Distinguishing Common Ratio Preferences from Common Ratio Effects Using Paired Valuation Tasks*

Christina McGranaghan Kirby Nielsen Ted O'Donoghue
Jason Somerville Charles D. Sprenger

September 24, 2023

Abstract

Without strong assumptions about how noise manifests in choices, we can infer little from existing empirical observations of the common ratio effect (CRE) about whether there exists an underlying common ratio preference (CRP). We propose to solve this inferential challenge using paired valuations, which yield valid inference under common assumptions. Using this approach in an online experiment with 900 participants, we find no evidence of a systematic CRP. To reconcile our findings with existing evidence, we present the same participants with paired choice tasks and demonstrate how noise can generate a CRE even for individuals without an associated CRP.

*McGranaghan: Department of Applied Economics and Statistics, University of Delaware, email: cmcgran@udel.edu; Nielsen: Division of the Humanities and Social Sciences, California Institute of Technology, email: kirby@caltech.edu; O'Donoghue: Department of Economics, Cornell University, email: edo1@cornell.edu; Somerville: Federal Reserve Bank of New York, email: jason.somerville@ny.frb.org; Sprenger: Division of the Humanities and Social Sciences, California Institute of Technology, email: sprenger@caltech.edu. For helpful comments and suggestions, we thank Ori Heffetz and Alex Rees-Jones, and seminar participants at the University of Michigan, Brown University, Loyola Marymount University, Claremont Graduate University, the University of California at Berkeley, MidExLab, briq, New York University, the University of California at Santa Cruz, Purdue University, University of Chicago, Penn State University, the 2022 Behavioral Economics Annual Meeting, the 2022 Stanford Institute for Theoretical Economics (Experiment Economics) Conference, and the 2023 East Coast Behavioral and Experimental Workshop. The views expressed in this paper are those of the authors alone and do not necessarily reflect the views Federal Reserve Bank of New York or the Federal Reserve System. The experiment reported in this paper was preregistered in the AEA RCT Registry in August 2021, under the ID AEARCTR-0008058. The experiment was reviewed and granted an exemption by the Institutional Review Board at the California Institute of Technology under protocol number 21-1073.

1 Introduction

The common ratio effect (CRE) refers to an empirical observation that, when choosing between a smaller amount that is more likely and a larger amount that is less likely, scaling down the probabilities by a common ratio makes people more prone to choose the riskier option. Allais (1953) first proposed the CRE as a plausible hypothetical counterexample to expected utility (EU). Later, Kahneman and Tversky (1979) provided experimental evidence of the CRE and used it to motivate one of the key features of their probability weighting function: "subproportionality." An extensive subsequent experimental literature has provided further empirical evidence of the CRE. Based on this evidence, the CRE is now commonly invoked as a violation of EU, and being able to explain it is a frequent litmus test for new models of choice under risk (for instance, see Loomes and Sugden, 1982; Gul, 1991; Bordalo et al., 2012; Cerreia-Vioglio et al., 2015).

There is an important challenge, however, to the standard interpretation of the CRE as a manifestation of non-EU preferences: If choices are stochastic rather than deterministic, then inference from paired choice tasks becomes problematic. Prior researchers have noted that EU with i.i.d. additive utility noise can naturally generate a CRE in paired choice tasks (Ballinger and Wilcox, 1997; Loomes, 2005; Hey, 2005; Wilcox, 2008; Blavatskyy, 2007, 2010; Bhatia and Loomes, 2017). Building on this literature, we demonstrate that, without strong assumptions about how noise manifests in choices, we can infer little from the existence or absence of a CRE in paired choice tasks about whether there exists an underlying *common ratio preference* (CRP).[1]

This paper proposes a solution to this inference challenge using *paired valuation tasks* instead of *paired choice tasks*. We first demonstrate theoretically that, under the same common assumptions about the structure of noise, paired choice tasks can yield a biased test of the EU null, whereas paired valuation tasks can yield unbiased tests. We then implement this paired-valuation approach in an experiment with 900 participants. We find no evidence of a systematic CRP at the aggregate level, although we observe substantial heterogeneity. To reconcile our findings with existing CRE evidence that uses paired choice tasks, we present the same participants with standard paired choice tasks. Individual heterogeneity in CRP, as measured by paired valuations, is highly predictive of whether an individual exhibits a CRE in paired choices. In addition, we demonstrate how appropriately chosen experimental parameters can generate a CRE in paired choice tasks even for individuals without an associated CRP (the same is true for *reverse* common ratio effects (RCRE) and preferences (RCRP)).

In Section 2, we develop theoretical results for testing the EU null—or, more generally, the null of no CRP—using paired choice tasks and paired valuation tasks when both are subject to noise. Importantly, we assume that behavior results from the same underlying preference and noise

[1] Throughout, we reserve the term "CRE" for empirical observations in experimental data, and we introduce the term "CRP" to refer to the underlying preference that these empirical observations are often taken to reveal. Our focus is whether one can empirically identify the existence of a CRP without making any assumptions about its source.

structure across both types of tasks, so different assumptions do not drive the results. We focus on commonly used assumptions about the noise structure—specifically, we highlight two cases, one where the noise is a simple disturbance to an underlying value and one where the noise reflects additive utility noise in the spirit of McFadden (1974, 1981).

To illustrate how noise can affect inference, consider the two choices below, which together form a *paired choice task*:

***AB* Choice**:	Lottery A: 100 percent chance of \$12	vs.	Lottery B: 50 percent chance of \$30
***CD* Choice**:	Lottery C: 20 percent chance of \$12	vs.	Lottery D: 10 percent chance of \$30

In the absence of noise, EU makes a strong prediction that individuals should prefer either lotteries A and C or lotteries B and D, since lotteries C and D are just lotteries A and B, respectively, scaled down by a common ratio of 0.2. Counter to this prediction, the CRE describes a systematic empirical pattern of people appearing more risk-tolerant in the scaled-down problem, resulting in aggregate choice frequencies where the share choosing A over B is larger than the share choosing C over D. Blavatskyy et al. (2023) provide a recent meta-study and document that out of 143 paired-choice experiments, 78 percent find that the share choosing A is greater than the share choosing C. The average difference in the shares across all 143 experiments is 22.0 percent.

Many researchers have interpreted the empirical finding of a CRE in paired choice tasks as a rejection of EU; most notably, Kahneman and Tversky (1979) use it to justify their assumption of subproportionality, one of the key properties of their probability weighting function. However, this interpretation may be invalid in the presence of noise. To illustrate, consider first the case of EU with i.i.d. additive utility noise that prior work has explored (Ballinger and Wilcox, 1997; Loomes, 2005; Hey, 2005; Wilcox, 2008; Blavatskyy, 2007, 2010; Bhatia and Loomes, 2017). Suppose the person's underlying EU preferences favor A and C. Despite this underlying preference, the person might sometimes choose B over A or D over C due to noise. The probability of choosing their preferred alternative depends on the *strength* of their preference. Under EU, when the probabilities of the initial choice are scaled down by a common ratio, the strength of the preference is scaled down by precisely the same ratio—that is, for a person who prefers A and C, their preference for A over B will be stronger than their preference for C over D. As a result, noise that is i.i.d. across the two choice tasks will be more impactful for the latter choice and will drive that choice probability further from one. Such noise thus implies that the probability of choosing A over B should be larger than the probability of choosing C over D—i.e., the person would exhibit a CRE even though they have underlying EU preferences and thus no CRP. An analogous intuition applies for a person whose EU preferences favor B and D, for whom additive i.i.d. noise implies that the probability of choosing A over B should be smaller than that for choosing C over D, indicating that they would exhibit an RCRE. In Section 2, we generalize this logic to other noise structures besides i.i.d. additive utility noise and to permit heterogeneity both in preferences and in the impact of noise.

While in principle the bias in paired choice tasks due to noise could go in either direction, there is evidence that, in practice, the prior literature's choice of experimental parameters has led to a systematic bias in the direction of finding a CRE. Of the 143 CRE paired-choice experiments reviewed by Blavatskyy et al. (2023), more than 75 percent exhibit a share of A choices greater than or equal to 50 percent, consistent with the case in which underlying preferences favor A and C and in which noise could thus yield a CRE even if there were no underlying CRP. In Section 2.4, we further detail the use of selected parameter values in the prior literature.

