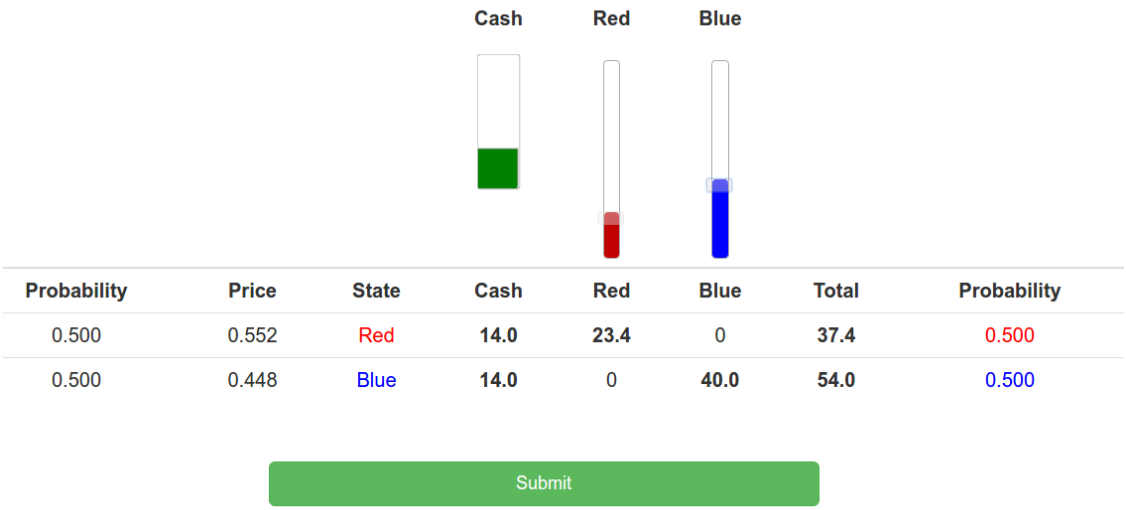

Figure 2: In treatment BJ, subjects choose an affordable allocation (x, y) by moving the sliders on the red and blue jars. The text below automatically updates so that x is shown in the "Total" column in the Red row, and y is shown below it in the Blue row. Clicking the Submit bar finalizes the allocation. In treatment BJn, the subject must empty the cash jar before the Submit bar becomes active.

to chose a single lottery (x, y) from a menu F of five or six alternatives, with equal state probabilities $\pi_X = \pi_Y = 0.5$. We modify their task by displaying the menu F as discrete set of points in a graph otherwise similar to Figure 1. Their menus typically graph as in Figure 3a: equally spaced points on a budget line spanned by the intercept for the cheaper security and the perfectly hedged portfolio $(x = y)$; their menus typically exclude stochastically dominated points (e.g. Crosetto and Filippin (2016)). Some of our BDEG trials, unlike the original, use unequal state probabilities. Holding constant prices and probabilities, comparing BDEG and BL choices isolates the impact of taking an evenly spaced discrete subset of the budget line.

Multiple price list (HL, BDHL). Perhaps the most widely used elicitation task in recent years is the multiple price list in text format (e.g., Holt and Laury, 2002). Each row in the list has the same two allocations but different rows have different state probabilities. Our HL treatments use Holt & Laury's original pair of allocations — the "safe" lottery $(x, y) = (2.00, 1.60)$ and the "risky" lottery $(x, y) = (3.85, 0.10)$. To streamline our design, we include only the six most relevant state probabilities, $\pi_X = 0.3, 0.4, 0.5, 0.6, 0.7, 0.8$.[3] Treatment HL stacks six rows of text, each

[3]The original list also included $\pi_X = 0.1, 0.2, 0.9, 1.0$ but 97% of subjects in the relevant treatment ("low real stakes," Holt and Laury (2002)) chose the safe lottery for $\pi_X = 0.1, 0.2$ and chose the risky lottery for $0.9, 1.0$. See Habib et al. (2017) for insight into the likely impact of dropping those lines from the list.

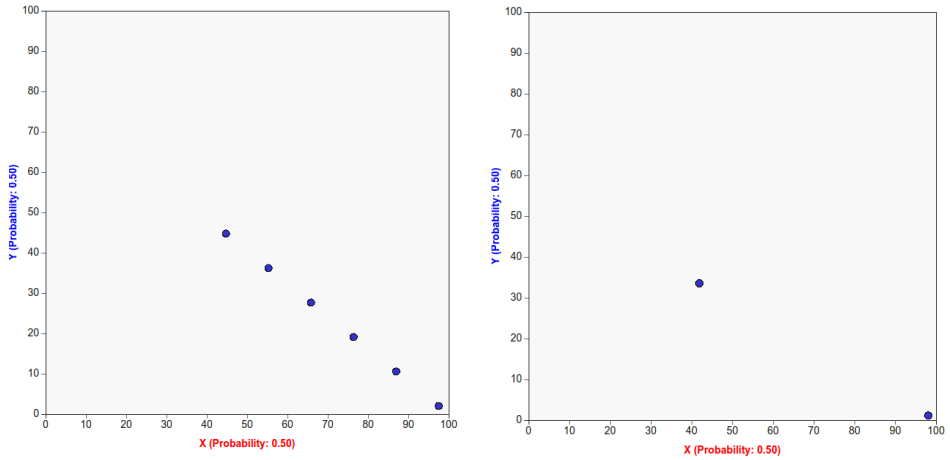

Figure 3: Discrete budget dots. Axis labels note π_X and π_Y; here, each is 0.5. (a) In treatment BDEG, the subject chooses an allocation of Arrow securities by clicking one of the six large dots on the given budget line, then clicking Confirm bar. (b) In treatment BDHL, subjects click one of two dots representing the two feasible HL allocations.

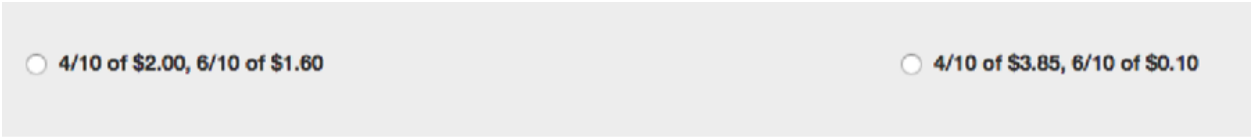

Figure 4: In treatment HL, subjects click a radio button to choose between two feasible allocations in each line; the X state probability (here 0.40) increases by 0.10 from one line to the next.

row representing choice between two lotteries as in Figure 4, with $\pi_X = 0.3$ in the top row and increasing by 0.1 in each successive row. Treatment BDHL takes the lotteries from one row (i.e., with a particular π_X value) from HL and displays the two feasible choices graphically, as in Figure 3(b), where the implicit price ratio is $p = -\frac{\Delta y}{\Delta x} = \frac{1.60-0.10}{3.85-2.00} \approx 0.81$. As further described below, successive trials vary the probabilities while keeping the price constant, and some sets of trials use an implicit price of .58 instead of the original 0.81. The comparison between HL and BDHL isolates (again) the impact of text vs. spatial representation of lotteries.

