Games and Economic Behavior 133 (2022) 58–76

Contents lists available at ScienceDirect

Games and Economic Behavior

www.elsevier.com/locate/geb

Varieties of risk preference elicitation ☆

Daniel Friedman [a], Sameh Habib [b], Duncan James [c,*], Brett Williams [d]

[a] Essex University and University of California, Santa Cruz, Santa Cruz, CA 95064, United States of America
[b] The Joint Committee on Taxation, United States of America
[c] Economics Department, Fordham University, United States of America
[d] University of California, Santa Cruz, Santa Cruz, CA 95064, United States of America

ARTICLE INFO

Article history:
Received 20 June 2021
Available online 17 February 2022

JEL classification:
C91
D81
D89

Keywords:
Risk aversion
Experiment
Elicitation
Multiple price list

ABSTRACT

We explore risk preference elicitation when subjects choose directly from an exogenously specified set of lotteries. Our choice tasks differ incrementally, e.g., from choosing between two lotteries to selecting a portfolio from a continuous set of bundled Arrow securities, and from text to spatial presentation. Each subject completes multiple instances of five different tasks, and responses for each task are summarized in parametric (CRRA) and non-parametric (normalized risk premium) measures of risk preference. Variation in task attributes explains much of the observed wide variation in elicited preferences and in correlations across task pairs.

© 2022 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

In order to explain and predict risky choice behavior, several generations of economists have proposed a variety of procedures or tasks intended to elicit individual risk preferences. This research program operates within the intersection of pure theory, applied design, experimental practice, and econometric inference. Deepening the understanding of risky choice behavior thus depends on deeper understanding of those component research tools, both individually and in their joint usage.

That researchers to date have generally found that preferences elicited using one task have very limited power to predict behavior in even a different elicitation task, much less the wider world, is both the challenge to which this paper responds, and the point of departure for our methodological approach. Specifically, for each subject we implement several elicitation tasks. Each task has in common that the subject chooses a lottery from an exogenous set of lotteries; otherwise, each elicitation task embodies a different particular bundle of task attributes. Each task attribute has two possible settings (e.g. spatial representation of information, or not). Thus each task can be characterized in a database as a list of "either/or"

☆ 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 input on early stages of this project, and to Nat Wilcox for input on data analysis. For helpful comments we thank Gabriele Camera, Dirk Engelmann, Paul Feldman, Mikhail Freer, Frank Heinemann, Eric Kimbrough, Jessie Li, Amma Panin, David Sungho Park, Stefan Trautmann (and students), Roel van Veldhuizen, 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.

* Corresponding author.
E-mail addresses: dan@ucsc.edu (D. Friedman), sameh.habib@jct.gov (S. Habib), dujames@fordham.edu (D. James), bwillia4@ucsc.edu (B. Williams).

https://doi.org/10.1016/j.geb.2022.02.002
0899-8256/© 2022 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

settings of attributes. This allows estimation of the marginal effect of each task attribute on risky choice behavior. Rather than viewing it as an impediment, we utilize the lack of agreement in behavior across elicitation tasks to assess the effects of task design on subject behavior.

We execute the joint usage of research tools as follows. Our theoretical framework is straightforward and recounted in Section 3. Our design is innovative; how to implement tasks previously employed by Gneezy and Potters (1997), Holt and Laury (2002), Eckel and Grossman (2002), and Choi et al. (2007) so as to vary incrementally the bundle of task attributes from one task to the next is recounted in detail in Section 4. In order to allow consistent estimation of the standard coefficient of relative risk aversion γ without the use of instrumental variable techniques, we exogenously vary prices, payoffs, and probabilities (parameters documented in Appendix D); Appendix F (instructions) completes the practical details. Our inferential approach is thorough and open-minded; we estimate risk preference parameters both by traditional means (e.g. crossover method of Holt and Laury (2002)) and newer means more attuned to the possibility of noise in behavior (e.g. Wilcox (2011)); the various inferential techniques are addressed in Section 5, with some additional discussion in Appendix A.

The resulting empirical findings document stark effects of design on behavior. Spatial representation of information strongly shifts choice behavior towards risk neutrality. Correlations between tasks are lowered by mismatches in task attributes, across the tasks being correlated. Our documentation of results begins with a baseline, in subsections 6.1 and 6.2, showing that indeed estimated individual subject γ varies across elicitation tasks. (Estimates obtained using a nonparametric metric, Revealed Risk Premium, defined in Section 5, are also presented.) Thereafter Section 6 traces the effect that each task attribute has on the level of estimated γ, and the effects that mismatches in bundles of attributes, across two tasks, have on the estimated correlation between those tasks. Section 6 also contains two robustness checks. One shows that Apesteguia and Ballester (2018) agents produce lower correlations than do human subjects. The other shows that use of the ORIV estimator of Gillen et al. (2019) does not eliminate the variation in correlation, from one task pair to another, which we suggest is driven by varying mismatches of attribute bundles across task pairs.

Applied researchers might utilize our findings by choosing their elicitation procedure so as to match best the task attributes across the predicting task and the predicted activity; our results on correlations across tasks suggest that this is a necessary condition for an elicitation procedure to generate useful control data. In the long run, decision theory might be able to encompass our empirical findings, and we also offer some speculation to this end.

2. Some relevant literature

Evidence has accumulated over many decades that individual humans make inconsistent choices across risk preference elicitation tasks.[1] Slovic (1962) compares nine psychometric tasks relating to choice under risk; surprisingly, there are about as many negative correlations as positive correlations across pairs of tasks. Lichtenstein and Slovic (1971, 1973) replace psychometric questionnaires with two incentivized elicitation tasks, binary choice and BDM (Becker et al., 1964), and famously find an inconsistency dubbed "preference reversal." Experimental economists Grether and Plott (1979) are not able to eliminate that inconsistency via their own methodological changes. Subsequently, Collins and James (2015) find that a majority of preference reversals disappear when the standard (selling) version of BDM, which elicits monetary values, is replaced by its dual, which elicits responses in probabilities. That supports the Slovic (1975) conjecture that preference reversals might be due in part to a clash in response modes; see also Slovic et al. (1988). That is, response mode differences, between entry of monetary values (interpreted as subjects' certainty equivalents) versus binary choice between lotteries (with probabilities visualized as pie charts), may lead to systematic differences in the consistency of subjects' responses across tasks.[2] The preference reversal literature suggests to us that mismatches across tasks more generally, not just in response mode, might generate inconsistencies in elicited preferences.

The set of incentivized risk-preference elicitation tasks has by now expanded far beyond binary choice and BDM. In order to avoid combinatorial explosion in task attributes, our experiment will include only elicitation procedures in which the subject selects a lottery from an exogenously specified set of lotteries,[3] and thereby exclude asking, bidding, and strategizing of any sort. Leading examples of such elicitation procedures include binary choice as above, and more recently as in Hey and Orme (1994); choice among a handful of lotteries as in Binswanger (1980) and Eckel and Grossman (2002); choice from a continuous budget line for Arrow securities, as in Choi et al. (2007), or Andreoni and Harbaugh (2009) or Andreoni et al. (2015); choices from multiple price lists as in Holt and Laury (2002); and even parameterization of a lottery by allocation between cash and a risky investment, as in Gneezy and Potters (1997).

We will thus focus our design upon those differences in operational mechanics that still remain within this set of tasks. For example, the Gneezy and Potters investment task and selection from a budget, as traditionally implemented, differ in

[1] A very abridged list of papers extending the cross-task inconsistency evidence includes Isaac and James (2000); Berg et al. (2005); Dave et al. (2010); Deck et al. (2013); Loomes and Pogrebna (2014); Collins and James (2015); Sprenger (2015); Pedroni et al. (2017); Zhou and Hey (2018); Crosetto and Filippin (2016); Charness et al. (2018, 2020).

[2] The idea here is that the original form of BDM requires subjects to return an answer in monetary units, while binary choice over probability-area pie charts induces a visual focus on probabilities, prior to subjects' selection of one of the pie charts. Conversely, accepting subject responses in units of probability would eliminate a procedural discordance.

[3] See Harbaugh et al. (2010) and Trautmann and van de Kuilen (2012) for evidence that such tasks might give expected utility theory its best shot.

that the former is displayed as a text/numeric problem, while the latter is displayed within a 2-D metric space (e.g., choice among ordered pairs of Arrow-Debreu securities).

Could such differences degrade response correlations between those tasks? The psychology literature provides suggestive analogies. In realms other than risky choice, numerous articles document the impact of visual apprehension of displayed numbers vs displayed area, and of visual cortex information processing, as well as the correlation (high or low) of spatial ability with other abilities. One early study on estimating the number of displayed objects is by none other than Jevons (1871); a particularly interesting recent study is Ross (2003). Dehaene and Cohen (1991) note that the ability to approximate (or at least to reject inaccurate approximations) can be present even in subjects with brain damage that has taken away the ability consciously to do arithmetic. They attribute this residual numerical sense to preconscious acquisition and processing of visual information. New research using ERP (event-related potential) identifies brain areas involved in assessing the number of objects (e.g. Fornaciai et al. (2017)). The EEG (electro-encephalogram) experiment of Van Rinsveld et al. (2020) suggests independent preconscious processing of some attributes of dot displays (e.g., number of dots or area spanned) but not others (e.g., dot size or density).

One also notes that in psychology, implementing similar tasks via different physiological channels can be considered to produce different tasks measuring different abilities, or engaging different parts of the brain. For example, a digit-span task measures a subject's ability to recall a sequence. But there are multiple ways to present a sequence to a subject. The original form of Hebb's (1961) task presents sequences of numbers to the subject, via speech; Corsi's (1972) block-tapping task presents an array of squares on a screen, and illuminates one square at a time in a sequence. The former is used to assess verbal memory span, the latter is used to assess visuo-spatial memory span; they are regarded as measuring different abilities. See Donolato et al. (2017) for a survey on differences in verbal and visuo-spatial recall.

The economics literature has had much less to say about the impact of different forms of information delivery. Some preliminary evidence is provided by Habib et al. (2017): subject behavior changes, to be less risk averse, when a Holt-Laury multiple price list is displayed spatially (as rotating cylinders) than when the same choices are displayed in text.

Another difference that some economists have found to matter is whether a random or a monotone sequence is used to present a set of lotteries; see Lévy-Garboua et al. (2012) and Habib et al. (2017). Psychologists studying the digit-span task also find that sequencing makes a difference; see Donolato et al. (2017).

3. Theoretical perspectives

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 if state X is realized and y points if state Y.

According to standard economic theory, only the opportunity set F (with associated probabilities) and the subject's preferences matter; how F is presented to the subject and how decisions are recorded should be irrelevant, so long as they are clear and unambiguous.

The Expected Utility Hypothesis (EUH) is a leading example of standard economic 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 and probabilities π_X and $\pi_Y = 1 - \pi_X$ so that subjects' choices reveal key aspects of their utility functions u.