As an alternative to the paired choice tasks above, consider instead the two valuations below, which together form a *paired valuation task*:

***AB* Valuation**: state an m_{AB} such that

$$\text{100 percent chance of } m_{AB} \sim \text{50 percent chance of \$30}$$

***CD* Valuation**: state an m_{CD} such that

$$\text{20 percent chance of } m_{CD} \sim \text{10 percent chance of \$30}$$

EU again makes a strong prediction that, in the absence of noise, individuals should state valuations that satisfy $m_{AB} = m_{CD}$ or, equivalently, $\Delta m \equiv m_{CD} - m_{AB} = 0$. In contrast, the logic of the CRE implies that individuals are more risk-tolerant for the CD comparison than for the AB comparison, and thus we would see $m_{CD} > m_{AB}$ or $\Delta m \equiv m_{CD} - m_{AB} > 0$.

Although valuation tasks are common in experiments, researchers have rarely used them in the context of the CRE. We demonstrate, however, that under the same commonly used assumptions about noise where paired choice tasks can yield biased tests of the null of no CRP, paired valuation tasks can yield unbiased tests. Specifically, if elicited valuations are unbiased measures of the underlying values, then we can test whether the mean of Δm equals 0. Alternatively, even if elicited valuations are biased measures of underlying values—e.g., due to utility curvature—then as long as the noise is symmetric around its median, we can instead use a sign test to assess whether there are equal proportions of positive and negative instances of Δm. Notably, both tests are robust to noise having a differential impact across the AB and CD tasks.

In Section 3, we describe the details of our experimental design. We recruit 900 participants from Prolific for an online experiment. In stage 1 of the experiment, we elicit paired valuations. For each participant, we elicit the value of m_{AB} that makes them indifferent between $(\$m_{AB}, 1)$ and $(\$30, p)$, and we separately elicit the value of m_{CD} that makes them indifferent between $(\$m_{CD}, r)$ and $(\$30, rp)$. Each participant reports these paired valuations for five values of p, and between subjects we consider three different values of r; hence, in stage 1, we elicit paired valuations for 15 combinations of (p, r). In stage 2 of the experiment, we present the same participants with paired choice tasks, with one paired choice task linked to each paired valuation task from stage 1. We use this connection between stages to validate the stage 1 valuations and to reconcile our findings with the prior literature.[2]

Section 4 describes our main results using data from the paired valuation tasks. We conduct our two tests for each of the 15 paired valuation tasks in stage 1. Out of the 15 means tests, we reject the null that the mean of Δm is zero in eight comparisons at the 5 percent level. All eight rejections indicate an RCRP rather than the standard CRP, and the means are small in magnitude. Out of the 15 sign tests, we find seven significant deviations from the null of equal proportions at the 5 percent level. Six of these are consistent with an RCRP, and there is only one test in which the deviation from equal proportions is in the direction of a CRP. Beyond the formal tests, we also find that in 14 of 15 cases, the median value of Δm is zero. Thus, our paired valuation tasks yield no evidence of a systematic CRP.

Our failure to find a systematic CRP in the aggregate does not imply that our data are consistent with EU. While the data indicate an aggregate central tendency of no CRP, we document substantial CRP heterogeneity—specifically, we find significant within-individual correlations of Δm across different (p, r) combinations, so some individuals appear to have a stable CRP while others seem to have a stable RCRP. Moreover, our m_{AB} valuations yield data consistent with probability weighting models and thus inconsistent with EU. Specifically, our m_{AB} elicitations are equivalent to the tasks researchers commonly use to estimate probability weighting functions, and they yield an inverse-S-shaped probability weighting function that matches those typically found in the literature. However, the probability weighting function implied by our m_{AB} valuations is wholly inconsistent with our elicited m_{CD} valuations—indeed, it would predict m_{CD} valuations consistent with a large CRP.

In Section 5, we analyze the connections between the valuations elicited in stage 1 and the corresponding choices made in stage 2. For each paired valuation task from stage 1 (i.e., for each of a participant's five (p, r) combinations), we choose a random value of M in stage 2 and then offer the participant a binary AB choice between $(\$M, 1)$ and $(\$30, p)$, and a binary CD choice between $(\$M, r)$ and $(\$30, rp)$. The connection between these linked valuations and choices allows us to assess whether there is differential noise across the AB and CD choices and to reconcile our main finding of no systematic CRP in paired valuation tasks with the vast literature that finds a CRE in paired choice tasks.

Two key predictions link a person's stage 1 valuations to their stage 2 choices. First, reflecting the impact of preferences, the stage 1 *value difference* Δm should predict whether an individual exhibits a CRE or an RCRE at stage 2. That is, individuals who have a CRP as measured through their stage 1 valuations should be more likely to exhibit a CRE in their stage 2 choices. Second, the impact of differential noise depends on the difference between the randomly chosen amount M and the stage 1 average indifference point; we refer to this measure as the *distance to indifference*. A sufficiently large positive distance to indifference means M is large enough that preferences favor

[2]Each participant sees five paired valuation tasks and five paired choice tasks of the form described here. For robustness, we also present each participant with another five paired valuation tasks and another five paired choice tasks that use a different structure, as we describe in Sections 2 and 3.

A and C. If, in addition, the choice noise is more impactful for the CD choice, then pattern AD will be more likely than pattern BC; that is, we would observe a CRE. Analogously, a sufficiently large negative distance to indifference means M is small enough that preferences favor B and D; in this case, if the choice noise is more impactful for the CD choice, then we would observe an RCRE.

When we take these two predictions to the data, we find strong support for them at both the individual and experiment level—where an "experiment" refers to the aggregate behavior of a subset of participants who faced the same paired choice task at stage 2.[3] Our finding that stage 1 value differences strongly predict stage 2 choices provides validation that our valuations are capturing underlying preferences. Our finding that the distance to indifference predicts stage 2 choices reveals the existence of differential noise. The latter finding further demonstrates how specific parameter combinations—in particular, those which induce a positive distance to indifference—can generate a CRE in paired choice tasks even if there is no underlying CRP. Indeed, in Section 5.4, we describe how our use of a broader and more balanced set of parameter combinations (relative to the prior literature) leads to little evidence of a CRE in our stage 2 paired choices.

Our analysis has several connections to the prior literature. First, a small number of papers do use paired valuation tasks in the context of the CRE (see, e.g., Castillo and Eil, 2014; Dean and Ortoleva, 2019; Schneider and Shor, 2017; Freeman et al., 2019). These papers address different research questions relative to ours, and none of them address the differential noise problem associated with paired choice tasks nor the fact that paired valuation tasks are robust to this problem.

Our paper also connects to research highlighting how systematic behaviors that appear to contradict standard models can emerge from standard models combined with noise. Even within our context, prior work has demonstrated how an observed CRE in paired choice tasks could be due to EU with additive utility noise (Ballinger and Wilcox, 1997; Loomes, 2005; Hey, 2005; Wilcox, 2008; Blavatskyy, 2007, 2010; Bhatia and Loomes, 2017). In recent years, this theme has re-emerged in other domains. For instance, apparent nonlinear probability weighting can emerge from underlying expected value preferences combined with cognitive imprecision in the perception of probabilities (Khaw et al., 2021; Frydman and Jin, 2023) or cognitive uncertainty about the optimal action choice (Enke and Graeber, forthcoming); an apparent preference for commitment could reflect the impact of noise on a binary choice whether to commit (Carrera et al., 2022); and recent work by Oprea (2022) and Enke et al. (2023) suggests that probability weighting and present bias might reflect heuristic reactions to the inherent complexity of the choice problems.

Our work also relates more generally to literature that explicitly accounts for noise in experimental design and the analysis of risky choice data (Harless and Camerer, 1994; Hey and Orme, 1994; Ballinger and Wilcox, 1997; Loomes and Sugden, 1998; Stott, 2006). More recently, a serious evaluation of noise in canonical experimental designs has led to a preference for new types of data collection, such as valuations (Bernheim and Sprenger, 2020; Carrera et al., 2022), repeated

[3] In other words, each combination of (M, p, r) used at stage 2 generates a different "experiment." Overall, we have 120 different experiments, with an average of 75 participants in each.

choices (Gillen et al., 2019), decision times (Alós-Ferrer et al., 2021), or measures of individual uncertainty and confidence (Butler and Loomes, 2011; Enke and Graeber, forthcoming). This sort of work connecting theory and data collection efforts with an encompassing view of noise carries promise for improving the reliability of the inference drawn from experimental observations and our understanding of the forces driving behavioral anomalies.