4.2 Experimental Design

Each subject faces a total of 66 lottery choices, organized into 56 trials in 11 blocks. The first and last block for each subject is the six-row (hence 6 lottery choices) HL elicitation task (with $p = \frac{p_x}{p_y} = 0.81$ implicitly) considered as a single trial. The other nine blocks consist of six consecutive trials, each with a single lottery choice. The middle block (block 6 of 11) for each subject is always the BL task with $\pi = 0.5$ and price sequence $p = 0.23, 0.58, 0.81, 0.93, 1.00, 1.23$. The remaining blocks (2-5 and

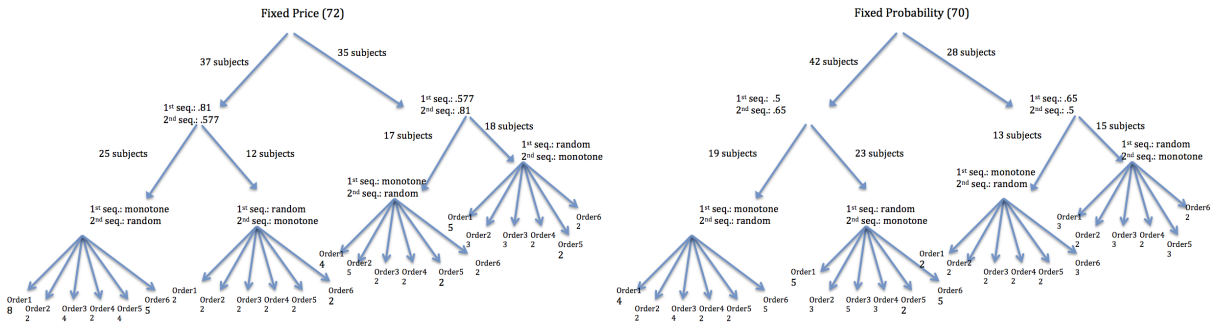

Figure 5: Experimental Design Summary. Sessions are balanced across several design dimensions: fixed price vs fixed probability, order of the two levels at which price or probability is fixed, monotone vs random sequencing of varying price (or probability), and treatment order across blocks.

7-10) use other elicitation tasks. In a given session, the tasks used in blocks 2-5 are the same as those used in blocks 7-10.

We place subjects in two sorts of sessions: fixed price and fixed probability. As summarized in Figure 5, about half of the fixed price sessions keep price at .81 in blocks 2-5, and keep it fixed at .58 in blocks 7-10, while the other fixed price sessions use $p = .58$ in blocks 2-5 and $p = .81$ in blocks 7-10. Within each interior block (2-10) of the fixed price sessions, each of the six probabilities (from $\pi_X = 0.3$ to 0.8) is used once. In half of these sessions a monotone increasing sequence of probabilities is used in blocks 2-5, while random sequences are used in blocks 7-10. In the other fixed price sessions random sequences are used in blocks 2-5 and the monotone sequence is used in blocks 7-10. Each fixed price session uses the elicitation tasks BL, BDHL, BJ, BJn once each in the first four blocks and 6 rounds each in the last four blocks. The monotone sequence sets of blocks also group trials according to elicitation task; e.g. the 6 rounds of BJ are consecutive, in addition to employing a monotone sequence of probabilities from earlier to later rounds of BJ. In contrast, the random sets of blocks randomize which task is implemented, from one round to the next, in addition to randomizing the sequence of probabilities within a task. Of the $4! = 24$ possible task sequences that could be employed in the set of monotone blocks, we selected a balanced subset denoted Order1 thru Order6, and used them with roughly equal frequency across sessions.

The fixed probability sessions have a parallel design. State probability π_X is fixed at 0.5 in blocks 2-5 and at .65 in blocks 7-10, or the reverse. Within each of these blocks we present the six varying prices $p = \frac{p_x}{p_y} = 0.23, 0.58, 0.81, 0.93, 1.00, 1.23$ in monotone increasing sequence or in random sequence (randomizing over task as well as price), and use Order1 through Order6 for the elicitation tasks across the monotone blocks. The four tasks used in these interior blocks is the same as in the fixed price sessions except that BDEG (which requires fixed probabilities) replaces BDHL (which requires fixed prices).

4.3 Implementation

A total of 142 subjects from the LEEPS lab subject pool participated in 18 sessions between October 2016 and March 2017. After subjects privately read instructions (a copy is attached as Appendix C), the conductor demonstrated the mechanics (e.g., sliders and confirm bar) of each elicitation institution, had subjects make practice decisions, illustrated the payoff procedure, and then conducted the paid trials.

After the 56 paid trials were completed, each subject was actually paid for a single trial, determined by a ball drawn from a bingo cage with 56 numbered balls. (If ball 1 or ball 56 came up, indicating a HL trial, then a roll of a six sided die determined the relevant line.) The subject then rolled a ten-sided die to determine which state (X or Y) of the chosen lottery paid that period. Each session lasted about 60 minutes, and the final payments $[min, max]$ range, including \$7 show-up fee, was \$[7.00, 17.00], with average payout roughly \$10.

5 Results

We first report preliminary results, which deal with variables relevant to standard decision theory. The second subsection presents our first two main results, examining the extent to which different elicitation tasks produce consistent results. The third subsection reports a simulation exercise using a noisy choice model. The remaining subsection investigates regularities in our data that lie outside the scope of standard choice theory. Readers who prefer to begin with raw data can turn now to the first part of Appendix B, which includes a visual summary of Block 2-10 choices for each of four sample subjects.

5.1 Preliminary Results

	BL	BJ	BJn	HL
Opportunities	2196	1438	1399	284
Violations	309	424	508	27
(Random)	872	571	558	257
Major Violations	16	9	19	-
(Random)	258	209	201	-

Table 1: Violations of FOSD. “Opportunites" is the number of trials for each task that allowed violations of First Order Stochastic Dominance. “Violations" is the number of such violations. “(Random)" gives, to nearest integer, the expected number of violations given iid uniformly distributed random choices in each task. A violation (x, y) at L is deemed "major" if $L \cdot \ln(\frac{x}{y}) \leq -1$.