For some elicitation tasks, F is a standard budget set: non-negative bundles that are affordable. Assuming only that higher payoffs are preferred to lower payoffs, there is no loss of generality in replacing the budget set 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 the elicitation (sub)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

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

or as

$$\ln \frac{u'(x)}{u'(y)} = -L, \text{ where} \tag{5}$$

$$L \equiv \ln \pi_X - \ln \pi_Y - \ln p_x + \ln p_y. \tag{6}$$

Thus, for whichever utility function u a subject may have, the expected utility hypothesis implies that the composite variable L is a sufficient statistic for prices and probabilities. Equation (5) holds at interior solutions, and corner solutions are also defined by L: when the usual non-negativity constraints $x, y \geq 0$ are included with (2), the 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 $\ln \frac{u'(0)}{u'(\frac{m}{p_y})} \leq -L$.

Before looking at useful special cases of EUH, we note that companion paper Williams and Habib (2021) shows that L remains a sufficient statistic in some popular generalizations of EUH. Appendix C demonstrates that our subjects indeed respond approximately symmetrically to prices and probabilities but also notes that the observed departures from symmetry are contrary to the traditional notion of diminishing sensitivity. Companion paper Williams (2021) uses L to show that choices that violate first order stochastic dominance are uncommon in our data, and that the vast majority of the violations are small.

3.1. Special cases

For a **risk neutral** agent we have $u'(x) = u'(y) = constant > 0$, and (3) becomes

$$\frac{\pi_Y}{p_y} = \frac{\pi_X}{p_x}. \tag{7}$$

Equation (7) can only be satisfied if $L = 0$. Otherwise one gets a corner solution — a 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 utility 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 utility 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 those expressions into (4) and taking logs yields

$$\ln \frac{x}{y} = \frac{-1}{\gamma}[\ln \pi_Y - \ln \pi_X - \ln p_y + \ln p_x] = \frac{1}{\gamma} L. \tag{8}$$

That is, regressing the log of the chosen allocation ratio on L will directly reveal (as the inverse slope) the subject's coefficient, γ, of relative risk aversion.

4. Laboratory procedures

Our design is built around three principles: task mutation, balance, and parameter variation. The most important principle is task mutation, wherein we move from one task to the next by means of a single change in the bundle of task attributes. Incremental change is what will allow us to estimate marginal effects associated with each task attribute, which has not been done before. Additionally, we balance the design in the sense of using different orderings of tasks, and different orderings of parameterizations within a task, thus controlling for order effects. Finally, we employ variation within each task, in the sense of employing a wide variety of parameters (state prices or probabilities) across trials.

4.1. Task mutation

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 Fig. 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 X state probability varies from 0.3 to 0.8.

Budget Jars (BJ and BJn). Fig. 2 shows an alternative elicitation procedure and user interface that presents the same feasible set as in Budget Line. 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), at a rate proportional to the price of that security, as the subject drags up (resp. down) the level in the red (security X) or blue (security Y) jar. The text below the jars spells out the state contingent payoffs (and state probabilities) at the current allocation. The subject clicks the Submit bar to finalize the current allocation. We refer to the task as Budget Jars no cash (BJn) when the submit bar is grayed out (not click-able) until the cash jar is empty. An advantage of the Budget Jars treatments, not exploited in the present paper, is that they can easily accommodate three or more Arrow securities. With two securities, the final BJn allocations map 1:1 onto the budget line.

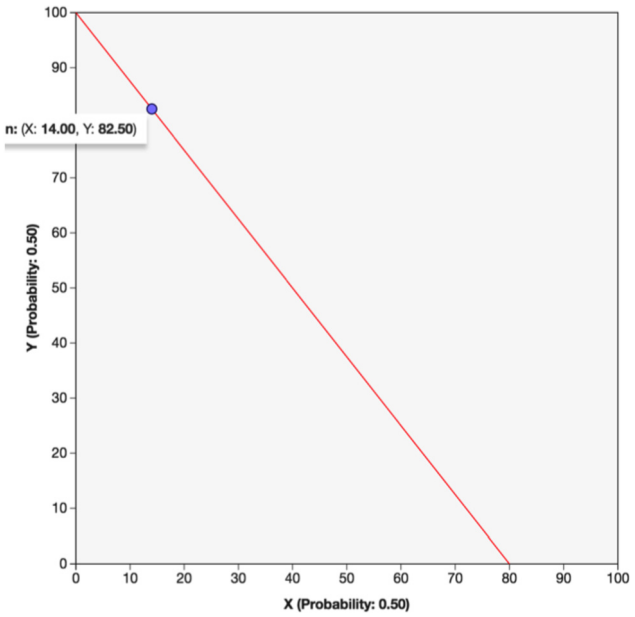

Fig. 1. In treatment Budget Line (BL), the subject chooses a portfolio of Arrow securities by clicking any point on a given budget line, then clicking Confirm bar (not shown). Text box shows values (x, y) at clicked point, here $(14.00, 82.50)$. Axis labels note π_X and π_Y; here, each is 0.50.

Fig. 2. In treatment Budget Jars (Budget Jars), 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 Budget Jars no cash (BJn), the Submit bar becomes active only when the cash jar is empty. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

The task Budget Jars is the same as Budget Jars (no cash), except that cash retention is allowed. The comparison between Budget Jars and Budget Jars (no cash) isolates the impact of allowing cash retention, while the comparison between Budget Line and Budget Jars (no cash) isolates the impact of 2D spatial representation of lotteries, versus otherwise.

From a choice-theoretic perspective (i.e., ignoring differences in the user interface), the Investment Game (IG) of Gneezy and Potters (1997) is a special case of our Budget Jars task, one where investment in the Y-security is capped at the amount implied by the initial cash endowment. In the same sense, IG is also equivalent to a restricted version of Budget Lines, wherein choice could only occur along a line segment connecting the $x = y$ diagonal to the axis associated with the cheaper security. Indeed, from a choice-theoretic perspective, the next task, Budget Dots Eckel-Grossman, is a discrete version of an appropriately parametrized IG.

Budget Dots (BDEG). Eckel and Grossman (2002, 2008) ask subjects to choose 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 graphically displaying the menu F as in Fig. 3a: equally spaced discrete points on a budget line spanned by the intercept for the cheaper security and a

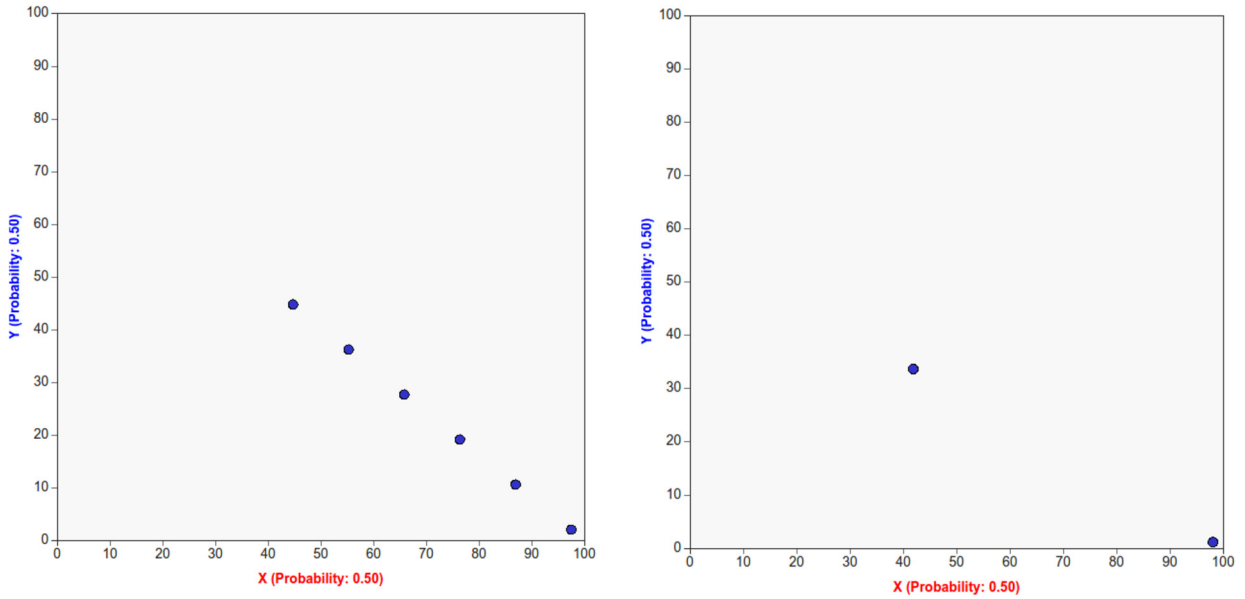

Fig. 3. Discrete budget dots. Axis labels note π_X and π_Y; here, each is 0.5. (a) In treatment Budget Dots Eckel-Grossman, the subject chooses an allocation of Arrow securities by clicking one of the six large dots on the given budget line, then clicking Confirm bar. (b) In treatment Budget Dots Holt-Laury, subjects click one of two dots representing the two feasible Holt-Laury allocations.

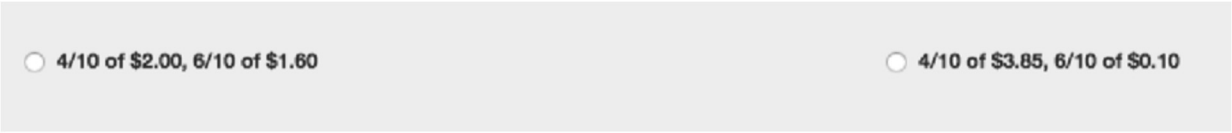

Fig. 4. In treatment Holt-Laury (HL), subjects click a radio button to choose between two feasible allocations in each line; the X state probability (here 0.40) increases by 0.10 from one line to the next.

perfectly hedged portfolio ($x = y$). Some of our Budget Dots Eckel-Grossman trials, unlike the original, use unequal state probabilities. Holding constant prices and probabilities, comparing Budget Dots Eckel-Grossman and Budget Line choices isolates the impact of taking a discrete subset of the budget line.

Multiple price list (HL, BDHL). Perhaps the most widely used elicitation task over the last two decades 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 Holt-Laury treatments use Holt and 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, $\pi_X = 0.3, 0.4, 0.5, 0.6, 0.7, 0.8$.[4] Treatment Holt-Laury stacks six rows of text in a single screen, each row representing choice between two lotteries as in Fig. 4, with $\pi_X = 0.3$ in the top row and increasing by 0.1 in each successive row. Treatment Budget Dots Holt-Laury takes the lotteries from one row (i.e., with a particular π_X value) from Holt-Laury and displays the two feasible choices graphically, as in Fig. 3b, 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 0.58 instead of the original 0.81. The comparison between Holt-Laury and Budget Dots Holt-Laury again isolates the impact of text vs. spatial representation of lotteries.