Our paper has two broad implications for future research that we expand on in our concluding Section 6. First, our methodological contribution applies far beyond the domain of the CRE. Canonical tests for many behavioral-economic phenomena use reduced-form comparisons of pairs of choice probabilities, with examples ranging from risk preferences to time preferences to context dependence. Our theoretical results readily extend to these settings, and valuations may be similarly suited to overcome the underlying inference problem.[4]

Second, our core empirical findings present an important challenge for the literature that seeks to understand risk attitudes. Like prior research, we find clear evidence against EU as an accurate descriptive model of behavior. However, unlike prior research, our core empirical finding contradicts the received wisdom that most people exhibit a systematic underlying CRP. If this finding turns out to be robust, then it calls for a reassessment of non-EU models built around the CRP as a motivating feature of preferences. First and foremost are models of probability weighting in the tradition of prospect theory (Kahneman and Tversky, 1979; Tversky and Kahneman, 1992), where a CRP (or subproportionality) is the central motivating concept behind the structure of probability weighting. Other models also use the CRE as a key motivating fact and thus build in a systematic CRP; these include disappointment aversion (Bell, 1985; Gul, 1991) and cautious expected utility (Cerreia-Vioglio et al., 2015), among others.

2 Underlying Theory and Proposed Tests

2.1 Paired Choices and Paired Valuations

The standard common-ratio test presents participants with *paired choice tasks* that take the following form:

$$\begin{array}{llll} \boldsymbol{AB}\ \textbf{Choice Task}: & \text{choose Lottery } A \equiv (M,1) & \text{or} & \text{Lottery } B \equiv (H,p) \\ \boldsymbol{CD}\ \textbf{Choice Task}: & \text{choose Lottery } C \equiv (M,r) & \text{or} & \text{Lottery } D \equiv (H,rp), \end{array}$$

where $H > M > 0$ and $p, r \in (0,1)$. The key feature is that the CD choice task is derived from the AB choice task by scaling down the probabilities for the non-zero outcomes by a common ratio r.[5]

[4] For a recent example, see Bernheim and Sprenger (2020), who use valuations to test the assumption of rank dependence (Quiggin, 1982). While not a central part of their analysis, they discuss how choice noise presents a challenge to research that tests axioms using pairs of choice tasks.

[5] To simplify notation, we adopt the convention of omitting the zero outcomes from lotteries. For example, Lottery B yields H with probability p and zero with the remaining probability of $1-p$. Most experimental implementations of paired choice tasks set the low outcome equal to zero as we do; however, the key theoretical points also hold for a

As highlighted by Allais (1953), paired choice tasks of this type are interesting because EU makes a sharp prediction. Normalizing $u(0) = 0$:

$$EU(A) - EU(B) \ > \ 0 \quad \Leftrightarrow \quad u(M) - pu(H) \ > 0 \quad \text{and} \tag{1}$$

$$EU(C) - EU(D) \ > \ 0 \quad \Leftrightarrow \quad r\left[u(M) - pu(H)\right] > 0. \tag{2}$$

Hence, EU predicts that a person should prefer either lotteries A and C or lotteries B and D. In contrast to this prediction, the empirical finding of a *common ratio effect* (CRE) involves deviations systematically in the direction of choosing lotteries A and D—in other words, people appear more risk-tolerant in the scaled-down choice. More precisely, letting $\widehat{\Pr}(X)$ be the proportion of participants who choose X, the common finding is $\widehat{\Pr}(AD) > \widehat{\Pr}(BC)$ or, equivalently, $\widehat{\Pr}(A) > \widehat{\Pr}(C)$.[6,7] We use the label *reverse common ratio effect* (RCRE) for the less common finding of deviations in the direction of choosing lotteries B and C, or a finding of $\widehat{\Pr}(A) < \widehat{\Pr}(C)$.

An alternative common-ratio test that researchers have used much less often presents participants with *paired valuation tasks.* Our main analysis will focus on *m-valuation tasks* in which we fix (H, p, r) and present participants with the following tasks:

$$\begin{array}{llll} \boldsymbol{AB} \textbf{ Valuation Task}: & \text{state an } m_{AB} \in [0, H] \text{ such that} & (m_{AB}, 1) \ \sim \ (H, p) \\ \boldsymbol{CD} \textbf{ Valuation Task}: & \text{state an } m_{CD} \in [0, H] \text{ such that} & (m_{CD}, r) \ \sim \ (H, rp). \end{array}$$

For paired valuation tasks, a finding of $\Delta m \equiv m_{CD} - m_{AB} > 0$ reflects a CRE. Again, a CRE involves people acting more risk-tolerant in the scaled-down task. Thus, an individual will demand a higher premium to accept the safer option in the CD task relative to what they demand in the AB task, or $m_{CD} > m_{AB}$. Analogously, a finding of $\Delta m < 0$ reflects an RCRE, and a finding of $\Delta m = 0$ would be consistent with EU (among other models).

A CRE, in choices or valuations, is an empirical finding; when interpreting this empirical finding in terms of underlying preferences, it is essential to account for noise. The following two subsections do so, first for the special case of EU and i.i.d. additive utility noise, and then for the more general case.

2.2 EU and i.i.d. Additive Utility Noise

Prior work has noted that EU with i.i.d. additive utility noise can naturally generate a CRE in paired choice tasks (Ballinger and Wilcox, 1997; Loomes, 2005; Hey, 2005; Wilcox, 2008; Blavatskyy,

non-zero low outcome.

[6] From here onward, we use $\widehat{\Pr}$ to denote empirically observed proportions, and Pr to denote model-predicted proportions. To ease notation, we suppress the choice set because it is typically self-explanatory; for example, $\widehat{\Pr}(A)$ is the proportion who choose A from the choice set $\{A, B\}$, and $\widehat{\Pr}(AD)$ is the proportion who choose combination AD from the choice set $\{AC, AD, BC, BD\}$.

[7] Note that $\widehat{\Pr}(A) = \widehat{\Pr}(AD) + \widehat{\Pr}(AC)$ and $\widehat{\Pr}(C) = \widehat{\Pr}(BC) + \widehat{\Pr}(AC)$, and thus $\widehat{\Pr}(AD) > \widehat{\Pr}(BC)$ is equivalent to $\widehat{\Pr}(A) > \widehat{\Pr}(C)$.

2007, 2010; Bhatia and Loomes, 2017). In this subsection, we illustrate the intuition behind this result.

Suppose there exist i.i.d. distributed noise draws ϵ_{AB} and ϵ_{CD} such that a person chooses lottery A over lottery B when $EU(A) - EU(B) > \epsilon_{AB}$, and chooses lottery C over lottery D when $EU(C) - EU(D) > \epsilon_{CD}$. Applying equations (1) and (2), $EU(C) - EU(D) = r[EU(A) - EU(B)]$. Hence, if F is the shared CDF for ϵ_{AB} and ϵ_{CD}, then the likelihoods of choosing lottery A over lottery B and of choosing lottery C over lottery D are, respectively,

$$\Pr(A) = F\left(EU(A) - EU(B)\right) \quad \text{and} \quad \Pr(C) = F\left(r\left[EU(A) - EU(B)\right]\right).$$

Figure 1 illustrates the implications for the case when F is a standard normal distribution. Panel A presents the choice probabilities. The solid line depicts $\Pr(A)$ as a function of $EU(A)-EU(B)$. If $EU(A)-EU(B) = 0$, in which case the person's underlying EU preferences are indifferent between lotteries A and B, then $\Pr(A) = 1/2$. As the person's underlying preference for lottery A gets stronger—i.e., as $EU(A) - EU(B)$ gets more positive—$\Pr(A)$ grows toward 1. As the person's underlying preference for lottery B gets stronger—i.e., as $EU(A) - EU(B)$ gets more negative—$\Pr(A)$ shrinks toward 0. The dashed and dotted lines in Panel A depict $\Pr(C)$ as a function of $EU(A) - EU(B)$ for the cases of $r = 0.5$ and $r = 0.25$. Each line has the same qualitative pattern as $\Pr(A)$. However, when we scale down the probabilities by a common ratio r, we also scale down the strength of preference by the same factor. As a result, the i.i.d. noise will be more impactful for the CD choice than for the AB choice. Therefore, for any specific value of $EU(A) - EU(B)$, $\Pr(C)$ is closer to 1/2 than $\Pr(A)$ is.