As noted in Section 3.4 above, stochastically dominated choices are inconsistent with EUH and many of its generalizations. Table 1 shows the overall prevalence of dominated choices in our experiment. The Table omits the BDEG and BDHL tasks because their feasible sets contain only undominated choices. Multicrossings in 6-row HL trials imply dominated choices, and these appear in the Table's last column. The other columns report first order stochastic dominance violations in the remaining tasks, where Proposition 1 applies. A violation is deemed "major" if its log ratio lies outside the rectangular hyperbola $\ln(\frac{x}{y}) \cdot L = -1$. Table 1 shows a fair number of minor violations of FOSD, but rather few major violations. Table B.3 in Appendix B looks at tighter criteria for major violations, and confirms that a large majority of actual violations are small, due to clicking just a few dozen pixels away from an undominated choice in the BL task, or to purchasing just a little of an asset that is more expensive but not more likely in the BJ tasks. To summarize,

Preliminary Result 1. Dominated choices are uncommon in all tasks, and only about 1% of observations in relevant tasks are major violations of FOSD.

With the reassurance provided by Table 1, subsequent analysis will include all observations, even those that violate FOSD.

Our next preliminary result concerns the two parameter DA model presented in section 3.2, estimated via the Nonlinear Least Squares procedure presented in Appendix B. Figure 6 compares the predictive power of that two parameter model to that of a simple CRRA model or, equivalently, to imposing the restriction $b = 0$. For each subject, we designate one of the observations as the prediction target, estimate both models on the remaining 53 of the 54 observations, predict the target observation, and compute the prediction error for each model. The table reports squared prediction errors summed over all 54 possible prediction targets.

The Figure shows that the majority of our 142 subjects have relatively small prediction errors (SSE < 150) that hardly differ between the two models and so fall almost on top of the diagonal line. The DA model does slightly better than the CRRA model for most (not all) subjects with large SSE's, but (by definition) neither model predicts their behavior very well. See Appendix B for more analysis of the two parameter model and more model comparisons. Thus we have

Preliminary Result 2. Little predictive accuracy is sacrificed by using the single parameter (γ) Constant Relative Risk Aversion model to summarize risk preferences instead of the two parameter (b, γ) Disappointment Aversion model.

Subsequent parametric analysis therefore will focus on the simple CRRA model.

We turn now to the predictions noted just before Section 3.1, that subjects will treat the composite variable $L = \ln(\pi_X) - \ln(\pi_Y) - \ln(P_x) + \ln(P_y)$ as a sufficient statistic and that they will

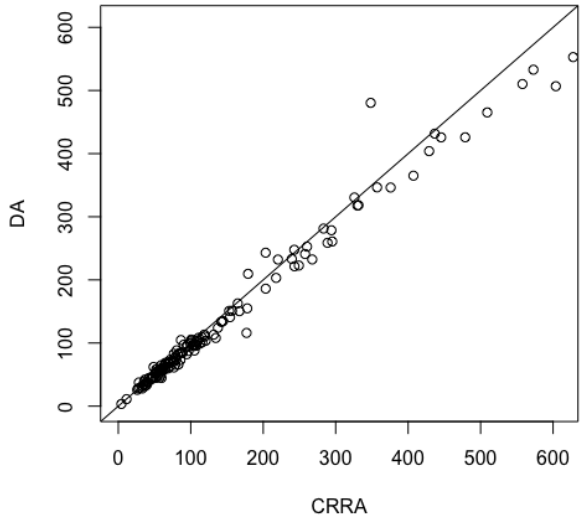

Figure 6: Predictive power comparison. Vertical axis is sum of squared prediction errors (SSE) for the (γ, b) DA model, and horizontal axis is SSE for the CRRA model ($b = 0$).

react symmetrically to prices and probabilities. To test those strong predictions, we run regressions on pooled data of all 142 subjects. The dependent variable is $\ln x/y$, the log portfolio ratio of the actual choice, trial by trial. Of course, that variable is not defined at corner choices, where $\ln x/y = \pm\infty$, so we truncate at $\ln x/y = \pm 4.0$, i.e., we lump together all observations within $\exp(-4) \approx 2\%$ of a corner. The key regression is

$$\ln(\frac{x}{y}) = \alpha_{lo} \ln(\frac{\pi_X}{\pi_Y}) + \alpha_{lp} \ln(\frac{P_x}{P_y}) + \varepsilon. \tag{11}$$

Table 2 presents the results for different subsets of the data. The estimates and F-test in the first column strongly reject the null hypothesis of symmetry ($\alpha_{lo} + \alpha_{lp} = 0$) in favor of the alternative that subjects respond more strongly to probabilities than to prices in the BJ and BJn tasks. By contrast, column (2) shows that symmetry can not be rejected in the BL task. In the increasingly inclusive combined data sets in the remaining columns, the symmetry hypothesis is always rejected in the direction opposite from the "diminishing sensitivity" prediction. Thus

Preliminary Result 3. The hypothesis that subjects react symmetrically to prices and probabilities is strongly rejected in the the budget jar data, and is also rejected in the combined data. Contrary to the diminishing sensitivity hypothesis, subjects tend to react more strongly to probabilities than to prices.

Two remarks are in order. First, even though standard decision theory deems the BL, BJ, and

	Dependent variable: ln(x/y)					
	(1)	(2)	(3)	(4)	(5)	(6)
logOdds	1.336***	1.240***	1.288***	1.320***	1.390***	1.473***
	(0.086)	(0.082)	(0.080)	(0.077)	(0.070)	(0.056)
logPrice	−0.889***	−1.345***	−1.101***	−1.130***	−1.233***	−1.334***
	(0.098)	(0.083)	(0.082)	(0.079)	(0.078)	(0.079)
Observations	3,408	2,556	5,964	6,804	7,668	9,372
R^2	0.430	0.443	0.430	0.441	0.460	0.477
F-stat	53.06***	2.14	16.16***	18.78***	13.48***	12.02***
Note:				*p<0.1; **p<0.05; ***p<0.01		

Table 2: Estimates of Equation (11), with errors clustered at subject level. Column (1) data consists of only BJ and BJn trials. Column (2) data consists of only BL trials, and (3) combines the data from the previous columns. Column (4) adds BDEG trials, and (5) adds BDHL trials as well. Column (6) also includes the HL list data, with each line treated as a separate trial. F-stats are for the parameter restriction $\alpha_{lo} + \alpha_{lp} = 0$.

BJn tasks to be equivalent, they provoke quite different behavior in our subjects. Our main results will expand on this theme, and try to explain it. Second, even though the reaction asymmetries are quite significant statistically, the probability responsiveness of about 1.3 to 1.5 in the more inclusive data sets is not dramatically different than the corresponding price responsiveness of 1.1 to 1.3. Subsequent analysis will therefore often use L to summarize prices and probabilities.