4.2. Balance

The design is best communicated by means of Fig. 5, which captures its branching, hierarchical nature. There are two trees, one for an environment in which probability varies while price is fixed and one for price varying while probability is fixed. Each tree branches according to possible orders in which tasks and environments were presented. Both trees have

[4] 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 rows from the list.

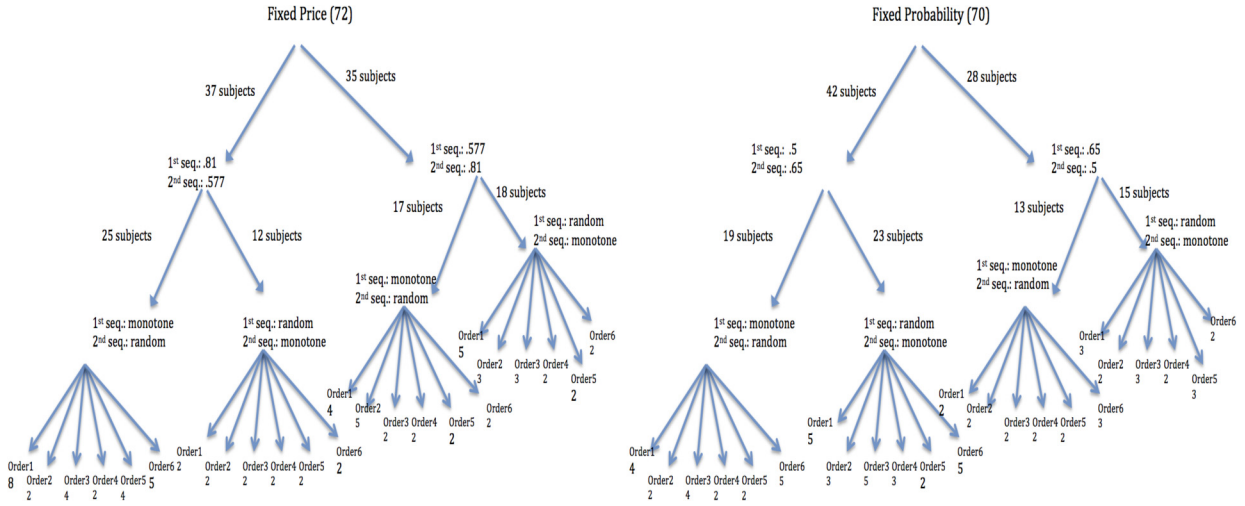

Fig. 5. Experimental Design Summary. Sessions are balanced across several design dimensions: fixed price vs fixed probability, order of the two levels at which price or probability is fixed, monotone vs random sequencing of varying price (or probability), and treatment order across blocks. Numbers displayed beneath each terminal node denote the number of subjects that encountered the unique path (set of treatments) to that node. For example, the leftmost node summarizes the treatments of the 8 subjects in fixed price sessions using price ratio 0.81 and monotone trial sequences in its first set of blocks (and price 0.57 and random in its second set), and used Order1 for the four tasks in each set of blocks.

24 terminal nodes, each representing the consequence of following a particular path (i.e. using a particular combination of order controls) through the tree.[5] See Appendix D for exhaustive narration of design and parameterization.

4.3. Parameter variation

We vary widely the parameters in the task environments. There are six relative prices: 0.23, 0.58, 0.81, 0.93, 1.0, and 1.23. The price 0.81 is derived from the ratio of state payoffs in the original version of Holt-Laury. When prices are varied, the probability π_X of the x-state is held constant at either .5 or .65; sessions of this sort are referred to as *fixed-probability*. When probabilities are varied, the price ratio is held constant at either 0.81 or 0.58; sessions of this sort are referred to as *fixed-price*. The x-state probabilities in these sessions are .3, .4, .5, .6, .7, and .8; as noted earlier, these are the middle six row probabilities in the original version of Holt-Laury.

Each session is divided into 11 blocks. The first and the last block are always a single instance of the six-row Holt-Laury elicitation task. The middle block (block 6 of 11) is always the Budget Lines task with π_X = 0.5 with six trials in the monotone price sequence $p = 0.23, 0.58, \ldots, 1.23$. All four remaining relevant tasks — Budget Lines, Budget Jars, Budget Jars No Cash, and either Budget Dots Eckel-Grossman (in fixed-probability sessions) or Budget Dots Holt-Laury (in fixed-price sessions) — are used in blocks 2-5. Each of these four blocks consists of six trials that use one of those tasks while varying prices or probabilities. The remaining four blocks 7-10 are organized the same way, subject to balancing between monotone and random sequencing and task ordering, as per Fig. 5. Thus each of the 9 middle blocks consists of 6 trials with a single lottery choice, while the first and last blocks each consist of a single trial with 6 lottery choices, so each subject faces a total of $2 \times 6 + 9 \times 6 = 66$ lottery choices, organized into $2 \times 1 + 9 \times 6 = 56$ trials.

4.4. 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 F), 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 Holt-Laury 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] Note that the bottom layer, denoted Order1,..., Order6, represents a selection from a different set of 24, the 4!=24 possible orderings of 4 tasks (Budget Lines, Budget Jars, Budget Jars No Cash, and either Budget Dots Holt-Laury or Budget Dots Eckel-Grossman). Of the 24 possible orderings, a balanced subset of 6 are implemented: in that subset each task gets to be both early and late in the sequence of 4 tasks.

5. Data analysis procedures

Our data analysis relies on revealed risk preference estimates, both parametric and nonparametric. Here we explain how we extract those estimates from raw data. Where appropriate, we exploit the power of linear (and/or limited dependent variable) regressions.

5.1. Extracting γ in BL, BJ, BJn and BDEG trials

A. Single trial extraction. Recall that γ is the coefficient of relative risk aversion for a decision maker with a CRRA utility function, and more generally is a parametric measure of risk preference. For each subject i in each trial ($t = 1, ..., 12$ for tasks $\tau =$ Budget Jars, Budget Jars (no cash) and Budget Dots Eckel-Grossman, and $t = 1, ..., 18$ for task $\tau =$ Budget Line) for which our controlled treatment satisfies $L_t \neq 0$, we invert equation (8) to produce a single trial revealed risk preference parameter[6]

$$\check{\gamma}_{it} = \frac{L_t}{\ln(x_{it}/y_{it})} \ . \tag{9}$$

Therefore, for each subject in the fixed probability treatment we have $\check{\gamma}_{it}$ as defined in equation (9) for trials $t = 1, ..., 54$, and similarly for trials $t = 1, ..., 42$ for fixed price subjects. Our analysis in section 6.3 uses single trial revealed risk preferences, $\check{\gamma}_{it}$, as it addresses the influence of design attributes on single trial subject choice.

B. Multiple trial extraction. Unlike section 6.3, our analysis in section 6.2 compares subjects' gammas across tasks, i.e. across bundles of attributes, rather than across individual attributes. Here we use our design's controlled variation in L_t to consistently estimate subject- and task-specific summaries of risk preferences. For each subject in each task we estimate equation (8) by OLS

$$\ln(x_{it}/y_{it}) = \beta_{i\tau} L_t + \varepsilon_{it} \tag{10}$$

for the allocations (x_{it}, y_{it}) chosen by subject i in trials $t = 1, ..., 12$ for tasks $\tau =$ Budget Jars, Budget Jars (no cash) or Budget Dots Eckel-Grossman, and for trials $t = 1, ..., 18$ for Budget Line. Revealed risk aversion in a given task then comes from the resulting coefficient estimate $\hat{\beta}_{i\tau}$ via

$$\hat{\gamma}_{i\tau} = 1/\hat{\beta}_{i\tau}. \tag{11}$$

Thus for each subject in the fixed probability treatment we have task-specific estimates $\hat{\gamma}_{i\tau}$ for the four tasks $\tau =$ Budget Line, Budget Jars, Budget Jars (no cash) and Budget Dots Eckel-Grossman, obtained by OLS regression. Similarly, for each subject in the fixed price treatment we have $\hat{\gamma}_{i\tau}$ for the three tasks $\tau =$ Budget Line, Budget Jars, Budget Jars (no cash).

Our regression-based $\hat{\gamma}_{i\tau}$ gives greater weight to trials t with larger absolute values of the control variable $L = \ln \pi_X - \ln \pi_Y - \ln p_x + \ln p_y$; Appendix A.4 explains why that OLS weighting provides more reliable estimates than equal weighting or averaging of single trial $\check{\gamma}_{it}$. Another advantage is that the regression (10) provides a standard error estimate for $\hat{\beta}_{i\tau}$ and the delta method extends it to $\hat{\gamma}_{i\tau}$; these s.e.'s are useful for subsequent analysis. We emphasize that the coefficient estimates in (10) are consistent and have no bias attributable to measurement error in the classical sense, as the right-hand side variable L_t is an exogenous and precisely controlled treatment.[7]

5.2. Extracting γ from HL and BDHL trials

Our binary choice tasks Holt-Laury and Budget Dots Holt-Laury data are not suitable for OLS regressions. For those tasks we apply a leading limited dependent variable model, with dependent variable being the indicator R_t (=1 if the subject chose the column A (risky) lottery and = 0 if column B (safe)) in trial t. For the independent variable, we follow Wilcox (2011), who normalizes the utility function to ensure a monotone conditional expectation function. Specifically, we apply logit estimation with explanatory variable

[6] We truncate $\check{\gamma}_t$ at ±4.0 to deal with outliers arising from choices approaching the diagonal $x_t = y_t$. For corner choices, we set $\check{\gamma}_t = 0$ as implied by the Kuhn-Tucker conditions.

[7] Three technical issues deserve brief mention. The left hand side of (10) is not defined at a strict corner allocation choice where x_t or $y_t = 0$; in such cases we replace the 0 by 10^{-3}. Appendix B.2 confirms that our broad results are unchanged when we adopt other conventions. Second, there is a potential Jensen's inequality issue in equation (11), but least absolute deviation (LAD) regressions reported in Appendix B.4 confirm that the issue is inconsequential. Finally, the regression treats the dependent variable $\ln(x_t/y_t)$ the same in the discrete treatment Budget Dots Eckel-Grossman as in the continuous choice treatments Budget Line, Budget Jars and Budget Jars (no cash). We do so because alternative regression procedures such as ordered probit fail to exploit the cardinal (not just ordinal) structure of the Budget Dots Eckel-Grossman choice set, and because with 6 choices, not just 2, the alternative procedures offer no actual advantage and are much more difficult to interpret; see Greene (2012), Chapter 18 for further discussion of this point.

$$v(A_t, B_t|\gamma) = \frac{EU_{A_t} - EU_{B_t}}{u(\overline{m}) - u(\underline{m})} \tag{12}$$

where EU_{A_t} (resp. EU_{B_t}) denotes the expected value of the CRRA utility function $u(c|\gamma) = \frac{c^{1-\gamma}}{1-\gamma}$ for the column A (resp column B) lottery in trial t. The denominator of v normalizes using the difference in utility between the maximum $\overline{m}$ and minimum payoff $\underline{m}$ in the two lotteries considered. Thus, for each subject i, using all 12 Holt-Laury choices (six lines each in Blocks 1 and 11), and similarly for the 12 Budget Dots Holt-Laury choices in the other blocks, we use non-linear least squares to estimate the logit function

$$R_t = (1 + e^{-\omega_{i\tau}[v(A_t, B_t|\gamma_{i\tau})]})^{-1} + e_t, \tag{13}$$

with precision (or scale) parameter, ω.[8]

5.3. Traditional γ-extraction methods for HL and EG

Traditional calculations of γ for Holt-Laury tasks do not involve regressions. As detailed in Appendix B.5, the traditional Holt-Laury calculation is to report the midpoint γ_{co} of a range for which someone maximizing the expectation of a CRRA utility function would cross over from the risky to the safe column, and human subjects who (contrary to noiseless EUH) cross multiple times often are simply dropped. The traditional Eckel-Grossman task calculation is essentially (again, see Appendix B.5 for details) the $\check{\gamma}$ (ratio of choice and L) value at the chosen point on the (discrete) budget line; with multiple trials, it is the simple average $\check{\gamma}$ across trials. In the next section we will present results using both traditional and logit (regression) extraction of γ for these two tasks.