Panel B of Figure 1 converts the choice probabilities from Panel A into a predicted difference $CRE-RCRE \equiv \Pr(A)-\Pr(C)$. Panel B illustrates that a person whose underlying EU preferences favor A and C will exhibit a CRE. Conversely, a person whose underlying EU preferences favor B and D will exhibit an RCRE. Panel B further highlights how the predicted magnitude of $CRE - RCRE$ depends on (i) the magnitude of the common ratio r and (ii) the strength of the person's underlying preference $EU(A) - EU(B)$. Finally, note that at extreme levels of $EU(A) - EU(B)$, $CRE - RCRE$ goes to zero as the preference component becomes so strong that it dominates the noise.

Figure 1 illustrates a fundamental problem in using paired choice tasks to test for an underlying CRP. In the following subsection, we develop a more general model of underlying preferences and noise that permits us to expand on this problem and illustrate how paired valuation tasks can be a solution.

2.3 More General Model of Underlying Preferences and Noise

We develop a framework to interpret data from both paired choice tasks and paired valuation tasks. Specifically, we assume a person has a realized indifference value that is determined from a

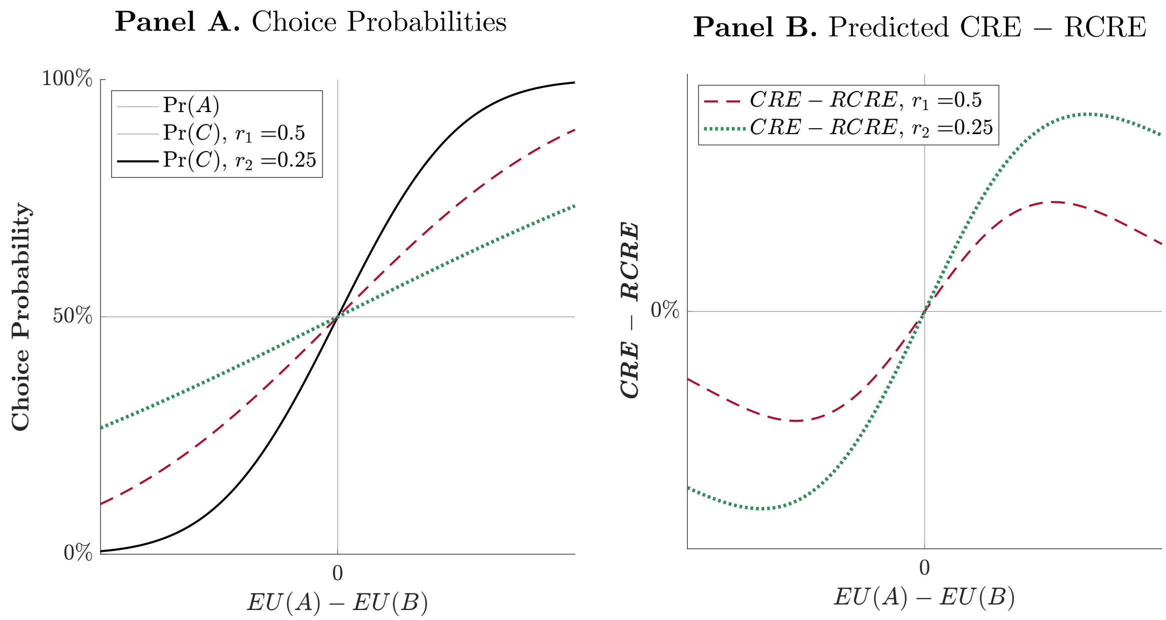

Figure 1: Predicted choice probabilities (panel A) and predicted $CRE - RCRE$ (panel B) as a function of $EU(A) - EU(B)$ under expected utility with additive i.i.d. noise. In both panels, the dashed red line reflects the case $r_1 = 0.5$ and the dotted green line reflects the case $r_2 = 0.25$. Both panels assume that the noise follows a standard normal distribution.

combination of their underlying preferences and noise. The person then makes a choice or states a valuation implied by this realized indifference value. The case of EU with i.i.d. additive utility noise studied in Section 2.2 is a special case of this framework.

Importantly, we impose the same structure of preferences and noise across both types of tasks. Hence, the different results we document are not due to different assumptions across the two types of tasks. Moreover, this framework implies a strong connection between paired valuation tasks and paired choice tasks. Motivated by this connection, our experiment collects data from the same participants on paired valuation tasks and linked paired choice tasks. We show in Section 5 that participants' valuations are predictive of their choices in a way that is consistent with this framework.[8]

Without loss of generality, we fix (H, p, r) and focus on behavior as a function of M. Assuming preferences are monotonic and continuous, for each (H, p, r) a person will have a pair of underlying indifference points (m^*_{AB}, m^*_{CD}) such that their (noise-free) preferences satisfy:

- Prefer $A \equiv (M, 1)$ over $B \equiv (H, p)$ if and only if $M \geqslant m^*_{AB}$, and

[8]Some researchers have debated whether valuations or choices provide better insight into people's underlying preferences. For instance, based on discrepancies between behavior in valuations versus binary choices, combined with an assumption that binary choices reflect true preferences, some have argued that valuations are unreliable (Brown and Healy, 2018; Freeman et al., 2019). Others have argued that binary choices, themselves, may be unreliable (Freeman and Mayraz, 2019). The systematic connections that we document in Section 5 suggest that neither conclusion is warranted.

- Prefer $C \equiv (M, r)$ over $D \equiv (H, rp)$ if and only if $M \geqslant m^*_{CD}$.[9]

Given these underlying indifference points, we assume that noise impacts choices and valuations as follows:

Assumption 1: Impact of Noise on Choices and Valuations

A person's *realized indifference points* (m_{AB}, m_{CD}) are $m_{AB} \equiv \Gamma(m^*_{AB}, \varepsilon_{AB})$ and $m_{CD} \equiv \Gamma(m^*_{CD}, \varepsilon_{CD})$, where $(\varepsilon_{AB}, \varepsilon_{CD})$ are noise draws from a continuous joint distribution with convex support, and Γ is increasing in both arguments and has $\Gamma(m, 0) = m$ for all m. Then:

- In an AB choice task, the person chooses $A \equiv (M, 1)$ over $B \equiv (H, p)$ if and only if $M \geqslant m_{AB} \equiv \Gamma(m^*_{AB}, \varepsilon_{AB})$,
- In a CD choice task, the person chooses $C \equiv (M, r)$ over $D \equiv (H, rp)$ if and only if $M \geqslant m_{CD} \equiv \Gamma(m^*_{CD}, \varepsilon_{CD})$,
- In an AB valuation task, the person states valuation $m_{AB} \equiv \Gamma(m^*_{AB}, \varepsilon_{AB})$, and
- In a CD valuation task, the person states valuation $m_{CD} \equiv \Gamma(m^*_{CD}, \varepsilon_{CD})$.

The distinction between underlying preferences and observed behaviors is integral to our analysis. To highlight this distinction, we say that a person has a *common ratio preference* (CRP) if they have $\Delta m^* \equiv m^*_{CD} - m^*_{AB} > 0$, and a *reverse common ratio preference* (RCRP) if they have $\Delta m^* < 0$. Assessing whether an observed CRE in choices or valuations is evidence of an underlying CRP is the key inferential challenge that we focus on in the remainder of this section.[10]

In Assumption 1, the function Γ permits a variety of models for how a person's underlying indifference points combine with choice noise to generate their realized indifference points. We highlight two special cases of Assumption 1:

Assumption 2a: $\Gamma(m, \varepsilon) = m + \varepsilon$, $\varepsilon_{CD} \overset{d}{=} k\varepsilon_{AB}$ for some $k > 0$, and $E(\varepsilon_{AB}) = E(\varepsilon_{CD}) = 0$.