5.2 Are revealed preferences consistent across elicitation tasks?

We now seek to compare systematically, across all elicitation tasks, the risk aversion revealed by individual subjects. For each subject we obtain γ estimates from the six-line HL trials (and the six BDHL trials at given price) using the traditional crossover method, and from BDEG trials using the standard indifference average method, as detailed in Appendix B. For the continuous tasks (BL, BJ, BJn) we exploit equation (7), treating L as a sufficient statistic for prices and probabilities. For each subject, we regress the observed log allocation ratio (truncated at ± 4) on L, and include task-specific interactive dummy variables to detect differences across elicitation tasks:

$$\ln(x/y) = (\beta_1 + \beta_2 BJ + \beta_3 BJn)\, L + \varepsilon, \tag{12}$$

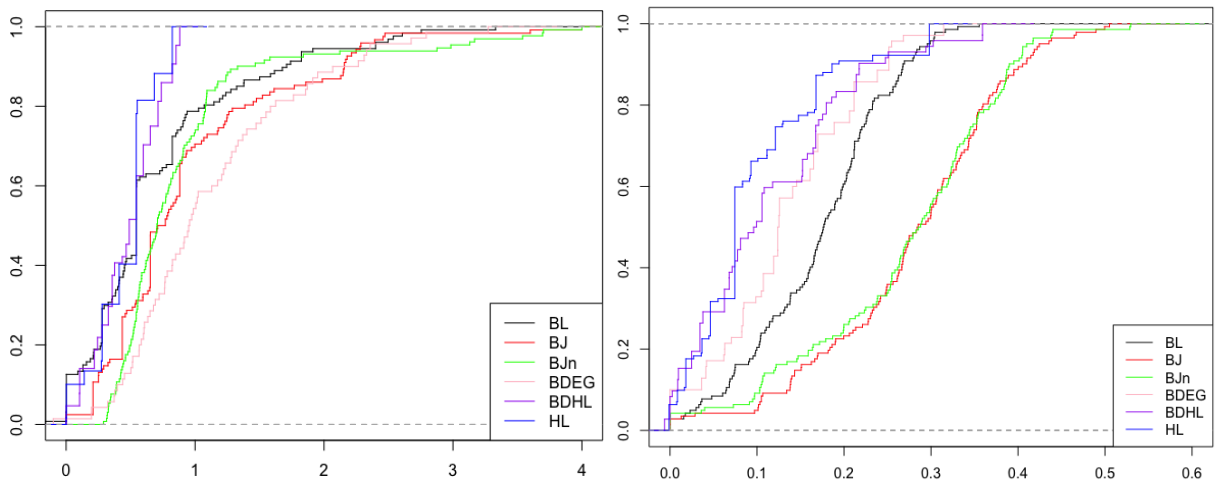

Figure 7: (a) Cumulative distribution functions of $\hat{\gamma}_{i\tau}$ for all subjects. HL, BDHL, and BDEG estimates use established procedures, and other estimates are from OLS fits of equations (12, 13) with errors clustered at task level. (b) Cumulative distribution functions of RRP for all subjects.

for the relevant seven blocks, which contain 42 observations: 18 from BL, and 12 each from BJ and BJn. Revealed risk aversion then is computed from the resulting coefficient estimates via

$$\hat{\gamma}_{i\tau} = 1/(\hat{\beta}_1 + \hat{\beta}_k), \tag{13}$$

where $k = 2$ for treatment $\tau =$ BJ and $k = 3$ for $\tau =$ BJn. Of course, for the omitted treatment $\tau =$ BL, the revealed value is just $1/\hat{\beta}_1$. Jensen's inequality suggests possibly biased γ's when we take the reciprocal of OLS estimates, but that problem disappears with least absolute deviation (LAD) estimates, whose fitted values are medians rather than means. With one exception noted below, Appendix B reports no major differences between the OLS estimates and the LAD estimates.

Panel (a) of Figure 7 collects the results for subjects in all sessions. The lowest γ estimates come from the HL and BDHL tasks; both are approximately uniformly distributed between 0 and 0.8. The median BL γ is similar, but estimates above the median skew towards greater revealed risk aversion. The BJ and BJn regression estimates have higher medians and have a similar upward skew. Here (but not in the LAD estimates or in the RRP estimates) the BDEG distribution is similar but with greater risk aversion revealed in its interquartile range.

Turning now to the non-parametric risk preference measure RRP, recall from equation (10) that, for BJ and BJn as well as BDEG, that measure is defined as the shortfall in expected value of the actual choice C relative to the risk neutral choice M, normalized by the standard deviation σ_M of the risk-neutral choice. In computing RRP, we treat each six-line HL trial as a compound lottery, i.e., C in equation (10) is the expected value of 1/6 chance of playing the chosen (safe or risky)

lottery in each of the 6 rows, and M and σ_M are similarly calculated for the six risk-neutral choices. We treat the six trials in each BDHL block in the same manner. The BJ and BJn distributions of RRP are close to each other in Panel b of Figure 7, and indicate that subjects respond to these tasks as if considerably more risk averse than in the HL and BDHL tasks. The distributions for the BL and BDEG tasks lie in between, a bit closer to the HL distributions.

Standard K-S tests confirm that the apparent differences in Figure 7 are statistically highly significant; see Appendix B. We fail to reject at the 1% level the null hypothesis that the distributions are the same only when comparing the γ distribution in BDEG task to that for the tasks with continuous budget sets, and for comparing the two budget jar distributions for γ and also for RRP. We can reject the null at the 5 or 10% level but not at 1% when comparing the distributions for HL and BDHL.

Main Result 1. Different elicitation tasks generally reveal substantially different distributions of risk preferences. The differences are quite pronounced for dissimilar tasks, and there is some variation even for rather similar tasks.

We now examine subject-level consistency. As stated near the beginning of Section 3, the expected utility hypothesis is that a human subject not only chooses as if maximizing some Bernoulli function, but also that her Bernoulli function is fixed. Such preference stability is crucial from a scientific perspective — the whole point of using some artificial task to elicit risk preferences is that those revealed preferences should enable the researcher to predict behavior in other risky settings of more direct economic interest. As a step towards checking preference stability, we now examine the power of individual subject's choices in one elicitation task to predict their behavior in other elicitation tasks.

Given the substantial differences already observed in the population distributions, the main concern now is the relative ranking among subjects. Does a subject who, say, reveals herself to be among the most (or least) risk averse subjects in one task tend to reveal a relatively high (or low) degree of risk aversion in other elicitation tasks? Specifically, do subjects' Spearman rank correlations across tasks approach $\rho = 1.0$?

Since they employ precisely the same feasible set $F =$ a budget line with the same state probabilities, the BL, BJ, and BJn tasks are the same according to standard decision theory, and BDEG can be thought of as a finite discrete subset version of BL. By the same token, the initial and final period HL trials are identical to each other and to the $p = 0.81$ blocks of BDHL. The expected utility hypothesis predicts identical rankings in all tasks, but the prediction seems especially compelling across these "identical" tasks.

	Fix-Price				Fix-Prob			
	BDHL	BL	BJ	BJn	BDEG	BL	BJ	BJn
HL	0.35	0.28	0.23	0.22	0.15	-0.07	0.03	0.16
BDx		0.75	0.46	0.54		0.46	0.56	0.53
BL			0.74	0.81			0.68	0.62
BJ				0.80				0.86

Table 3: Within-subject Spearman rank correlation of $\hat{\gamma}$. The row BDx refers to BDHL in the Fix-Price Data and to BDEG in the Fix-Prob data.