5.4. Extracting a nonparametric measure: relative risk premium

Even if the utility function is not in the CRRA family, the reciprocal γ of the estimated slope coefficient still can serve as a measure of revealed risk aversion. Some researchers may nevertheless prefer a model-free, nonparametric summary measure that captures risk preferences. Unfortunately there apparently is no established such measure that is defined and comparable across all of our elicitation tasks. We considered several possibilities[9] and eventually settled on a normalized (or relative) risk premium, RRP, 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

$$RRP = \frac{M - C}{\sigma_M}, \tag{14}$$

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.[10] It normalizes subject responses in terms of a common benchmark, the risk-neutral choice for any given feasible set.

For tasks Budget Line, Budget Jars, Budget Jars (no cash) and Budget Dots Eckel-Grossman, we simply apply the definition (14) directly to each trial. Some of our analysis uses these single-trial estimates, denoted $R\check{R}P$, but much of the analysis concerns subject- and task-specific summaries. These summaries, denoted $RRP_{i\tau}$, are simple averages of all instances of $R\check{R}P$ for a particular subject i and task τ.

A single Budget Dots Holt-Laury trial, or a single row of a Holt-Laury trial, does not produce a useful preference estimate. Consistent with standard practice, we extract a single estimate from the 6 rows of a Holt-Laury trial, which we treat as a compound lottery. That is, C in equation (14) 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 6 risk-neutral choices. We use exactly the same

[8] For 58 of 142 subjects, the logit regression returns a unique fitted CRRA parameter $\hat{\gamma}_{i\tau}$ for task τ = Holt-Laury which we use for subsequent analysis, and likewise for 40 of 72 subjects who faced task τ = Budget Dots Holt-Laury. Another 71 subjects for task Holt-Laury (and 29 for Budget Dots Holt-Laury) get fits with arbitrarily high precision ω and $\hat{\gamma}_{i\tau}$ indeterminate within a narrow range; this happens when all 12 choices are perfectly consistent with maximizing CRRA expected utility for γ within that range. In such cases, we assign $\hat{\gamma}_{i\tau}$ to be the midpoint of that range, consistent with the original procedure of Holt and Laury (2002). There are also 11 subjects in Holt-Laury trials, and the remaining 3 subjects in Budget Dots Holt-Laury trials, who always choose the safe option. Since this is consistent with any $\gamma \geq 1.34$, we assign to them the upper truncation value $\hat{\gamma}_{i\tau} = 4.0$. Finally, there are 2 subjects whose Holt-Laury choices are so erratic that we can't get fits of (13) with positive precision. Their choices also seem erratic for other tasks, and so we eliminate them from subsequent analysis. Consequently the next section will consider subjects indexed $i = 1, \ldots, 140$, of whom 70 participated in Fixed-Price sessions and the other 70 in Fixed Probability sessions.

[9] 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.

[10] RRP would also be positive for a risk-*seeking* subject willing to forego some expected payoff in order to increase dispersion.

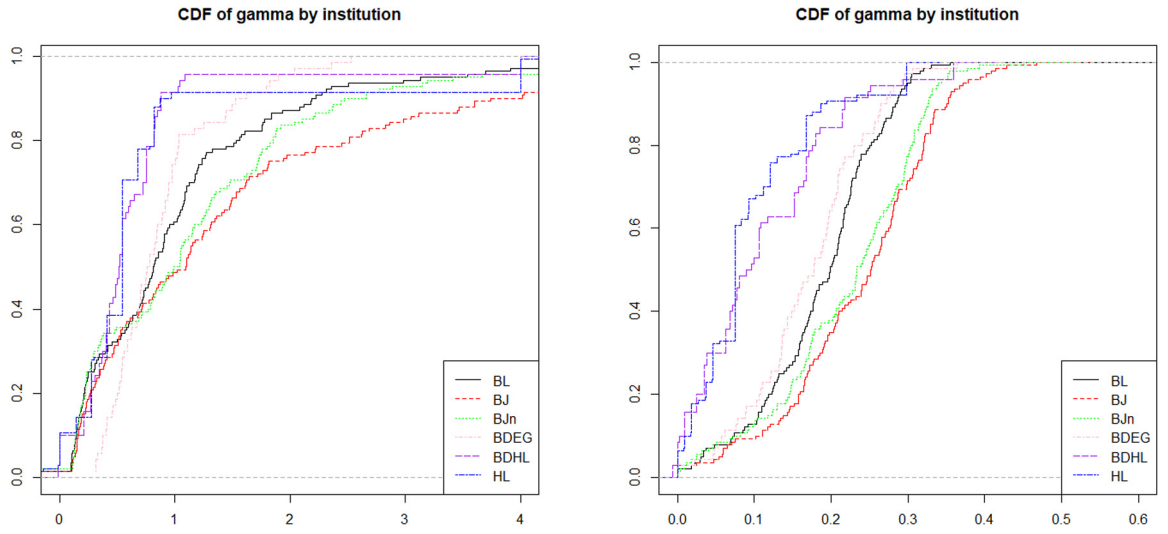

Fig. 6. Empirical cumulative distribution functions for all elicitation tasks. Panel (a): Relative risk aversion parameter $\hat{\gamma}_{jk}$. Panel (b): Relative risk premium RRP_{jk}.

convention to extract an $R\check{R}P$ estimate from each set of 6 Budget Dots Holt-Laury trials. Again we take simple averages of the $R\check{R}P$'s to obtain subject- and task-specific estimates $RRP_{i\tau}$'s for $\tau =$ Holt-Laury and Budget Dots Holt-Laury. Of course, here the average is over 2 instances, rather than 12 instances for $\tau =$ Budget Jars, Budget Jars (no cash) and Budget Dots Eckel-Grossman and 18 instances for Budget Line.

6. Results

We begin by examining the extent to which different elicitation tasks reveal consistent preferences. Section 6.1 compares tasks in terms of their revealed preference distributions across individuals, while Section 6.2 examines the ordinal consistency of individual subjects across tasks. We then seek regularities beneath the inconsistencies that we observe. Section 6.3 defines a set of attributes of elicitation tasks; some attributes can play a role in standard choice theory but others can not. Both sorts of attributes are shown to have an impact on revealed preference distributions, and on correlations. The impact of task attribute mismatches on correlation is shown in Section 6.4. Section 6.5 checks robustness of the preceding results in various ways: via an application of the ORIV procedure (Gillen et al., 2019), and by means of a simulation exercise using a noisy choice model.

Readers who prefer to begin with raw data should 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.

6.1. Are revealed preference distributions consistent across elicitation tasks?

Fig. 6a graphs the empirical cumulative distribution function (cdf) of the individual- and task-specific estimates $\gamma_{i\tau}$ specified in Section 5.1B - 5.2. These are shown separately for each of the six elicitation tasks τ, and each cdf includes all remaining subjects i who faced that task — the 70 fixed-probability subjects for Budget Dots Eckel-Grossman, the 70 fixed-price subjects for Budget Dots Holt-Laury, and all 140 subjects for the other four tasks. The lowest γ estimates come from the Holt-Laury and Budget Dots Holt-Laury tasks; apart from the 11 (resp. 3) subjects who always choose the safe option in Holt-Laury (resp. Budget Dots Holt-Laury), both cdf's are roughly uniformly distributed between 0 and 1.0. The Budget Line, Budget Jars and Budget Jars (no cash) regression estimates have higher medians and each has an upward skew.

Panel b of the same Figure paints a similar picture for the Relative Risk Premia $RRP_{i\tau}$ specified in Section 5.4. Despite a different range and different conceptual foundations than estimated risk aversion parameters, the RRPs have cumulative distribution functions with similar orderings. Holt-Laury and Budget Dots Holt-Laury reveal the least risk aversion while Budget Line, Budget Jars and Budget Jars (no cash) reveal the most, with Budget Dots Eckel-Grossman mostly in between.

Standard Kolmogorov-Smirnov (K-S) tests in Table 1 reject equivalence in distribution between most pairs of tasks in Fig. 6a.

Result 1. Different elicitation tasks reveal substantially different distributions of risk preferences.

Table 1
Kolmogorov-Smirnov test p-values for equality across task pairs of $\hat{\gamma}$ distributions. The row BDx refers to Budget Dots Holt-Laury in the Fix-Price Data and to Budget Dots Eckel-Grossman in the Fix-Prob data.

	Fix-Price				Fix-Prob			
	BDHL	BL	BJ	BJn	BDEG	BL	BJ	BJn
HL	0.225	0.000	0.000	0.000	0.000	0.001	0.000	0.00
BDx		0.000	0.000	0.000		0.019	0.002	0.000
BL			0.743	0.957			0.011	0.032
BJ				0.747				0.476

Table 2
Within-subject Spearman rank correlation of $\hat{\gamma}$'s. The row BDx refers to Budget Dots Holt-Laury in the Fix-Price Data and to Budget Dots Eckel-Grossman in the Fix-Prob data. () denote p-values, approximated via student's t-distribution (Zar, 1972).