Assumption 2b: $\Gamma(m, \varepsilon)$ is potentially nonlinear in m and ε, but $\varepsilon_{CD} \overset{d}{=} k\varepsilon_{AB}$ for some $k > 0$, and ε_{AB} is symmetric about 0.

Assumption 2a is consistent with assumptions researchers frequently use when analyzing valuations data, where they model noise as an additive disturbance to an underlying value. Assumption

[9] To simplify the exposition, the text assumes that the person prefers and chooses the safer option when indifferent, but given we also assume continuous noise, this assumption is immaterial.

[10] We use the term "preference" to capture all stable drivers of behavior (as opposed to transient confusion or noise); in addition to reflecting "true" tastes, these drivers might also include stable heuristics. We also note that, by CRP and RCRP, we do not mean that people care directly about common-ratio transformations of probabilities. Instead, we use these terms to capture the idea that underlying risk attitudes react systematically to these transformations. Based on the existing CRE evidence, many theories—most notably, models of probability weighting—assume that underlying preferences have a systematic CRP.

2b is consistent with assumptions researchers frequently use when analyzing choice data, where they instead model noise as a symmetric additive perturbation of utility in the spirit of McFadden (1974, 1981). To illustrate, we describe when Assumptions 2a and 2b would hold under two prominent models of underlying preferences.

Example 1: Expected Utility and Prospect Theory

Suppose that a person evaluates a lottery (x, q) with $x > 0$ as $\pi(q)u(x)$. This formulation corresponds to both original prospect theory as in Kahneman and Tversky (1979) and cumulative prospect theory as in Tversky and Kahneman (1992), where $\pi(\cdot)$ is a probability weighting function and $u(\cdot)$ is a value function defined over gains and losses. This formulation also reduces to EU when $\pi(q) = q$ for all q and $u(\cdot)$ is a Bernoulli utility function. Under this formulation, the underlying indifference points satisfy

$$\begin{aligned} u(m^*_{AB}) = \pi(p)u(H) \qquad &\Leftrightarrow \qquad m^*_{AB} = u^{-1}\left(\pi(p)u(H)\right) \\ \pi(r)u(m^*_{CD}) = \pi(rp)u(H) \qquad &\Leftrightarrow \qquad m^*_{CD} = u^{-1}\left(\frac{\pi(rp)}{\pi(r)}u(H)\right). \end{aligned}$$

When working with valuations data, one might incorporate noise by assuming that observed valuations satisfy $m_{AB} = m^*_{AB} + \varepsilon_{AB}$ and $m_{CD} = m^*_{CD} + \varepsilon_{CD}$. This formulation satisfies Assumption 2a as long as $\varepsilon_{CD} \overset{d}{=} k\varepsilon_{AB}$ for some $k > 0$ and $E(\varepsilon_{AB}) = E(\varepsilon_{CD}) = 0$—e.g., if ε_{AB} and ε_{CD} are both mean-zero normal or logistic distributions with possibly different variances.[11]

Alternatively, for either valuations or choice data, one might incorporate additive utility noise by instead assuming that the realized indifference points satisfy

$$\begin{aligned} u(m_{AB}) = \pi(p)u(H) + \epsilon_{AB} \qquad &\Leftrightarrow \qquad m_{AB} = u^{-1}\left(u(m^*_{AB}) + \epsilon_{AB}\right) \\ \pi(r)u(m_{CD}) = \pi(rp)u(H) + \epsilon_{CD} \qquad &\Leftrightarrow \qquad m_{CD} = u^{-1}\left(u(m^*_{CD}) + \epsilon_{CD}/\pi(r)\right) \end{aligned}$$

where ϵ_{AB} and ϵ_{CD} reflect additive utility noise.[12] This formulation fits Assumption 1 with $\Gamma(m, \varepsilon) = u^{-1}(u(m) + \varepsilon)$, $\varepsilon_{AB} = \epsilon_{AB}$, and $\varepsilon_{CD} = \epsilon_{CD}/\pi(r)$. This formulation further satisfies Assumption 2b as long as ϵ_{AB} is symmetric about 0 and $\epsilon_{CD} \overset{d}{=} k'\epsilon_{AB}$ for some $k' > 0$—e.g., if the additive utility noise terms ϵ_{AB} and ϵ_{CD} are both mean-zero normal or logistic distributions.[13] Finally, note that the case of EU with i.i.d. additive utility noise that we described in Section 2.2 fits Assumption 2b with $\varepsilon_{CD} = \varepsilon_{AB}/r$, and so $k = 1/r$.

[11] For instance, this approach is used by Tversky and Kahneman (1992) and Bruhin et al. (2010).

[12] To help clarify our exposition, we use ϵ to denote additive perturbations to utility, and ε to denote the noise described in Assumption 1. The latter equations use $\pi(p)u(H) = u(m^*_{AB})$ and $\pi(rp)u(H) = \pi(r)u(m^*_{CD})$.

[13] For instance, this approach is used by Camerer and Ho (1994), Hey and Orme (1994), and Wu and Gonzalez (1996).

The framework of Assumption 1, with Assumptions 2a and 2b as special cases, allows us to demonstrate the problems with using paired choice tasks, and when paired valuation tasks might be robust to those problems. Proposition 1 establishes conditions under which paired choice tasks yield a biased test of the null of $\Delta m^* = 0$. All proofs appear in Online Appendix A.

Proposition 1 *(Paired Choice Tasks Can Yield Biased Tests of* $\Delta m^* = 0$*):* Consider a person who has $m^*_{AB} = m^*_{CD} \equiv m^*$ and thus $\Delta m^* = 0$, and suppose that $\varepsilon_{CD} \overset{d}{=} k\varepsilon_{AB}$ for some $k > 0$, and define $Z \equiv \Pr(\varepsilon_{AB} < 0) = \Pr(\varepsilon_{CD} < 0)$.

(1) If $M - m^* > 0$ and thus the person prefers A and C, then:

  (a) $k > 1$ implies $\Pr(A) > \Pr(C) > Z$ (CRE);

  (b) $k < 1$ implies $\Pr(C) > \Pr(A) > Z$ (RCRE); and

  (c) $k = 1$ implies $\Pr(A) = \Pr(C) > Z$.

(2) If $M - m^* < 0$ and thus the person prefers B and D, then:

  (a) $k > 1$ implies $\Pr(A) < \Pr(C) < Z$ (RCRE);

  (b) $k < 1$ implies $\Pr(C) < \Pr(A) < Z$ (CRE); and

  (c) $k = 1$ implies $\Pr(A) = \Pr(C) < Z$.

(3) If $M - m^* = 0$, then $\Pr(A) = \Pr(C) = Z$ for all k.

Proposition 1 captures and then generalizes the conclusions from Section 2.2. In our general framework, the case of EU with i.i.d. additive utility noise has $k = 1/r > 1$ (see Example 1). Moreover, under EU, $M - m^* > 0$ implies $EU(A) - EU(B) > 0$ whereas $M - m^* < 0$ implies $EU(A) - EU(B) < 0$. Parts (1)(a) and (2)(a) of Proposition 1 therefore imply the same predictions as illustrated in Figure 1, reiterating the core intuition that the AB preference is stronger than the CD preference under EU, and thus i.i.d. noise will have a larger impact on the CD choice. Hence, if EU preferences favor A and C, then the prediction is $\Pr(A) > \Pr(C)$. In contrast, if EU preferences favor B and D, then the prediction is $\Pr(A) < \Pr(C)$.[14]

Proposition 1 generalizes these conclusions in several ways. For the case of EU, Proposition 1 characterizes predictions under different noise structures besides i.i.d. additive noise, including when the noise is a disturbance to the underlying indifference values and when there is additive utility noise but with different variances across the two choices. Our framework and Proposition 1 also apply to any non-EU model that exhibits $\Delta m^* = 0$. The key feature that determines predictions is whether the noise is more impactful for the CD choices (i.e., $k > 1$), in which case the predictions are as in Section 2.2. The predictions flip if the noise is more impactful for the AB choices (i.e., $k < 1$). The implication is that, when using paired choice tasks, a person with $\Delta m^* = 0$ could exhibit either a CRE or an RCRE depending on the combination of whether (i)

[14] By assuming a standard normal distribution, Figure 1 also has the feature that either $\Pr(A) > \Pr(C) > 1/2$ or $\Pr(A) < \Pr(C) < 1/2$. Proposition 1 establishes that the 1/2 threshold is replaced by $Z \equiv \Pr(\varepsilon_{AB} < 0)$ for more general noise distributions.

the offered M leads to an underlying preference for A and C versus B and D, and (ii) choice noise has a larger impact on the AB choice versus the CD choice.