	Fix-Price				Fix-Prob			
	BDHL	BL	BJ	BJn	BDEG	BL	BJ	BJn
HL	0.50	0.41	0.33	0.37	0.21	0.08	0.06	0.06
BDx		0.61	0.51	0.59		0.61	0.57	0.44
BL			0.71	0.82			0.58	0.47
BJ				0.80				0.69

Table 4: Within-subject Spearman rank correlation of RRP.

Table 3 collects the rank correlations for the parametric measure γ of revealed risk preference. For reference, the correlation is 0.63 between the γ's elicited in first and last trials (both HL). In the fixed price sessions we get higher correlations among the continuous budget set tasks, in the 0.7 - 0.8 range, but the correlation between HL and BDHL is only .35. Even restricting to the $p_x = 0.81$ subset of BDHL (where the tasks are identical according to standard decision theory), that correlation is .36; see Appendix B. The correlation between BL and BDHL is much higher, at 0.75. Is the rough similarity between BL and BDHL in graphical appearance somehow more important than the exact similarity between HL and BDHL in choice sets? We will investigate such questions in the next two subsections.

In the fixed probability sessions, the nice BJ-BJn correlation of 0.86 drops to 0.62 when we substitute the (theoretically identical) spatial interface task BL for BJ, and slips a bit further to 0.53 when we substitute the discrete task BDEG for BL. The correlation drops all the way to 0.16 when we substitute the HL task for BDEG, and it is essentially zero (measured at -0.07) when we then swap BL for BJn (reintroducing a spatial versus text mismatch in interface at this last step). Thus predictive power dissipates completely as we make a series of small changes in tasks that should, according to standard theory, each have minimal impact.

Table 4 tells a generally similar story for the non-parametric measure RRP. In the Fixed Price data, the HL-BDHL correlation is $\rho = 0.50$, and is roughly .5 to .8 among the continuous budget tasks, but their correlations are around .6 with BDHL and .3 - .4 with HL. Again, in the Fixed Probability data, some correlations are near zero while the nicest correlation (again for BJ-BJn)

dissipates as we move up the last column.

Main Result 2. Individual subjects reveal risk preferences that are poorly correlated across tasks with dissimilar bundles of attributes, but that are highly correlated for some tasks with very similar bundles of attributes.

5.3 Stochastic choice simulations

Main Results 1-2 seem inconsistent with standard decision theory, but so far we have not accounted for the fact that humans can not be perfectly accurate and perfectly consistent. Suppose that each individual subject actually has a stable personal Bernoulli function, but responds to each elicitation task with more or less noise. Might that account for our data?

We ran simulations to find out, modeling noisy choice via the random coefficient approach of Wilcox (2008) and Apesteguia and Ballester (2018) with task-specific noise amplitudes. The simulation proceeds as follows; see Appendix A.4 for details. The mean γ estimate in BL, BJ, BJn trials is deemed to be the "true" value γ_i^* for a given subject, which we sort from lowest ($i = 1$) to highest ($i = 142$). For each elicitation task $k = 1, ..., 6$, we construct a noise distribution D_k by collecting (over all subjects i) the deviations in estimated γ in each trial from the subject-specific mean for the task k employed in that trial. Each Monte Carlo simulation trial cycles through the 142 γ_i^*'s as follows. For each of the 56 trials faced by subject i, we map γ_i^* to the i^{th} subject-specific γ observed for that trial's task k, and add an iid draw from D_k. For the CRRA utility function with the resulting γ_{ikt}, we find the expected utility maximizing feasible choice $(x, y) \in F_k$. Thus a single Monte Carlo trial produces simulated data of 142 subjects x 56 choices, which we process in exactly the same way as actual data, yielding a full set of γ and RRP histograms and correlations.

	Fix-Price				Fix-Prob			
	BDHL	BL	BJ	BJn	BDEG	BL	BJ	BJn
HL	47	45	51	54	0	1	7	34
BDx		100	99	100		76	100	100
BL			100	100			100	100
BJ				100				100

Table 5: Human subjects' γ correlations as percentiles of a 1000 trial Monte Carlo simulation. The row BDx refers to BDHL in the Fix-Price Data and to BDEG in the Fix-Prob data.

Table 5 shows where the human subjects' correlations lie in the distribution of 1000 Monte Carlo simulated correlations. Our conventions on noise amplitudes enable the simulations to reproduce nicely the human correlations between HL text and the other tasks in the FixPrice data: these percentiles are all in the vicinity of 50. But those same conventions produce unrealistically low

simulated correlations among the BL, BJ, BJn tasks: the actual correlations lie at or above 99% of the simulated correlations. In the FixProb data the actual correlations are mostly outliers: relative to the simulated correlations, the actual correlations among BL, BJ and BJn are again too high, and here the actual correlations with HL are too low. Other conventions explored in Appendix B shift the percentiles around somewhat, but the conclusion is always the same: random coefficient CRRA agents do not generate correlations that are anything like those of our human subjects. **Main Result 3.** The variation in correlation across pairs of elicitation tasks can not be explained by a random coefficient model of expected utility maximization.

In retrospect, this result is inevitable due to the very different observed correlations for tasks that differ in ways deemed irrelevant by standard (or even noisy) choice theory. As noted, simulated correlations of HL text with other tasks comfortably nest the observed correlations in FixPrice sessions. But switching from text to the spatial display of BDHL dramatically increases the observed correlations while, of course, having negligible impact on the simulated correlations. The shift from FixProb to FixPrice protocol (varying price instead of probability within blocks) also has a major impact on the human data, e.g., lowering the HL correlations with the other tasks. Again, since the simulations respond systematically only to L and not per se to its components, there is no scope for simulations to reproduce that impact.

5.4 How task attributes influence revealed risk preferences

	HL	BDHL	BDEG	BL	BJ	BJn
Spatial	-	✓	✓	✓	-	-
2Dots	✓	✓	-	-	-	-
6Dots	-	-	✓	-	-	-
Cash	-	-	-	-	✓	-
FixProb	-	-	✓	*	*	*
Random	-	*	*	*	*	*
Px58	-	*	*	*	*	*

Table 6: Task attributes. A ✓in the column for a given task indicates that the attributes in that row is always present, a − indicates an attribute never present, and a * indicates an attribute present in some but not all trials using the task.

We are now ready to consider how attributes of our choice tasks that are irrelevant according to standard choice theory might nevertheless play a systematic role. Table 6 rounds up seven possible candidate attributes, of which four pertain to the presentation format and three pertain to the wider environment. The attribute Spatial refers a budget line display in the 2 dimensional space of Arrow portfolios, either allowing choice anywhere on the line (in BL) or on a finite subset of it (BDHL

and BDEG). The attribute 2Dots refers to tasks with only binary choices, either via radio buttons in text (HL) or via two points in Arrow-Debreu 2-space (BDHL). The attribute 6Dots refers to the other discrete choice set which is represented in Arrow-Debreu 2-space, and Cash refers to the attribute (used only in treatment BJ) allowing retention of cash. The environmental attributes are Fix[ed]Prob[ability] sessions (versus Fixed Price), Random (versus monotone) ordering of task and price or probability sequences, and whether the trial uses $p_x = 0.58$.