	Fix-Price				Fix-Prob			
	BDHL	BL	BJ	BJn	BDEG	BL	BJ	BJn
HL	0.443 (0.000)	0.305 (0.010)	0.352 (0.002)	0.328 (0.006)	0.070 (0.604)	0.086 (0.500)	−0.033 (0.784)	0.212 (0.085)
BDx	-	0.566 (0.000)	0.526 (0.000)	0.493 (0.000)	-	0.618 (0.000)	0.612 (0.000)	0.623 (0.000)
BL	-	-	0.747 (0.000)	0.737 (0.000)	-	-	0.610 (0.000)	0.585 (0.000)
BJ	-	-	-	0.783 (0.000)	-	-	-	0.716 (0.000)

6.2. *Are individual subjects' choices consistent across elicitation tasks?*

Result 1 entails different revealed median (and mean) levels of risk preferences across elicitation tasks, and that undermines the ability of those tasks to predict behavior, such as insurance purchase, that depends on the actual level of risk aversion. However, the Result still leaves open the possibility that we can successfully predict some sorts of individual behavior using relative position. For example, suppose that an individual's risk preference measure is at the 15th percentile in one task. We could successfully predict her behavior in another context with known distribution of behavior if we could reliably say that she will again be near the 15th percentile. From this perspective, the crucial question is whether subjects' Spearman rank correlations across tasks approximate $\rho = 1.0$.

Since they employ precisely the same feasible set $F =$ a budget line with the same state probabilities, the Budget Line, Budget Jars, and Budget Jars (no cash) tasks are the same according to standard choice theory, while Budget Dots Eckel-Grossman can be thought of as a finite discrete version of Budget Line. In the same sense, the initial and final period Holt-Laury trials are identical to each other and to the $p = 0.81$ blocks of Budget Dots Holt-Laury. The expected utility hypothesis predicts identical rankings in all tasks, but the prediction seems especially compelling across those tasks deemed identical by standard choice theory.

Table 2 collects the rank correlations for the task-specific $\hat{\gamma}$ distributions. In the fixed price sessions we get impressively large values, around 0.75, among the continuous budget set tasks. However, the correlation between risk preference revealed in Holt-Laury and those in the continuous tasks is at best only around 0.35. Budget Dots Holt-Laury-elicited preferences are more highly correlated with those from continuous tasks (~ 0.5) than with preferences elicited from the deemed identical Holt-Laury task ($\rho = 0.44$). In the fixed probability sessions, the correlations among Budget Line, Budget Jars, Budget Jars (no cash) and Budget Dots Eckel-Grossman are between 0.58 and 0.72, but the correlation between any of them and Holt-Laury is near zero.

In a sense to be made more precise later, correlations between preferences across pairs of tasks deteriorate as the task attributes across pairs diverge. Starting in the bottom corner of the Fix-Price panel of Table 2, switching one of the tasks from continuous (BL) to discrete (Budget Dots Holt-Laury), correlations drop from 0.737 or 0.747 to 0.493 or 0.526. Correlations drop another $\sim$0.20 after switching from Budget Dots Holt-Laury (spatial) to Holt-Laury (same opportunities presented in text). Finally, when we switch from the left panel (fixed state price, varying state probability) to the right panel (fixed state probability, varying state price)), the correlation between Holt-Laury and the other tasks common to both panels (BL, Budget Jars, Budget Jars (no cash)) drops another $\sim$0.3, to near zero.

Table 3 shows task-pair correlations for the non-parametric measure RRP that roughly parallel those for the parametric measure $\hat{\gamma}$.[11] In the Fix-Price data, $\rho = 0.48$ for Holt-Laury and Budget Dots Holt-Laury; ρ's are roughly .7 to .8 among the

[11] Appendix B.6 reports cross- and own-correlations of task-specific $\hat{\gamma}$'s with RRPs. In the Fixed-Prob data, the correlation of $\hat{\gamma}$ with RRP is 0.86 for the Budget Jars data, and the other own-correlations range from 0.89 to 0.98. In the Fix-Price data, the Holt-Laury own correlations is .90 and the others are all at least 0.98.

Table 3
Within-subject Spearman rank correlation of Revealed Risk Premium.

	Fix-Price				Fix-Prob			
	BDHL	BL	BJ	BJn	BDEG	BL	BJ	BJn
HL	0.48	0.39	0.33	0.36	0.17	0.08	0.00	0.13
BDx		0.59	0.53	0.60		0.70	0.57	0.56
BL			0.72	0.83			0.67	0.49
BJ				0.79				0.78

Table 4
Task attributes. A ✓ in the column for a given task indicates that the attribute 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.

	HL	BDHL	BDEG	BL	BJ	BJn
Spatial	-	✓	✓	✓	-	-
2Dots	✓	✓	-	-	-	-
6Dots	-	-	✓	-	-	-
Cash	-	-	-	-	✓	-
FixProb	-	-	✓	*	*	*
Random	-	*	*	*	*	*
Px58	-	*	*	*	*	*
Pr65	-	-	*	*	*	*

continuous budget tasks, but continuous budget tasks' respective correlations with Budget Dots Holt-Laury are around .6 and are .3 - .4 with Holt-Laury. In the Fixed Probability data, correlations between Holt-Laury and anything else are again near zero.

Result 2. Individual subjects' revealed risk preferences are poorly correlated across tasks with dissimilar attributes, but are highly correlated across some task pairs with similar attributes.

Remark. Section 6.3 will formalize "similar" and "dissimilar" task attributes; for now those designations are informal and suggestive. In Section 6.5 we will see that our high correlations (observed for similar tasks) are actually higher than correlations obtained by Gillen et al. (2019) using the ORIV procedure, while our low correlations (observed for dissimilar tasks) are comparable to Gillen et al.'s lower correlations.

6.3. The impact of task attributes on revealed preference distributions

The preceding results suggest that task attributes — whether or not deemed relevant by standard decision theory — may affect subjects' revealed risk preferences. To assess that possibility more directly, we regress trial-by-trial revealed risk aversion, either $\check{\gamma}$ or $R\check{R}P$, on indicator variables flagging the presence or absence of the task attributes defined in Table 4. The design attribute Spatial refers to a budget line display in the 2 dimensional space of Arrow portfolios, either allowing choice anywhere on the line (in Budget Line) or on a finite subset of it (Budget Dots Holt-Laury and Budget Dots Eckel-Grossman). The design attribute 2Dots refers to tasks with only binary choices, either via radio buttons in text (Holt-Laury) or via two points in Arrow-Debreu 2-space (Budget Dots Holt-Laury). The design 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 Budget Jars) allowing retention of a perfectly hedged asset, cash. The environmental attributes are Fix[ed]Prob[ability] sessions (versus Fixed Price), HLprob (Holt-Laury trials administered to subjects in FixProb sessions), Random (versus monotone) ordering of task and price or probability sequences, whether the trial uses $p_x = 0.58$ in the Fix-Price data, and whether the trial uses $\pi_X = .65$ in the Fix-Prob data.

Table 5 reports the regression results. Controls include subject-level fixed effects as well as order (trial sequencing) indicator variables reported in Appendix B.7. 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., -0.287 for $\check{\gamma}$ in the fixed price environment. The results also suggest an overall tendency for subjects to under-respond to shifts in the choice environment. The coefficients on Pr65 and Px58 would be 0 if each subject made choices in Pr65 (resp. Px58) trials that were rationalizable by the same utility function as in her Pr50 (resp. Px81) trials, but instead these coefficients are large and significantly positive, indicating more risk-averse choices remaining closer to the perfect hedge, the diagonal $x = y$.

The Budget Dots Eckel-Grossman attribute of restricting choice to 6 dots on the same side of the diagonal increases the elicited $\check{\gamma}$ by 0.29, while any effect of the discrete restriction in Holt-Laury or Budget Dots Holt-Laury to 2 dots (one of which is near a corner) is more modest. Another intriguing finding is that the Random sequence attribute (whether in

Table 5
Full sample OLS coefficient estimates. Dependent variables are parametric and nonparametric single-trial risk preference elicitations ($\check{\gamma}$ and $R\check{R}P$) for tasks Budget Line, Budget Jars, Budget Jars (no cash) and Budget Dots Eckel-Grossman and individual estimates ($\hat{\gamma}$ and RRP) for tasks Holt-Laury and Budget Dots Holt-Laury. Regressions include subject level fixed effects with errors clustered at the subject level (s.e. in parentheses) and at the subject-task level [s.e. in brackets]. Asterisks *,** and *** respectively indicate p-values $<$ 0.10, 0.05 and 0.01. Regressions also include the order controls reported in Appendix B.7.

	$\check{\gamma}$	RRP
Spatial	−0.287 (0.087)*** [0.074]***	−0.033 (0.006)*** [0.006]***
Cash	0.034 (0.088) [0.077]	0.013 (0.007)* [0.006]**
2Dots	0.072 (0.099) [0.099]*	−0.114 (0.012)*** [0.010]***
6Dots	0.292 (0.074)*** [0.072]***	−0.037 (0.007)*** [0.007]***
Pr65	0.504 (0.066)*** [0.063]***	0.125 (0.008)*** [0.005]***
Px58	0.236 (0.063)*** [0.062]***	0.048 (0.005)*** [0.005]***
2Dots*Px58	−0.206 (0.082)** [0.102]***	0.018 (0.015) [0.014]
RandomFirst	0.211 (0.135) [0.126]*	0.010 (0.017) [0.013]
RandomSecond	0.198 (0.108)* [0.123]**	0.006 (0.015) [0.012]
HLprob	0.020 (0.174) [0.175]	0.003 (0.017) [0.017]
FixProb	−0.144 (0.140) [0.146]	−0.054 (0.018)* [0.014]***
FixProb*Spatial	0.125 (0.123) [0.105]	0.006 (0.012) [0.009]
FixProb*Cash	0.166 (0.140) [0.123]	−0.009 (0.010) [0.010]
FixProb:randomFirst	−0.109 (0.176) [0.161]	−0.021 (0.021) [0.016]
FixProb:randomSecond	−0.145 (0.157) [0.172]	0.002 (0.020) [0.015]
Observations	6,725	6,725
R^2	0.256	0.647

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. Note that $R^2 = 0.256$ for the gamma regression, comparable to the R^2 obtained by Gillen et al. using (unidentified) factors. For the RRP regression, R^2 is even higher, at .647.

Result 3. Much of the variation in location across task-specific distributions of elicited risk preferences can be explained by task attributes such as spatial presentation versus not, and restrictions on choice sets or by shifts in the choice set.

6.4. The impact of task attributes on correlations

Can differences in attributes also explain differences in rank correlation? To assess this possibility we partitioned our 140 subjects into 8 cohorts, each corresponding to one of the 8 branches in the second layer of Fig. 5. Thus all subjects in a given cohort face exactly the same sequence of trial blocks, random or monotone, with the same sequences of prices and probabilities. The dependent variable in each cohort is an exhaustive set of Spearman rank correlations, one for each pair of trials. Possibly the tasks are the same in the two trials, but in most pairs the tasks differ ("mismatch") in one or more attributes. The explanatory variables include indicators that flag a match (0) or mismatch (1) on a particular feature in the pair of trials.[12]

Table 6 reports regression results. Seven of eight cohorts exhibit negative entries for M-Spatial, some significant; thus rank correlation of elicited preferences across a pair of tasks is unlikely to be higher when one task is spatial and the other is text-based. Cash mismatch more often has a negative impact than a positive impact. Mismatches in the fixed price (.81 vs .58) or in the fixed probability (.50 vs .65) also tend to reduce correlations, often substantially, and the reduction is significant at the 1% level in half the cohorts. The indicator Random:either is a useful control whose sign varies across cohorts. In sum,

Result 4. Mismatches in spatial or cash attributes across pairs of tasks often lower the rank correlations of elicited risk preferences. Mismatches in parameterization do so even more reliably.