When using data from paired choice tasks to test the null of $\Delta m^* = 0$, researchers typically assess whether $\widehat{\Pr}(A) = \widehat{\Pr}(C)$. Proposition 1 implies that the theoretical prediction of $\Pr(A) = \Pr(C)$ holds only when $k = 1$. This knife-edge case seems unlikely to hold in practice, and thus paired choice tasks are likely to yield a biased test.

Whereas paired choice tasks yield a biased test in all but this knife-edge case, Proposition 2 shows that unbiased tests are possible using paired valuation tasks under a much broader set of conditions.

Proposition 2 *(Paired Valuation Tasks Can Yield Unbiased Tests of* $\Delta m^* = 0$*):* Consider a person who faces a paired valuation task, and let m_{AB} and m_{CD} be their stated valuations.

(1) If $\Gamma(m, \varepsilon) = m + \varepsilon$ and $E(\varepsilon_{AB}) = E(\varepsilon_{CD})$, then $E(\Delta m) = \Delta m^*$.
(2) If a person has $m^*_{AB} = m^*_{CD} \equiv m^*$, and if the joint distribution $(\varepsilon_{AB}, \varepsilon_{CD})$ is symmetric around some median vector $(\varepsilon', \varepsilon')$,[15] then $\Pr(\Delta m > 0) = \Pr(\Delta m < 0) = 1/2$.

Part (1) of Proposition 2 describes conditions under which we can test the null of $\Delta m^* = 0$ using a means test, specifically, testing whether $\widehat{E}(\Delta m) = 0$.[16] When $E(\varepsilon_{AB}) = E(\varepsilon_{CD}) = 0$, as under Assumption 2a, m_{AB} and m_{CD} are unbiased measures of the underlying indifference points, and thus their difference is an unbiased measure of Δm^*. Furthermore, even if the errors have non-zero means, in which case m_{AB} and m_{CD} are biased measures of the underlying indifference points, their difference remains an unbiased measure of Δm^* as long as the errors have the same mean.[17]

A test based on $\widehat{E}(\Delta m)$ becomes biased if Γ is a nonlinear function of m, which can arise when the noise is modeled as additive utility noise (see Example 1). Indeed, in Online Appendix B.1 we show that, for the case of EU with additive i.i.d. utility noise, concave utility implies $E(\Delta m) > 0$. Thus a test based on $\widehat{E}(\Delta m)$ would be biased towards rejecting the null of $\Delta m^* = 0$ in favor of a CRP.

Given this potential concern, part (2) of Proposition 2 describes conditions under which we can test the null of $\Delta m^* = 0$ using a sign test. This test compares whether the observed proportions of $\Delta m > 0$ and $\Delta m < 0$ are the same. The intuition behind why this approach provides an unbiased test of $\Delta m^* = 0$ is easiest to see when noise is symmetric around zero (i.e., $\varepsilon' = 0$). For this case,

[15] Formally, if $(\varepsilon_{AB}, \varepsilon_{CD})$ has joint distribution F with PDF f, then symmetry around $(\varepsilon', \varepsilon')$ implies $f(\varepsilon' + z_{AB}, \varepsilon' + z_{CD}) = f(\varepsilon' - z_{AB}, \varepsilon' - z_{CD})$ for all (z_{AB}, z_{CD}). This property holds, for instance, for a bivariate normal with mean $(\varepsilon', \varepsilon')$ and any correlation.

[16] Analogous to our use of $\widehat{\Pr}$ to denote empirically observed proportions and Pr to denote model-predicted proportions, we use $\widehat{E}$ to denote empirically observed averages and E to denote model-predicted averages.

[17] Some researchers have raised concerns that certain methods for eliciting valuations, such as price lists, may lead to systematic errors of over- or under-valuation. Our results imply that a test based on $\hat{E}(\Delta m) = 0$ remains valid provided these systematic errors have the same mean across the pair of valuations.

a person with $m^*_{AB} = m^*_{CD}$ will exhibit $\Delta m > 0$ if and only if $\varepsilon_{CD} > \varepsilon_{AB}$. The symmetry of the noise then implies that, for any $\bar{\varepsilon}$, it is equally likely to get $\Delta m < 0$ due to $\varepsilon_{AB} = \bar{\varepsilon}$ and $\varepsilon_{CD} < \bar{\varepsilon}$ as it is to get $\Delta m > 0$ due to $\varepsilon_{AB} = -\bar{\varepsilon}$ and $\varepsilon_{CD} > -\bar{\varepsilon}$. Averaging over all $\bar{\varepsilon}$, it is equally likely to get $\Delta m > 0$ and $\Delta m < 0$.[18]

Proposition 1 outlines scenarios under which paired choice tasks yield a biased test of the null of $\Delta m^* = 0$. In contrast, Proposition 2 outlines scenarios under which paired valuation tasks can yield an unbiased test. Corollary 1 highlights how Assumptions 2a and 2b, which reflect assumptions commonly made in the literature, satisfy both scenarios.

Corollary 1: Under Assumption 2a, paired choice tasks can yield a biased test of $\Delta m^* = 0$, whereas paired valuations can yield an unbiased test based on the mean of Δm. Under Assumption 2b, paired choice tasks can yield a biased test of $\Delta m^* = 0$, whereas paired valuations can yield an unbiased test based on the sign of Δm.

2.4 Evaluating Prior Experimental Tests

Figure 2 depicts the set of theoretical predictions for a population with $\Delta m^* = 0$ under Assumption 2a together with data from existing experimental tests of the CRE. The theoretical predictions, which we formally derive in Online Appendix B.2, permit heterogeneity in preferences and in the impact of noise.

Panel A focuses on data from paired-choice experiments, where observed behavior in a population is $\widehat{\Pr}(A)$ and $\widehat{\Pr}(C)$. If, for everyone, the impact of noise were the same for both the AB and the CD choices (i.e., $k = 1$ in Proposition 1), then the set of predicted $(\Pr(A), \Pr(C))$ combinations consistent with a population in which everyone has $\Delta m^* = 0$ would be represented by the 45-degree line. Once we allow for the possibility of differential noise, the set of predicted $(\Pr(A), \Pr(C))$ combinations consistent with $\Delta m^* = 0$ expands considerably to become the gray shaded region.[19]

Panel B focuses on data from paired-valuation experiments, where observed behavior in a population is $\widehat{E}(m_{AB})$ and $\widehat{E}(m_{CD})$. Under Assumption 2a, the set of predicted $(E(m_{AB}), E(m_{CD}))$

[18] Our formal test uses the following logic. If $\Pr(\Delta m > 0) = \Pr(\Delta m < 0) = 1/2$ for every observation, the likelihood of observing at most n instances of $\Delta m > 0$ out of N observations is equal to $G(n, N)$, where G denotes the cumulative distribution function for a binomial distribution with a 50 percent success rate. Hence, if we observe n_+ instances of $\Delta m > 0$ and n_- instances of $\Delta m < 0$, the p-value for a two-sided sign test under the null of $\Delta m^* = 0$ is $2 * G(\min\{n_+, n_-\}, n_+ + n_-)$.

[19] The gray shaded region reflects the set of possible $(\Pr(A), \Pr(C))$ predictions that one can construct when one permits any distribution of preferences for A versus B and imposes no restrictions on noise other than it being median zero. The area below (above) the 45-degree line and above (below) $\Pr(C) = 1/2$ can be constructed using homogeneous preferences that favor A and C (B and D) and noise that is more impactful for the CD choice. The mirror images of these regions can be constructed using noise that is more impactful for the AB choice. The remaining shaded area can be constructed using heterogeneity in the preferences for A versus B, effectively permitting convex combinations of the above areas. Prior researchers (Ballinger and Wilcox, 1997; Wilcox, 2008) have provided examples that combine heterogeneity and noise to generate a population outcome with $\Pr(A) > 1/2 > \Pr(C)$; our analysis in Online Appendix B.2 characterizes the complete set.