	γ		RRP	
	(1)	(2)	(3)	(4)
spatial	−0.268***	−0.268***	−0.111***	−0.111***
	(0.093)	(0.039)	(0.007)	(0.010)
Cash	0.032	0.032***	0.015**	0.015**
	(0.085)	(0.000)	(0.007)	(0.006)
2Dots	−0.223*	−0.223***	−0.164***	−0.164***
	(0.134)	(0.039)	(0.014)	(0.037)
6Dots	0.377***	0.377***	−0.024***	−0.024***
	(0.075)	(0.050)	(0.008)	(0.008)
Px58	0.224***	0.224***	0.042***	0.042***
	(0.064)	(0.078)	(0.005)	(0.006)
FixProb	−0.062	−0.062	−0.067***	−0.067*
	(0.159)	(0.060)	(0.021)	(0.035)
randomSecond	0.221*	0.221***	0.030*	0.030
	(0.134)	(0.059)	(0.017)	(0.025)
randomFirst	0.230	0.230***	0.045**	0.045
	(0.175)	(0.048)	(0.020)	(0.030)
Observations	6,752	6,752	6,815	6,815
R^2	0.250	0.250	0.658	0.658
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 7: OLS regression coefficients (and standard errors) for risk preference measures γ and RRP. Regressions include subject level fixed effects with errors clustered at the subject level (columns 1 and 3) and task level (columns 2 and 4). All regressions also include order controls and interactions, reported in the Appendix.

Table 7 reports regressions of revealed risk aversion on attribute dummies. The first two columns report results for the γ measure and the other two columns for the RRP measure. Controls include subject-level fixed effects (not shown) and a long list of order controls and interactions that are reported explicitly in Appendix B. One can see that several attributes indeed have economically

and statistically significant effects. For example, the Spatial coefficient suggests that, consistent with the results in Habib et al. (2017), subjects responding to budget lines (or dots) drawn in a two-dimensional space reveal less risk aversion than subjects responding to text. The effect is substantial, e.g., -0.268 for γ in the fixed price environment. The BDEG attribute of restricting choice to 6 dots on the same side of the perfect hedge increases the elicited γ by almost 0.4, while the discrete restriction in HL or BDHL to 2 dots (one of which is near the corner of the budget line) reduces the elicited γ by about 0.2. The use of the lower price ratio (0.58) in BDHL increases γ by 0.2. Such differences are huge in terms of economic behavior: the difference between risk neutrality and highly risk averse "square root" utility is, of course, 0.5.

Another intriguing result is that the Random sequence attribute (whether in early or late blocks of trials) seems to induce greater revealed risk aversion, reminiscent of the Lévy-Garboua et al. (2012) result on variations in implementation of Holt-Laury. The impact of discreteness (e.g. "Dot6") may be less surprising, but nevertheless seems worth documenting, especially in light of the conjectures of Crosetto and Filippin (2016).

Main Result 4. Much of the variation in location across task-specific distributions of risk preferences can be explained by task attributes, particularly spatial vs text presentation of the choices, discrete vs continuous choice sets, random vs monotone price or probability sequences, and high vs low implicit price ratios.

Can differences in attributes across tasks explain how correlations vary among task pairs? To investigate, we partitioned our 142 subjects into 8 cohorts within which each subject faced the same protocol (fixed price or fixed probability, same ordering in early vs late blocks of price (0.81 or 0.57) or prob and of monotone or random), computed the γ revealed in each trial and, across the subjects in each cohort, computed all trial-pair correlations. That dependent variable was regressed on indicator variables. The Spatial, Cash, Prob and Price independent variables are constructed such that the indicator is one if the trial pairs are mismatched on a given attribute. For DD the indicator is one if both elements of the trial pair have a discrete choice set; while for CD the indicator is one if only one element of the task pair has a discrete choice set. The "both tasks continuous" case is the holdout/reference set. The Random:either variable flags whether either element of the trial pair was generated in a block of trials which was ordered randomly.

Table 8 reports the results. For all eight cohorts we see negative entries for Spatial, significant at the 5% level for three of them, indicating that the revealed risk aversion correlation of a pair of tasks tends to be lower when the opportunities are presented in text in one task and spatially in the other. Mismatches in the fixed price (.81 or .58) or in the fixed probability (.50 or .65) also

	Fix Prob				Fix Price			
cohort N	LE19	EE13	LL23	EL15	EE25	LL18	LE17	EL12
spatial	−0.001	−0.038**	−0.012	−0.024	−0.049***	−0.018	−0.034	−0.046**
	(0.014)	(0.015)	(0.013)	(0.016)	(0.015)	(0.018)	(0.022)	(0.023)
cash	−0.018	0.047***	0.003	0.018	0.025	0.040**	0.032	−0.015
	(0.015)	(0.016)	(0.013)	(0.017)	(0.016)	(0.019)	(0.022)	(0.024)
CD	0.031**	0.016	0.048***	0.016	−0.057**	0.057**	0.118***	0.146***
	(0.014)	(0.015)	(0.013)	(0.017)	(0.022)	(0.027)	(0.032)	(0.034)
DD	0.045	0.099***	0.168***	0.049	0.307**	0.489***	0.358*	0.515**
	(0.031)	(0.033)	(0.028)	(0.035)	(0.134)	(0.161)	(0.193)	(0.204)
prob	−0.063***	−0.010	0.004	−0.122***				
	(0.018)	(0.015)	(0.012)	(0.021)				
price					−0.002	0.023	−0.087***	−0.103***
					(0.017)	(0.020)	(0.023)	(0.024)
random:either	0.091***	−0.139***	−0.123***	0.184***	0.013	−0.075***	0.144***	0.140***
	(0.020)	(0.016)	(0.014)	(0.023)	(0.018)	(0.021)	(0.024)	(0.026)
Constant	0.149***	0.273***	0.244***	0.061***	0.090***	0.142***	0.203***	0.215***
	(0.016)	(0.017)	(0.015)	(0.018)	(0.017)	(0.021)	(0.025)	(0.027)
Observations	1,432	1,432	1,432	1,432	947	947	947	947
R^2	0.021	0.065	0.084	0.049	0.026	0.031	0.059	0.065
Note:								*p<0.1; **p<0.05; ***p<0.01

Table 8: Impact of trial attribute mismatches on rank correlations, by protocol cohort. Column labels identify the cohort treatment and cohort size, e.g., LE19 refers to the N=19 subjects who saw the higher fixed probability (or higher price, in the last 4 columns) in the late (L) blocks and saw monotone ordered trials in the early (E) blocks. The indicator CD (resp. DD) is 1 if the correlation is between a continuous (resp. discrete) task and a discrete task.

tend to reduce correlations, often substantially, and the reduction is significant at the 1% level in half the cohorts. The DD coefficients are all positive, the majority of them significantly so at the 5% level; relative to the omitted case CC (both tasks have continuous feasible sets), having discrete feasible sets in both tasks tends to increase correlations. The increase is impressively large, around 0.3 to 0.5, in the Fix Price cohorts (which pertain to HL and BDHL but not to BDEG). The CD coefficients are also mostly positive but smaller. Cash mismatch has a significantly positive impact in two cohorts. The indicator Random:either is a useful control whose sign varies across cohorts.