Discreteness of choice sets impacts correlations in an intriguing manner. A discrete-discrete pair will tend to have higher correlation than an otherwise similar discrete-continuous pair, which will in turn exhibit, all else equal, a higher correlation

[12] Specifically, M-spatial, M-cash, M-prob and M-price have value 1 if the tasks in the trial pair are mismatched respectively on the spatial/not attribute, the cash/not attribute, the fixed probability (0.5 or 0.65) and the fixed price (0.81 or 0.58), and otherwise are zero. For example, a trial pairing with tasks Budget Jars (no cash) and Budget Line would have coding M-spatial=1 (since Budget Jars (no cash) is text-based while Budget Line is represented in 2-dimensional space) and M-cash=0 (since neither allows residual cash holdings). In this example, M-prob is constructed only for FixProb cohorts, and there depends on the specific trials being compared; and likewise M-price is constructed only for FixPrice cohorts, where again its value depends on the specific pair of trials. The indicator DD takes on the value 1 if both tasks in a trial pair have a discrete choice set (0 otherwise), while the indicator CD takes on the value 1 if only one of the two tasks has a discrete choice set; the omitted (reference) indicator is CC, comprising the cases where both tasks are continuous. The Random:either indicator is set at 1 if either trial in a pair was generated in a block of trials which was ordered randomly.

Table 6
Impact of attribute mismatches on correlations. Independent variable is the rank correlation of $\hat{\gamma}$'s (or $\hat{\gamma}$ for Holt-Laury and Budget Dots Holt-Laury) across trial pairs within the given cohort. Mismatch variables M-Z indicate mismatch in attribute Z in the trial pair. CD and DD are dummies for whether either institution has a continuous vs discrete choice set; the holdout dummy is CC where both trials have continuous choice sets. Column labels identify the cohort treatment and cohort size for the dependent variable, 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.

	Fix Prob				Fix Price			
cohort N	LE19	EE13	LL23	EL15	EE25	LL18	LE17	EL12
M-spatial	−0.007	−0.018	0.006	−0.020	−0.059**	0.008	−0.031	−0.019
	(0.020)	(0.020)	(0.013)	(0.024)	(0.024)	(0.013)	(0.025)	(0.019)
M-cash	0.050**	−0.043	−0.086***	−0.029	0.023**	−0.105***	0.020	−0.018
	(0.022)	(0.027)	(0.016)	(0.023)	(0.011)	(0.009)	(0.014)	(0.025)
M-prob	−0.055***	−0.017*	0.012	−0.139***				
	(0.016)	(0.009)	(0.016)	(0.027)				
M-price					0.007	0.015	−0.076***	−0.086***
					(0.010)	(0.014)	(0.016)	(0.021)
CD	0.034	−0.009	0.029*	0.015	0.011	0.076***	0.084***	0.067
	(0.021)	(0.013)	(0.015)	(0.025)	(0.013)	(0.018)	(0.023)	(0.041)
DD	0.106**	0.038	0.151***	0.064	0.515***	0.448***	0.375**	0.550***
	(0.044)	(0.032)	(0.015)	(0.090)	(0.145)	(0.149)	(0.149)	(0.096)
random1	0.084***	−0.150***	−0.136***	0.223***	0.002	−0.063***	0.132***	0.128***
	(0.021)	(0.034)	(0.011)	(0.030)	(0.015)	(0.019)	(0.024)	(0.018)
Constant	0.128***	0.297***	0.271***	0.038	0.087***	0.189***	0.210***	0.194***
	(0.021)	(0.051)	(0.009)	(0.029)	(0.026)	(0.012)	(0.027)	(0.008)
Observations	1,225	1,225	1,225	1,225	990	990	990	990
R^2	0.031	0.077	0.127	0.072	0.042	0.080	0.052	0.050

Note: *p<0.1; **p<0.05; ***p<0.01.

than a continuous-continuous pair (the omitted dummy in the regression). The DD (both tasks discrete) coefficients are all positive, the majority of them significantly so at the 5% level. The increase is impressively large, around 0.375 to 0.55, in the Fix Price cohorts (which pertain to Holt-Laury and Budget Dots Holt-Laury but not to Budget Dots Eckel-Grossman). The CD coefficients are also mostly positive but smaller; relative to the omitted case CC, it seems that having discrete feasible sets in either task tends to increase correlations. The estimates suggest that a discrete-discrete pairing such as Holt-Laury-Holt-Laury (i.e. an Holt-Laury re-test) would have 0.375 to 0.55 of its fitted rank correlation of 0.6 to 0.75 attributable to the discreteness of its choice space. Thus

Result 5. Single-trial correlations involving discrete tasks tend, other things equal, to be higher than single trial correlations among tasks with continuous choice sets.

Our interpretation is that estimated correlations between single trials, as in Table 8, are artificially inflated for coarse discrete tasks, due to the limited number of possible responses in those tasks, and consequent increased chance of identical choices. Note the contrast to Table 2, where task-specific correlations are higher among continuous tasks. Correlations between γ's estimated by regressions on repeatedly sampled responses, as in Table 2, are more reflective of subject behavior and not simply the granularity of the choice space.

6.5. Robustness and noise

Traditional extraction of γ. Recall that Table 2 is based on risk preference parameter estimates γ extracted using our regression methods presented in Section 5. Appendix A.4 notes that the regressions give greater weight to observations where the price/probability control variable L_t is larger in absolute value, and argues that the rank ordering of estimated gammas is less responsive to behavioral noise when $|L_t|$ is large. For that reason, we believe that the rank correlations reported in Table 2 are more reliable than those using traditional extraction methods, which equally weight observations irrespective of $|L_t|$.

For comparative purposes Table 7 reports Spearman rank correlation coefficients based on the traditional extraction method. Comparing the left panel of Table 7 to its Table 2 counterpart shows that where estimated correlations differ between the two tables, they are lower than their regression-based counterparts. The right panel of Table 7 has low correlations in the first row, similar to their counterparts, while the bottom row entries are lower than their Table 2 counterparts.

The upshot is that Result 2 above would only be strengthened if we were to replace our regression-based methods by more traditional (but, we believe, less reliable) methods to extract γ. In using methods which generate higher cross-task

Table 7
Within-subject Spearman rank correlations. Holt-Laury and BDx γ's obtained via traditional calculations. All subjects are included for whom the traditional calculation is ever feasible. Other tasks are as in Table 2.

	Fix Price				Fix Prob			
	BDHL	BL	BJ	BJn	BDEG	BL	BJ	BJn
HL	0.29	0.24	0.32	0.27	0.19	0.04	0.06	0.20
BDx	-	0.50	0.47	0.47	-	0.47	0.45	0.48

Table 8
Within-subject Pearson correlations of γ. Left panel shows results presented in GSY, before ("Raw") and after applying the ORIV procedure. Middle panel shows results after fixing a coding error and removing censoring from the GSY data. Right panel shows results using a corresponding subset of our data (two elicitation rounds each of Budget Jars and Holt-Laury and one elicitation of Budget Dots Eckel-Grossman) as explained footnote 13.

	GSY			GSY (minus censoring & miscoding)			Our Data		
		Project	Lottery		Project	Lottery		BJ	BDEG
Raw	Lottery	0.27		Lottery	0.27		BDEG	0.23	
	MPL	0.18	0.22	MPL	0.08	0.11	HL	−0.17	−0.02
		Project	Lottery		Project	Lottery		BJ	BDEG
ORIV	Lottery	0.55		Lottery	0.55		BDEG	0.62	
	MPL	0.37	0.42	MPL	0.14	0.15	HL	−0.35	−0.03

correlations, we have given EUT its best shot at predictive success across tasks. Thus, our conclusion that correlations decay as tasks become less similar is robust with respect to γ -extraction method.

Comparison to ORIV. Some readers might wonder how the results in this present study relate to those of Gillen et al. (2019), denoted GSY below. Their results and ours complement each other in ways that create insight into best practices in experimental design, and into the impact of attenuation bias (first addressed by Spearman (1904)).

GSY emphasize that measurement error should be taken into account when interpreting experimental data, and apply a procedure called ORIV to estimating Pearson correlations among three quantitative risk preference elicitation tasks. The tasks in the GSY study — Lottery, MPL and Project — correspond respectively to our Budget Dots Eckel-Grossman, Holt-Laury and Budget Jars. The GSY data consist entirely of subject responses, i.e., of dependent variables, with two observations each in MPL and Project tasks and a single observation in the Lottery task. This creates the need for ORIV, an instrumental variable estimation technique which aims to redress that downward bias in coefficient estimates which is due to regressing dependent variables on other dependent variables. The ORIV procedure produces population-level correlations, attenuation-corrected for the measurement error in GSY's right-hand side variables.

In contrast, exogenous variation in prices and probabilities in our experiment allows us to construct a classical explanatory variable L. That variable is controlled experimentally, and as such is without measurement error. Any measurement error in our data is confined to the dependent variable, so our individual-level risk parameter estimates are consistent and unbiased.

To illustrate the application of ORIV to our data, we select a subset of our trials that most closely matches the parameterizations in the GSY study,[13] and (as in GSY) we use traditional calculations to extract a γ from each trial. As explained in Friedman et al. (2019), there is a minor error in the code used to generate the results reported in GSY, and they also censor the MPL data by reclassifying all risk-seeking choices as risk-neutral. Table 8 presents the GSY results with and without the coding error and censoring, as well as the results of ORIV applied to the most relevant subset of our own data.

We draw three lessons from this exercise. First, applying ORIV to a subset of our data and to GSY's data (minus censoring and mis-scaling) reveals a common pattern in correlations: relatively high correlation between γ's elicited from the Budget Jars and the Budget Dots Eckel-Grossman tasks (or their analogues) and rather low correlations between either Budget Jars or Budget Dots Eckel-Grossman and Holt-Laury (or their analogues). Recall that Budget Jars and Budget Dots Eckel-Grossman are the same task, barring discretization and visualization. Second, the correlation obtained by use of ORIV and endogenous regressors need not be higher than its analogue using our combination of exogenous regressors, simple regression, and Spearman rank correlation of the resulting estimates; for example, in Table 2 we obtain a correlation of 0.612 between Budget Dots Eckel-Grossman and Budget Jars (higher than the 0.55 obtained via ORIV by Gillen et al.). The task pairs

[13] GSY vary the parametrization of their two trials of Project (Gneezy-Potters) as follows. Trial 1 has 0.4 chance of high state, and an implied budget line slope of 0.5 (since cash — the bundle (1, 1) — trades for triple payout in state X, i.e., for the bundle (3,0), so $-\Delta y/\Delta x = 0.5$). Trial 2 has 0.5 chance of high state, and an implied budget line slope of 2/3. Thus for Budget Jars we select our two trials where (a) probability of x state is 0.65 and slope is 0.58 and (b) probability of x state is 0.5 and slope is 0.23. This allows us to move the probability of the x-state and relative price of the x-security in offsetting directions, across the two selected trials, following GSY. For Budget Dots Eckel-Grossman we directly match the state probabilities and implied budget line slope for GSY's single observation of Lottery: GSY set the state probability at .5 and the implied slope at .62, so we select the Budget Dots Eckel-Grossman trial with state probability .5 and price .58. We used the standard parameterization in both of our two Holt-Laury trials.

exhibiting the lower ORIV-estimated correlations in Table 8 yield comparable correlations in the top row, right panel of Table 2. Third, the righthand panel of Table 8 reminds us that applying ORIV does not always move correlations closer to 1; as seen here, a negative raw correlation can be amplified by ORIV.