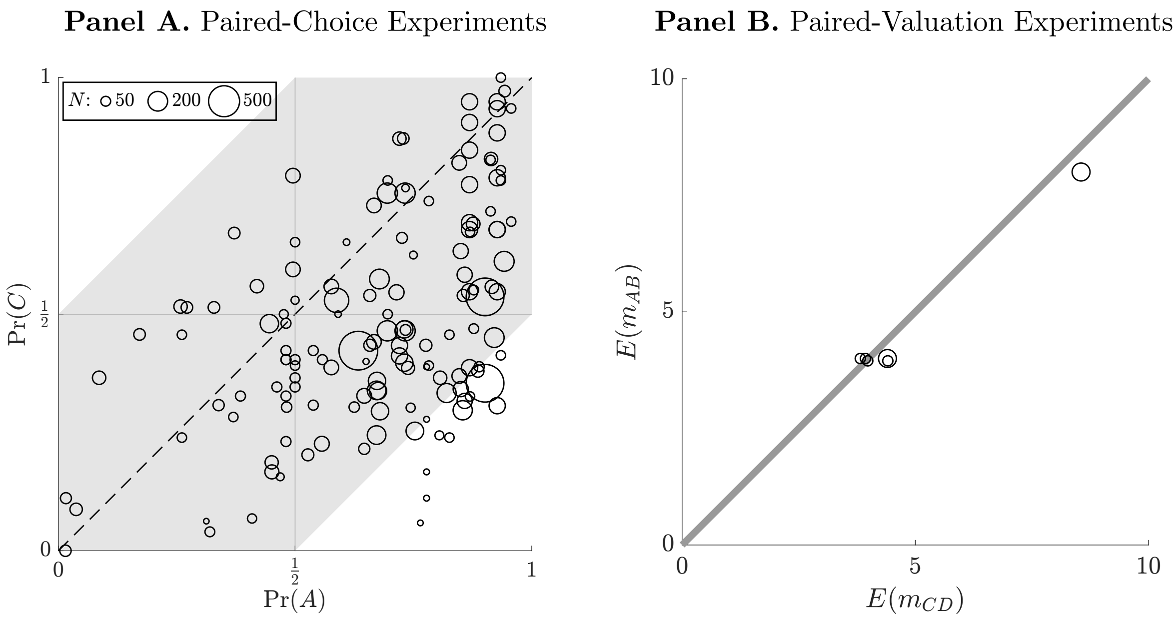

Figure 2: Predictions and observations for paired-choice experiments (panel A) and paired-valuation experiments (panel B). In each panel, points below the 45-degree line are combinations that indicate a CRE, while points above the 45-degree line are combinations that indicate an RCRE. The shaded grey region in panel A denotes predicted $(\Pr(A), \Pr(C))$ combinations consistent with $\Delta m^* = 0$ under Assumption 2a; the solid grey line in panel B denotes predicted $(E(m_{AB}), E(m_{CD}))$ combinations consistent with $\Delta m^* = 0$ under Assumption 2a (see Online Appendix B.2 for derivations). The black circles denote empirical observations from previous experiments. Panel A depicts 143 CRE paired-choice experiments surveyed by Blavatskyy et al. (2023) scaled by the number of observations; panel B depicts six similarly scaled CRE paired-valuation experiments identified by us (see Online Appendix B.3 for details).

combinations that are consistent with a population in which everyone has $\Delta m^* = 0$ is represented by the grey bold-faced 45-degree line, even with differential noise.

The black circles in Panel A depict observed $(\widehat{\Pr}(A), \widehat{\Pr}(C))$ combinations from 143 CRE paired-choice experiments across 39 studies identified in the meta-study by Blavatskyy et al. (2023). The typical test for a CRE in paired choice tasks is to assess whether $\widehat{\Pr}(A) > \widehat{\Pr}(C)$, and panel A reveals that the vast majority of experiments (112 experiments, or 78 percent) exhibit this pattern—hence, the widespread perception that there exists a systematic CRE. At the same time, panel A also reveals that almost all experiments (129 experiments, or 90 percent) fall within the gray area and thus are consistent with $\Delta m^* = 0$, i.e., no CRP or RCRP, once one permits the possibility of differential noise.[20]

The black circles in panel B of Figure 2 depict observed $(\hat{E}(m_{AB}), \hat{E}(m_{CD}))$ combinations from

[20] While the existence of a CRE is typically assessed by testing whether $\widehat{\Pr}(A) - \widehat{\Pr}(C) > 0$, a more stringent test is whether $\widehat{\Pr}(AD) > 1/2$. Such a finding would be inconsistent with Proposition 1 if the noise has a median of zero and thus $Z = 1/2$. Among the 143 experiments in panel A, there are 20 experiments with $\widehat{\Pr}(AD) \geqslant 1/2$. While this provides some evidence against the no-CRP null, one would probably not conclude from this evidence that there is a systematic CRP.

six CRE paired-valuation experiments across two studies that we identified (see Online Appendix B.3 for details about the data). Panel B demonstrates that the data from paired valuation tasks do not fall far from the 45-degree line. Moreover, comparing panel B to panel A reveals that the literature has tested for a CRE using almost exclusively paired choice tasks.

Another notable feature of Figure 2(A) is that more than 75 percent of prior paired-choice experiments found $\widehat{\Pr}(A) \geqslant 1/2$. Interpreted in terms of Proposition 1, if the noise is indeed more impactful for the CD choices, then the vast majority of prior studies reflect instances where noise would yield a CRE even if there were no underlying CRP. Of course, whether $\widehat{\Pr}(A) > 1/2$ in any particular study depends largely on the experimenter's choice of parameters M, H, and p. In other words, prior studies have focused on parameters that predominantly yield $\widehat{\Pr}(A) > 1/2$.

A closer look at the data from Blavatskyy et al. (2023) reveals that there has been relatively little variation in experimental parameters in the prior literature. Indeed, 33.6 percent of the experiments in Figure 2(A) use exactly the same combination ($p = 0.8, r = 0.25$) that was used in Problems 3 and 4 in Kahneman and Tversky (1979). More notably, the ratio $M/(pH)$ is a natural measure of the likely preference for A versus B: The larger this ratio, the more likely it is that participants prefer A over B. 115 of the 143 experiments in Figure 2(A) have $M/(pH) \geqslant 0.75$; among these, the average (sample-weighted) $CRE - RCRE$ is 23.9 percent, compared to only 8.8 percent among the other 28 experiments.

Unlike the prior literature, when we collect data on paired choice tasks in stage 2 of our study, we use a broader and more balanced set of experimental parameters. In Section 5.4, we investigate the impact of doing so on what one might conclude from paired choice tasks.

2.5 Robustness: h-Valuation Tasks

Our primary analysis focuses on the m-valuation tasks described in Sections 2.1 and 2.3 in which we fix (H, p, r) and elicit the m that makes people indifferent. We prefer these tasks for two reasons. First, they have a natural bounded domain of $m \in [0, H]$ that is the same for all p and r. Second, the AB variants of the m-valuation tasks are equivalent to the valuation tasks that researchers typically use to estimate probability weighting functions; thus, one way to validate our approach is to compare our observed AB valuations to what the prior literature has found.

However, as a robustness check, we also consider *h-valuation tasks* in which we fix (M, p, r) and elicit an $h_{AB} \geqslant M$ and an $h_{CD} \geqslant M$ such that

$$\begin{aligned} & (M, 1) \sim (h_{AB}, p) \\ \text{and} \qquad & (M, r) \sim (h_{CD}, rp). \end{aligned}$$

Online Appendix B.4 provides a full development of the theory underlying h-valuation tasks that is analogous to the theory developed in Section 2.3 for m-valuation tasks. Here, we highlight two important points that we use in our analysis. First, in terms of the underlying indifference values (h^*_{AB}, h^*_{CD}), a CRP implies $h^*_{AB} > h^*_{CD}$. We therefore define $\Delta h^* \equiv h^*_{AB} - h^*_{CD}$ so that, analogous to $\Delta m^* > 0$, a CRP is reflected by $\Delta h^* > 0$. Our empirical object of interest is thus $\Delta h \equiv h_{AB} - h_{CD}$. Second, for a fixed (p, r), the m- and h-valuation tasks measure approximately the same preference. Hence, for any given (p, r), there should be a positive correlation between m_z/H in an m-valuation task and M/h_z in the corresponding h-valuation task as both measure a proportional risk premium.[21] We assess this correlation empirically in Section 4.2 as one way to validate our valuation-task data.