Main Result 5. Mismatches in attributes across pairs of tasks often lower the correlation of gamma estimates, in particular for mismatches in the price ratio or state probability, and whether price is displayed spatially. Other things equal, correlations between tasks with discrete choice sets (especially HL and BDHL) tend to be higher than their continuous choice counterparts.

6 Discussion

We report a laboratory experiment in which each of our 142 subjects responds to a variety of direct choice lottery tasks, each of which is intended to elicit personal risk preferences. The tasks vary incrementally, in the state probabilities and prices as well as in attributes such as how choices are presented. For each task and for each subject, we summarize the elicited preferences parametrically, via the coefficient γ of relative risk aversion, and also non-parametrically via RRP, the revealed relative risk premium.

The previous section presented three preliminary and five main results. Taken together, these results suggest two broad conclusions. First, the EUH — that each subject chooses among lotteries as if maximizing the expectation of some personal Bernoulli function — is a decent approximation when each elicitation task is considered in isolation. Violations of first order stochastic dominance restrictions are not infrequent, but most violations are tiny (Biais et al. (2017) reach a roughly similar conclusion in independent recent work); including a parameter for disappointment aversion does little to increase predictive power; and responses to prices and probabilities are not very far from symmetric (although we find a statistically significant deviation in the direction opposite from that predicted by "diminishing sensitivity," a well-known component of behavioral choice theory).

The second conclusion is that, by contrast, the EUH is a rather poor approximation when comparing behavior across elicitation tasks. The EUH allows individual subjects to differ in their elicited risk preferences, and indeed our subjects do differ considerably. However, the EUH (and generalizations such as Prospect Theory) do not allow for substantial differences across elicitation tasks; indeed, for elicited personal Bernoulli functions (or value functions) to be scientifically meaningful, they must have some sort of stability across tasks. To the contrary, we find large differences across elicitation tasks in the population distributions of γ and of RRP, and more importantly we also find considerable instability in the ordering of individual subjects within a distribution. A simulation exercise suggests that including noise in the choice process can not account for the observed patterns in how revealed individual risk preference correlate across elicitation tasks.

The last part of our study seeks regularities across the diverse task-specific revealed distributions and correlations. Task attributes deemed irrelevant by standard choice theory — such as whether the feasible lotteries are presented spatially (via budget lines or budget dots) or in text, and whether price or probability sequences are presented in monotone or random order — turn out to have a substantial impact on both distributions and correlations. These results unify and extend the findings of earlier studies such as Isaac and James (2000), Berg et al. (2005), Loomes and

Pogrebna (2014), Pedroni et al. (2017), and Deck et al. (2013). For example, Habib et al. (2017) find a particular spatial representation (using rotatable cylinders instead of text) pushes revealed preferences towards risk neutrality, but we now see that spatial representation quite generally matters in presenting state probabilities and prices. Likewise, expanding on Lévy-Garboua et al. (2012), we now see that that a monotone (as opposed to random) sequencing of probabilities and prices across trials also generally pushes subjects' behavior towards risk neutrality.

The cumulative impact of such attributes is huge. We saw impressively large correlations (e.g., 0.86) decay all the way to zero as we successively switched from one elicitation task to another that differed only in an attribute that is irrelevant according to standard choice theory.

What are the implications for applied researchers? First, don't expect individual revealed risk preferences to have much predictive power in different circumstances. Second, to best predict individual behavior in some later context, it seems prudent to choose an elicitation task whose attributes most closely resemble those of the target context. For example, to predict portfolio choice in a text-based context, it might be better to use something like our budget jars instead of a multiple price list or budget dots, although one of the latter tasks might be the better choice in some other context.

Our experiment design, with incremental variation of attributes across multiple within-subjects tasks, is well suited for detecting attribute impacts. It is less suited, however, for identifying a specific theory that might explain and predict the impacts of variables not recognized by consequentialist choice theories. We hope that our findings will provide helpful clues to future researchers who seek to develop process-oriented theories in the spirit of, for example, Pleskac and Busemeyer (2010), and Massaro and Friedman (1990), or other sorts of trans-consequentialist theories. On the empirical side, we hope that our study encourages new laboratory (and perhaps field) experiments investigating the persistence and robustness of behavioral differences across wider sets of tasks (not necessarily direct choice) with different formats but identical opportunities and payoffs.

References

James Andreoni and William T. Harbaugh. Unexpected utility: Experimental tests and five key questions about preferences over risk. 2009.

James Andreoni, Michael A. Kuhn, and Charles Sprenger. Measuring time preferences: A comparison of experimental methods. *Journal of Economic Behavior and Organization*, 116:451–464, 2015.

Jose Apesteguia and Miguel A Ballester. Monotone stochastic choice models: The case of risk and time preferences. *Journal of Political Economy*, 126(1):74–106, 2018.

Gordon M. Becker, Morris H. DeGroot, and Jacob Marschak. Measuring utility by a single-response sequential method. *Behavioral Science*, 9(3):226–232, 1964.

Joyce Berg, John Dickhaut, and Kevin McCabe. Risk preference instability across institutions: A dilemma. *Proceedings of the National Academy of Sciences*, 102(11):4209–4214, 2005.

Bruno Biais, Thomas Mariotti, Sophie Moinas, and Sébastien Pouget. Asset pricing and risk—sharing in a complete market: An experimental investigation. 2017.

Hans P. Binswanger. Attitudes toward risk: Experimental measurement in rural India. *American Journal of Agricultural Economics*, 62(3):395–407, 1980.

Pavlo R. Blavatskyy. Stronger utility. *Theory and Decision*, 76(2):265–286, 2014.

Colin F. Camerer and Teck-Hua Ho. Violations of the betweenness axiom and nonlinearity in probability. *Journal of Risk and Uncertainty*, 8(2):167–196, 1994.

Syngjoo Choi, Raymond Fisman, Douglas Gale, and Shachar Kariv. Consistency and heterogeneity of individual behavior under uncertainty. *The American Economic Review*, 97(5):1921–1938, 2007.

Sean M. Collins and Duncan James. Response mode and stochastic choice together explain preference reversals. *Quantitative Economics*, 6:825–856, 2015.