These lessons have implications for experimental design and data analysis. If the goal of a study is to obtain the highest possible correlations between different tasks, then we would recommend a design using exogenous regressors, together with sufficient repetition to allow estimation at the individual level. As for which correlation to emphasize, we, along with many other recent authors, prefer the Spearman rank to the Pearson correlation when comparing elicitation tasks with different ranges (or shapes). For example, GSY justify censoring negative MPL estimates by noting that some of their other tasks do not permit negative estimates of γ. The issue of differing ranges (or shapes, e.g., skewness) is hard to resolve cleanly for Pearson correlations but does not arise for rank correlations.

There are also insights into the construction and interpretation of attenuation corrections. The underlying problem that ORIV is intended to correct is that correlation estimates are downward biased when right-hand-side variables are measured with error. In Appendix A.2 we show how to construct an attenuation adjustment for Pearson correlations for our γ's directly, without using ORIV. It turns out that the "true" Pearson correlation between tasks A and B is $\rho_{AB} = \rho_{\hat{A}\hat{B}}\,[R_A R_B]^{-0.5}$, where $\rho_{\hat{A}\hat{B}}$ is the raw correlation between the noisy observed γ's in the two tasks, and the attenuation correction factor $[R_A R_B]^{-0.5} \geq 1$ is defined in terms of reliability ratios that we can implement using standard errors of regressions and cross-sectional variances of the γ's. Appendix B.8 applies that formula to our data and obtains estimates that generally exceed 1, so the maximum likelihood estimates are 1.0 — seemingly a much stronger result than that reported in the righthand panel of Table 8. It seems to us that such an attenuation correction is more defensible in addressing the hypothesis that correlations are 1.0 than in providing superior point estimates of true correlations.

The underlying question is: when are attenuation corrections appropriate? They may not helpful for the practical goal of predicting behavior. (See Fuller (1987) Chapter 1 for a nuanced discussion.) On the other hand, an attenuation correction is appropriate for assessing more philosophical questions about the relation between unobservable latent variables. In particular, it can be appropriate to use such corrections to test whether, as assumed in the Expected Utility Hypothesis, each individual subject indeed has a stable personal Bernoulli function, but perhaps responds to each elicitation task with more or less noise.

Simulation exercise. Although prediction is the main goal of the current paper, for the rest of this subsection we take up the philosophical question just raised. Are our data consistent with human subjects each maintaining a stable personal Bernoulli function, but responding with task-specific noise? Attenuation bias corrections may provide indirect evidence, but now we take a more direct approach, and report simulations of noisy choice using flexible extensions of the random coefficient models of Wilcox (2008) and Apesteguia and Ballester (2018).

Each run in our simulation consists of a set of automated agents, each of them making the same 66 lottery choices as a unique human subject in our experiment. As in a random coefficient model, each agent is assigned a "true" value of γ, which we set equal to an overall γ estimated for its human counterpart. We extend the random coefficient model to allow the "true" γ to be task-specific, but to occupy the same percentile within the population distribution in each task. For each lottery choice faced by its human counterpart, the automated agent draws independently a noisy version of its task-specific "true" γ, perturbed by task-specific noise, and chooses an action that maximizes expected utility with respect to that perturbed γ. Thus each run of the simulation produces a data set parallel to our actual choice data, and a thousand such runs (each with its own realized perturbations) give us a distribution of correlations in which can place our human subjects' actual correlations.

The results are striking. For the most plausible amplitudes of task-specific noise, the correlations between Budget Dots Holt-Laury and other tasks in the human data fall nicely between the tails of the distribution of simulated correlations. However, the other human-data correlations are extreme outliers. Human subjects' correlations between continuous choice-set tasks (Budget Lines and both variants of Budget Jars) are far too high, and their correlations between traditional Holt-Laury and other tasks are often far too low to arise from the random coefficients simulations. See Appendix A.3 for further discussion of the simulations, including a detailed step-by-step presentation of the algorithms that define the automated agents.

Result 6. The observed variation in correlations of revealed preferences across elicitation tasks can not be explained by a flexible random coefficient model of expected utility maximization.

An underlying reason for this negative result is that our simulated agents (and indeed, any agents rooted in standard choice theory) respond only to changes in the feasible set (e.g., prices and probabilities) and not to changes in the way the feasible set is presented (e.g., in text vs spatially). In contrast, as we saw in Sections 6.3 and 6.4, humans can and do respond to task attributes not recognized by standard choice theory.

7. Discussion

Subjects in our experiment respond in the expected direction to changes in Arrow lottery prices and probabilities, but the *degree* of responsiveness — which economists encapsulate as revealed risk preference — is inconsistent across elicitation

tasks. As we move from one elicitation task to another, the population distribution of revealed risk preferences shifts and changes shape (Result 1). Perhaps more importantly, correlations of revealed risk preferences across widely used elicitation tasks are quite small, indicating little predictive power (Result 2). These findings are robust to alternative specifications and seem not to be due simply to task-specific noise (Result 6).

A major take-away message of our study is that there is some regularity to the inconsistencies. Task attributes — the way choices are presented and responses are entered, and the way choices are parametrized — have a substantial impact, whether or not the attributes have a role in standard economic theory (Result 3). The power of revealed preferences to predict behavior in a new setting improves when elicitation task attributes better match the attributes of that setting (Results 4, 5).

How might our results inform applied economic work? Many sorts of applications involve individual decision making under risk, and investigators may need to control for individual risk preferences. To illustrate the implications of our results, we now consider two specific applications.

Development. For many decades, development economists have studied the relationship between risk preferences and take-up decisions regarding technology or credit; such work goes back at least to Moscardi and De Janvry (1977), Dillon and Scandizzo (1978) and Binswanger (1980). Standard current practice still is to ask respondents to choose a preferred lottery from a short menu as in a single instance of the Eckel-Grossman task.

Our Results 3, 4 and 5 suggest possible improvements by better matching of attributes across control task and ability to be predicted. Suppose, for example, the investigator wants to predict a binary choice such as building or not building stone terraces to reduce soil erosion as in Teklewold and Köhlin (2011), in the presence or absence of subsides, and wants to control for individual risk preferences. Then it may help for the risk preference elicitation task also to involve binary choice (between lotteries). One could also match the representation of payoffs in each state (e.g., crop yields following heavy or light rains) as pictograms (e.g., more or fewer bundles of wheat) as well as numerical values. Such matching should improve correlation between the control task and field scenarios, thereby enabling sharper estimates of policy impact.

A second suggestion is for investigators, to the extent possible, to include multiple trials with varying prices or probabilities. A single multiple price list trial might take about the same time as two or three Eckel-Grossman (or Binswanger) style trials or as five or ten Hey-Orme style binary lottery trials. A regression using just a dozen well chosen binary lottery trials might sharpen the estimate of a respondent's revealed risk aversion, as noted in Sections 5 and 6 and Appendix B; estimation would be further enhanced with a larger, but still feasible, number of trials. Such regressions should again increase the usefulness of revealed risk preference as a right hand side control variable.

Investing. Brokerages and asset management firms typically ask customers to fill out questionnaires intended to assess risk preferences or "risk tolerance;" see Appendix E for an example in the public domain. Also, online brokers now compete to offer customers better visualizations of historical return distributions as well as 2D options strategy payoff diagrams; again see Appendix E for examples. Better assessment of a given client's attitudes on financial risk-taking surely will help the firm's risk management teams, margin desks, and client relations.

Our results suggest several possible improvements to current practice. First, as exemplified in Appendix E, typical assessment questionnaires for passive retail investors currently have a discrete multiple choice format. Since typical choice tasks for these customers (e.g. adjusting the balance in a portfolio between an index fund and Treasury bills) are continuous, our Result 5 suggests that the assessment should also be continuous. A second suggestion for risk tolerance assessment is that the choices considered include an all-cash position, but not start with an all-cash default. Our results suggest that people do behave differently when a cash default is present,[14] and positioning the respondent in 100% cash at the start of a task creates an artificial and avoidable mismatch with the outside world. A third suggestion is to use a spatial interface, at least when focused on long-term financial goals; Table 7 suggests that this will reduce possible bias towards overly risk averse revealed preferences.

Our results also have implications for active investment platforms that want to screen their customers, or just to prepare them properly. (A platform may wish to do so to reduce the number of client insolvencies, or to encourage customers to make more frequent use of the platform, among other reasons.) If so, then screening or training modules utilizing continuous choice sets and spatial displays would be helpful for reasons just noted. Spatial display of trading information is now standard for options trading (see Appendix E).

Within the set of tasks we have investigated, our suggestions correspond most closely to Budget Lines, which are continuous and spatial, and include the all-cash portfolio but do not initialize there. A similar task for investment management clients could be implemented via user-friendly displays that, at each slider position, show the relative frequencies of meeting or not meeting a client-specified financial target or, alternatively, perhaps show a histogram of projected returns.

A caveat is in order. Preferences (and thus preferred financial portfolios) will change over time as life circumstances change; a new assessment may be in order given changes in family size, in financial obligations (e.g., a mortgage) or in family obligations, for example. Chapter 7 of Friedman et al. (2014) includes a lengthy discussion of related issues. Fortunately, as illustrated in Appendix E, many risk tolerance assessments currently in use in the field do attempt to capture this kind of information.

[14] That a default setting would have an impact is consistent with findings from organ donation (Johnson and Goldstein (2003)) and from savings programs (Beshears et al. (2009)).

Broader Perspectives. Our recommendations concerning applied work are based on the following *a fortiori* argument. Our results demonstrate that, even in a tightly controlled laboratory environment, mismatches between tasks can degrade correlations. A reasonable conjecture is that uncontrolled factors in field settings might arise and degrade them even further. Thus, applied researchers should avoid unnecessary mismatches in design, and might regard the sort of correlations we find in our lab data as upper bounds on the correlations they can expect from field data.