James C. Cox, Bruce Roberson, and Vernon L. Smith. Theory and behavior of single object auctions. In Vernon L. Smith, editor, *Research in Experimental Economics, Vol. 2*, pages 1–44. JAI Press, 1982.

James C. Cox, Vernon L. Smith, and James M. Walker. Theory and individual behavior of first-price auctions. *Journal of Risk and Uncertainty*, 1(1):61–99, 1988.

James C. Cox, Vjollca Sadiraj, and Ulrich Schmidt. Paradoxes and mechanisms for choice under risk. *Experimental Economics*, 18(2):215–250, 2015.

Barbara Erin Crockett and Sean Crockett. Endowments and risky choice. 2018.

Paolo Crosetto and Antonio Filippin. A theoretical and experimental appraisal of four risk elicitation methods. *Experimental Economics*, 19(3):613–641, 2016.

Paolo Crossetto and Antonio Filippin. The "bomb" risk elicitation task. *Journal of Risk and Uncertainty*, 47(1):31–65, 2013.

Cary Deck, Jungmin Lee, Javier A Reyes, and Christopher C Rosen. A failed attempt to explain within subject variation in risk taking behavior using domain specific risk attitudes. *Journal of Economic Behavior & Organization*, 87:1–24, 2013.

Catherine Eckel and Philip J. Grossman. Sex differences and statistical stereotyping in attitudes toward financial risk. *Evolution and Human Behavior*, 23(4):281–295, 2002.

Catherine Eckel and Philip J. Grossman. Forecasting risk attitudes: An experimental study using actual and forecast gamble choices. *Journal of Economic Behavior and Organization*, 68(1):1–17, 2008.

David Freeman, Yoram Halevy, and Terri Kneeland. Eliciting risk preferences using choice lists. 2016.

Daniel Friedman, R. Mark Isaac, Duncan James, and Shyam Sunder. *Risky Curves: On the Empirical Failure of Expected Utility*. Routledge, 2014.

Uri Gneezy and Jan Potters. An experiment on risk taking and evaluation periods. *The Quarterly Journal of Economics*, 112(2):631–645, 1997.

David Grether. Financial incentive effects and individual decisionmaking. Technical Report 401, California Institute of Technology Social Science Working Paper, 1981.

Faruk Gul. A theory of disappointment aversion. *Econometrica*, 59(3):667–686, 1991.

Sameh Habib, Daniel Friedman, Sean Crockett, and Duncan James. Payoff and presentation modulation of elicited risk preferences in MPLs. *Journal of the Economic Science Association*, 3(2): 183–194, 2017.

William T. Harbaugh, Kate Krause, and Lise Vesterlund. The fourfold pattern of risk attitude in choice and pricing tasks. *Economic Journal*, 120(3):595–611, 2010.

Jan Heufer. Nonparametric comparative revealed risk aversion. *Journal of Economic Theory*, 153: 569–616, 2014.

John Hey and Chris Orme. Investigating generalizations of expected utility theory using experimental data. *Econometrica*, 62(6):1291–1326, 1994.

Charles A. Holt and Susan K. Laury. Risk aversion and incentive effects. *The American Economic Review*, 92(5):1644–1655, 2002.

Leonid Hurwicz. On informationally decentralized systems. In C. B. McGuire and Roy Radner, editors, *Decision and Organization*, pages 297–336. North Holland, Amsterdam, 1972.

R. Mark Isaac and Duncan James. Just who are you calling risk averse? *Journal of Risk and Uncertainty*, 20(2):177–187, 2000.

Duncan James. Incentive compatible elicitation procedures. Perth, Australia, December 2011. 19th International Congress on Modelling and Simulation.

Steven J. Kachelmeier and Mohamed Shehata. Examining risk preferences under high monetary incentives: Experimental evidence from the People's Republic of China. *The American Economic Review*, 82(5):1120–1141, 1992.

Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–292, 1979.

Edi Karni and Zvi Safra. 'preference reversal' and the observability of preferences by experimental methods. *Econometrica*, 55(3):675–85, 1987.

C. W. Lejuez, Jennifer P. Read, Christopher W. Kahler, Jerry B. Richards, Susan E. Ramsey, Gregory L. Stuart, David R. Strong, and Richard A. Brown. Evaluation of a behavioral measure of risk taking: The balloon analogue risk task (bart). *Journal of Experimental Psychology*, 8(2): 75–84, 2002.

Louis Lévy-Garboua, Hela Maafi, David Masclet, and Antoine Terracol. Risk aversion and framing effects. *Experimental Economics*, 15(1):128–144, 2012.

Sarah Lichtenstein and Paul Slovic. Reversals of preference between bids and choices in gambling decisions. *Journal of Experimental Psychology*, 89(1):46–55, 1971.

Sarah Lichtenstein and Paul Slovic. Response-induced reversals of preference in gambling: An extended replication in Las Vegas. *Journal of Experimental Psychology*, 101(1):16–20, 1973.

Graham Loomes and Ganna Pogrebna. Measuring individual risk attitudes when preferences are imprecise. *The Economic Journal*, 124(576):569–593, 2014.

Andreu Mas-Colell, Michael D. Whinston, and Jerry R. Green. *Microeconomic Theory*. Oxford University Press, 1995.

Dominic W Massaro and Daniel Friedman. Models of integration given multiple sources of information. *Psychological review*, 97(2):225, 1990.

Andreas Pedroni, Renato Frey, Adrian Bruhin, Gilles Dutilh, Ralph Hertwig, and Jörg Rieskamp. The risk elicitation puzzle. *Nature Human Behavior*, 1:803–809, 2017.

Timothy J Pleskac and Jerome R Busemeyer. Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. *Psychological review*, 117(3):864, 2010.

Vernon L. Smith. Experiment economics: Induced value theory. *American Economic Review, Papers and Proceedings*, 66(2):274–279, 1976.

Vernon L. Smith. Theory, experiment and economics. *Journal of Economic Perspectives*, 3(1): 151–169, 1989.

Charles Sprenger. An endowment effect for risk: Experimental tests of stochastic reference points. *Journal of Political Economy*, 123(6):1456–1499, 2015.

Stefan T. Trautmann and Gijs van de Kuilen. Prospect theory or construal level theory? diminishing sensitivity vs psychological distance in risky decisions. *Acta Psychologica*, 139(1):254–260, 2012.

Amos Tversky and Daniel Kahneman. Prospect theory: An analysis of decision under risk. *Journal of Risk and Uncertainty*, 5(4):297–323, 1992.

Nathaniel T Wilcox. Stochastic models for binary discrete choice under risk: A critical primer and econometric comparison. In *Risk aversion in experiments*, pages 197–292. Emerald Group Publishing Limited, 2008.

Wenting Zhou and John Hey. Context matters. *Experimental economics*, 21(4):723–756, 2018.