On the question of whether a typical person has unified risk preferences, we would view our evidence as more negative than positive. Our simulation results suggest that even perturbations, as per Apesteguia and Ballester, to fixed underlying risk preferences cannot explain the pattern of correlations we observe across tasks. Furthermore, the location of revealed risk preferences can be shifted via exogenous controls – the task attributes – regardless of attenuation corrections and their interpretations.

What might be the underlying causes of sensitivity to task attributes not recognized by standard economic choice theory? As noted in Section 2, decades of work in perceptual psychology speak to that question. That literature reports that similar information delivered by different sensory channels can invoke different sorts of information processing and result in different sorts of behavior. Our findings suggest that the same may be true in the domain of risky choice.

Rather than viewing our results as pessimistic with respect to existing decision theories, we hope that our findings will inspire decision theorists to build new sorts of models. Continuing work on capturing decision processes (see Schram and Ule (2019) for examples) and in axiomatizing process-based models (Blavatskyy (2014) provides an example) may point the way towards better prediction of choice under risk.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.geb.2022.02.002.

References

Andreoni, James, Harbaugh, William T., 2009. Unexpected Utility: Experimental Tests and Five Key Questions About Preferences over Risk.

Andreoni, James, Kuhn, Michael A., Sprenger, Charles, 2015. Measuring time preferences: a comparison of experimental methods. J. Econ. Behav. Organ. 116, 451–464.

Apesteguia, Jose, Ballester, Miguel A., 2018. Monotone stochastic choice models: the case of risk and time preferences. J. Polit. Econ. 126 (1), 74–106.

Becker, Gordon M., DeGroot, Morris H., Marschak, Jacob, 1964. Measuring utility by a single-response sequential method. Behav. Sci. 9 (3), 226–232.

Berg, Joyce, Dickhaut, John, McCabe, Kevin, 2005. Risk preference instability across institutions: a dilemma. Proc. Natl. Acad. Sci. USA 102 (11), 4209–4214.

Beshears, John, Choi, James J, Laibson, David, Madrian, Brigitte C, 2009. The importance of default options for retirement saving outcomes: Evidence from the United States. In: Social security policy in a changing environment. University of Chicago Press, pp. 167–195.

Binswanger, Hans P., 1980. Attitudes toward risk: experimental measurement in rural India. Am. J. Agric. Econ. 62 (3), 395–407.

Blavatskyy, Pavlo R., 2014. Stronger utility. Theory Decis. 76 (2), 265–286.

Charness, Gary, Eckel, Catherine, Gneezy, Uri, Kajackaite, Agne, 2018. Complexity in risk elicitation May affect the conclusions: a demonstration using gender differences. J. Risk Uncertain. 56 (1), 1–17.

Charness, Gary, Garcia, Thomas, Offerman, Theo, Villeval, Marie Claire, 2020. Do measures of risk attitude in the laboratory predict behavior under risk in and outside of the laboratory? J. Risk Uncertain. 60 (2), 99–123.

Choi, Syngjoo, Fisman, Raymond, Gale, Douglas, Kariv, Shachar, 2007. Consistency and heterogeneity of individual behavior under uncertainty. Am. Econ. Rev. 97 (5), 1921–1938.

Collins, Sean M., James, Duncan, 2015. Response mode and stochastic choice together explain preference reversals. Quant. Econ. 6, 825–856.

Corsi, P., 1972. Memory and the medial temporal region of the brain. Unpublished doctoral dissertation. McGill University, Montreal, QB.

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

Dave, Chetan, Eckel, Catherine C., Johnson, Cathleen A., Rojas, Christian, 2010. Eliciting risk preferences: when is simple better? J. Risk Uncertain. 41 (3), 219–243.

Deck, Cary, Lee, Jungmin, Reyes, Javier A., Rosen, Christopher C., 2013. A failed attempt to explain within subject variation in risk taking behavior using domain specific risk attitudes. J. Econ. Behav. Organ. 87, 1–24.

Dehaene, Stanislas, Cohen, Laurent, 1991. Two mental calculation systems: a case study of severe acalculia with preserved approximation. Neuropsychologia 29 (11), 1045–1074.

Donolato, Enrica, Giofrè, David, Mammarella, Irene C., 2017. Differences in verbal and visuospatial forward and backward order recall: a review of the literature. Front. Psychol. 8, 663.

Dillon, John L., Scandizzo, Pasquale L., 1978. Risk attitudes of subsistence farmers in Northeast Brazil: A sampling approach. Amer. J. Agr. Econ. 60 (3), 425–435.

Eckel, Catherine, Grossman, Philip J., 2002. Sex differences and statistical stereotyping in attitudes toward financial risk. Evol. Hum. Behav. 23 (4), 281–295.

Eckel, Catherine, Grossman, Philip J., 2008. Forecasting risk attitudes: an experimental study using actual and forecast gamble choices. J. Econ. Behav. Organ. 68 (1), 1–17.

Fornaciai, Michele, Brannon, Elizabeth M., Woldorff, Marty G., Park, Joonkoo, 2017. Numerosity processing in early visual cortex. NeuroImage 157, 429–438.

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

Friedman, Daniel, Habib, Sameh, James, Duncan, Williams, Brett, 2019. Experimenting with Measurement Error: Comment.

Fuller, Wayne A., 1987. Measurement error models. In: Wiley Series in Probability and Mathematical Statistics. New York: Wiley, 1987.

Gillen, Ben, Snowberg, Erik, Yariv, Leeat, 2019. Experimenting with measurement error: techniques with applications to the caltech cohort study. J. Polit. Econ. 127 (4), 1826–1863.

Gneezy, Uri, Potters, Jan, 1997. An experiment on risk taking and evaluation periods. Q. J. Econ. 112 (2), 631–645.

Greene, William H., 2012. Econometric Analysis, 7th ed. Prentice Hall.

Grether, David M., Plott, Charles R., 1979. Economic theory of choice and the preference reversal phenomenon. Am. Econ. Rev. 69 (4), 623–638.

Habib, Sameh, Friedman, Daniel, Crockett, Sean, James, Duncan, 2017. Payoff and presentation modulation of elicited risk preferences in MPLs. J. Econ. Sci. Assoc. 3 (2), 183–194.

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

Hebb, Donald Olding, 1961. Distinctive features of learning in the higher animal. In: Brain Mechanisms and Learning, vol. 37, p. 46.
Heufer, Jan, 2014. Nonparametric comparative revealed risk aversion. J. Econ. Theory 153, 569–616.
Hey, John, Orme, Chris, 1994. Investigating generalizations of expected utility theory using experimental data. Econometrica 62 (6), 1291–1326.
Holt, Charles A., Laury, Susan K., 2002. Risk aversion and incentive effects. Am. Econ. Rev. 92 (5), 1644–1655.
Isaac, R. Mark, James, Duncan, 2000. Just who are you calling risk averse? J. Risk Uncertain. 20 (2), 177–187.
Jevons, W. Stanley, 1871. The Power of Numerical Discrimination.
Johnson, Eric J., Goldstein, Daniel, 2003. Do Defaults Save Lives?. American Association for the Advancement of Science.
Lévy-Garboua, Louis, Maafi, Hela, Masclet, David, Terracol, Antoine, 2012. Risk aversion and framing effects. Exp. Econ. 15 (1), 128–144.
Lichtenstein, Sarah, Slovic, Paul, 1971. Reversals of preference between bids and choices in gambling decisions. J. Exp. Psychol. 89 (1), 46–55.
Lichtenstein, Sarah, Slovic, Paul, 1973. Response-induced reversals of preference in gambling: an extended replication in Las Vegas. J. Exp. Psychol. 101 (1), 16–20.
Loomes, Graham, Pogrebna, Ganna, 2014. Measuring individual risk attitudes when preferences are imprecise. Econ. J. 124 (576), 569–593.
Moscardi, Edgardo, De Janvry, Alain, 1977. Attitudes toward risk among peasants: an econometric approach. Amer. J. Agr. Econ. 59 (4), 710–716.
Pedroni, Andreas, Frey, Renato, Bruhin, Adrian, Dutilh, Gilles, Hertwig, Ralph, Rieskamp, Jörg, 2017. The risk elicitation puzzle. Nat. Hum. Behav. 1, 803–809.
Ross, John, 2003. Visual discrimination of number without counting. Perception 32 (7), 867–870.
Schram, Arthur, Ule, Aljaž, 2019. Handbook of Research Methods and Applications in Experimental Economics.
Slovic, Paul, 1962. Convergent validation of risk taking measures. J. Abnorm. Soc. Psychol. 65 (1), 68.
Slovic, Paul, 1975. Choice between equally valued alternatives. J. Exp. Psychol. Hum. Percept. Perform. 1 (3), 280.
Slovic, Paul, Fischoff, Baruch, Lichtenstein, Sarah, 1988. Response Mode, Framing, and Information-processing Effects in Risk Assessment. In: Decision Making: Descriptive, Normative, and Prescriptive Interactions. Bell, David E., Raiffa, Howard, Tversky, Amos (Eds.). Cambridge University Press, pp. 152–166.
Spearman, Charles, 1904. "General Intelligence," Objectively Determined and Measured. The American Journal of Psychology 15 (2), 201–292.
Sprenger, Charles, 2015. An endowment effect for risk: experimental tests of stochastic reference points. J. Polit. Econ. 123 (6), 1456–1499.
Teklewold, Hailemariam, Köhlin, Gunnar, 2011. Risk preferences as determinants of soil conservation decisions in Ethiopia. J. Soil Water Conserv. 66 (2), 87–96.
Trautmann, Stefan T., van de Kuilen, Gijs, 2012. Prospect theory or construal level theory? Diminishing sensitivity vs psychological distance in risky decisions. Acta Psychol. 139 (1), 254–260.
Van Rinsveld, Amandine, Guillaume, Mathieu, Kohler, Peter J., Schiltz, Christine, Gevers, Wim, Content, Alain, 2020. The neural signature of numerosity by separating numerical and continuous magnitude extraction in visual cortex with frequency-tagged EEG. Proc. Natl. Acad. Sci. USA 117 (11), 5726–5732.
Wilcox, Nathaniel T., 2008. Stochastic models for binary discrete choice under risk: a critical primer and econometric comparison. In: Risk Aversion in Experiments. Emerald Group Publishing Limited, pp. 197–292.
Wilcox, Nathaniel T., 2011. 'Stochastically more risk averse:' a contextual theory of stochastic discrete choice under risk. J. Econom. 162 (1), 89–104.
Williams, Brett, 2021. Violations of First Order Stochastic Dominance.
Williams, Brett, Habib, Sameh, 2021. A Note on Disappointment Aversion in Risk Elicitation Tasks.
Zar, Jerrold H., 1972. Significance testing of the Spearman rank correlation coefficient. J. Am. Stat. Assoc. 67 (339), 578–580.
Zhou, Wenting, Hey, John, 2018. Context matters. Exp. Econ. 21 (4), 723–756.