3 Experimental Methodology

Our experimental design closely mirrors our theoretical framework and consists of two main stages.[22] At the beginning of the experiment, each participant is randomly assigned a common-ratio factor $r \in \{0.2, 0.4, 0.6\}$. This value remains constant throughout the experiment. In stage 1, participants complete ten paired valuation tasks for a total of 20 valuations. In stage 2, participants complete ten paired choice tasks for a total of 20 binary choices. Each paired choice task corresponds to one of the paired valuation tasks from stage 1. Participants complete all 20 valuations before proceeding to the 20 binary choices, and we randomize the order of questions within each stage. Figure 3 provides a high-level overview of the experiment timeline.

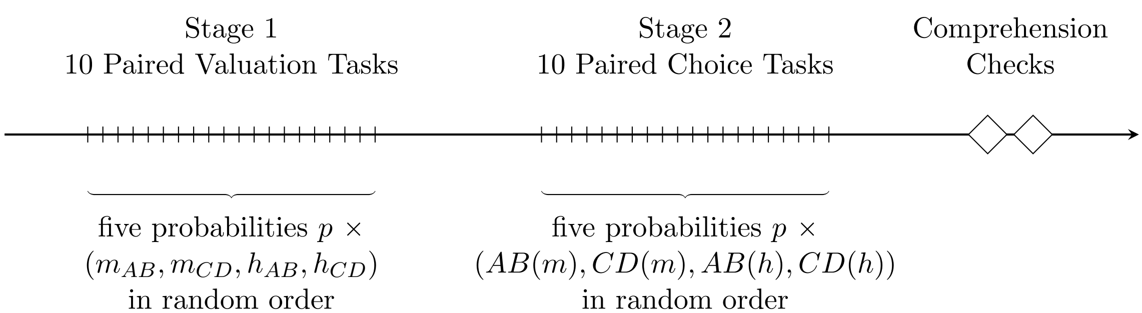

Figure 3: Experiment Timeline

3.1 Stage 1: Paired Valuation Tasks

Table 1 provides an example of one paired m-valuation task. Each valuation in the pair consists of a series of choices in a multiple-price-list format. For all m-valuations, we fix the larger outcome at $H = \$30$, and we elicit an indifference value m by varying M in \$1 increments from \$0 to \$30. In the AB variant, the left-hand option remains fixed at a p chance of \$30, while the right-hand option is a 100 percent chance of M that starts with $M = \$0$ and increases by \$1 per row. The

[21] For instance, under EU with $u(x) = x^\alpha$, $m_z/H = M/h_z = p^{1/\alpha}$ for both $z = AB$ and $z = CD$. More generally, there need not be an exact equivalence, but there should be a positive correlation.

[22] Our experiment was preregistered in the AEA RCT Registry in August 2021, under the ID AEARCTR-0008058.

paired CD variant has the same structure, except we reduce both probabilities by the participant's common ratio r. For each variant, we take the average value of M at the switching rows to be our measure of the realized indifference point, which we denote by m_{AB} or m_{CD}.

Table 1: Paired m-Valuation Tasks

Panel A: AB Variant

Option B		**Option** A
$1-p$ CHANCE OF \$0, p CHANCE OF \$30	OR	100% CHANCE OF \$0
$1-p$ CHANCE OF \$0, p CHANCE OF \$30	OR	100% CHANCE OF \$1
...	...	...
$1-p$ CHANCE OF \$0, p CHANCE OF \$30	OR	100% CHANCE OF \$30

Panel B: CD Variant

Option D		**Option** C
$1-rp$ CHANCE OF \$0, rp CHANCE OF \$30	OR	$1-r$ CHANCE OF \$0 r CHANCE OF \$0
$1-rp$ CHANCE OF \$0, rp CHANCE OF \$30	OR	$1-r$ CHANCE OF \$0 r CHANCE OF \$1
...	...	...
$1-rp$ CHANCE OF \$0, rp CHANCE OF \$30	OR	$1-r$ CHANCE OF \$0 r CHANCE OF \$30

Note: Structure of multiple-price lists for the AB and CD variants of a paired m-valuation task. Each participant faces five such pairs, all with the same r, but with five different values for p.

For the h-valuations, we fix the smaller outcome and elicit the larger outcome as the indifference point. Specifically, we fix the value of the smaller outcome at $M = \$(p \cdot 30)$, and we elicit an indifference value h by varying H in \$1 increments from $\$(p \cdot 30)$ to $\$(p \cdot 30 + 30)$. For instance, for $p = 0.2$, we fix $M = \$6$ and vary H from \$6 to \$36. The left-hand option is again fixed: It offers $\$(p \cdot 30)$ for sure in the AB variant and an r chance of $\$(p \cdot 30)$ in the CD variant. The right-hand option starts with an outcome of $\$(p \cdot 30)$ and increases by \$1 per row. For each variant, we take the average value of H at the switching rows to be our measure of the realized indifference point, which we denote by h_{AB} or h_{CD}.

For each price list, we enforce a unique switching point—which naturally corresponds to our theoretical framework in which individuals reveal a unique indifference point. For convenience, when participants click a row in the left panel, it highlights the left-hand option in that row and all rows above. Analogously, when they click a row in the right panel, it highlights the right-hand option in that row and all rows below. They can adjust their choices as much as they want before submitting their final choices for that valuation. Online Appendix Figures C.1 to C.5 provide example screenshots of the m- and h-valuations for both the AB and CD variants.

We elicit valuations for five different probabilities, $p \in \{0.1, 0.2, 0.5, 0.8, 0.9\}$. For each p, participants complete the AB and CD variants for both the m- and h-valuation tasks. Participants therefore complete a total of 20 valuations in stage 1: five probabilities $(0.1, 0.2, 0.5, 0.8, 0.9)$ × two variants (AB and CD) × two types of valuation tasks (m and h). Thus, for each participant, we elicit the realized indifference points $(m_{AB}, m_{CD}, h_{AB}, h_{CD})$ for the five different levels of p. We randomize the order in which participants complete these 20 valuations. Hence, while the valuations are paired from our perspective, there is no obvious sense in which they are paired from the participants' perspective.[23]

3.2 Stage 2: Paired Choice Tasks

For each of the ten paired valuation tasks from stage 1, each participant faces a corresponding paired choice task that isolates one specific row from stage 1. Table 2 provides an overview of the ten AB binary choices that participants see in stage 2. For the m-choices (in panel A), the larger outcome is always $H = \$30$ as in the corresponding price list, and we randomly draw a value for M corresponding to a randomly selected row of the price list. Each participant sees one AB variant for each value of p; and, for each value of p, we randomly draw a value of M from the values listed in the table. For the h-choices, the smaller outcome is always $M = \$(p \cdot 30)$, and we randomly draw a value for H. Again, each participant sees one AB variant for each value of p, and for each, we randomly draw a value of H from the values listed in the table. We chose the possible values for M and H based on pilot data, with the aim that one extreme would yield a majority choosing lottery A while the other extreme would yield a majority choosing lottery B.[24]

For each of the ten AB variants shown in Table 2, the participant also sees the paired CD variant in which we hold fixed the values for M and H but scale down the probabilities by that participant's common ratio r. Thus, each participant makes a total of 20 binary choices: five probabilities $(0.1, 0.2, 0.5, 0.8, 0.9)$ × two variants (AB and CD) × two choice tasks (m and h). We randomize the order in which participants see the 20 binary choices and the relative position on the screen (left or right) of the two options within each binary choice. Online Appendix Figures C.6 to C.10 provide example screenshots of the m- and h-choices for both the AB and CD variants.[25]

[23] Given our use of multiple-price lists, a possible concern is that, for each list, participants are not providing an (effective) indifference point as our model assumes, but rather are choosing their switching rows based on the resulting overall compound lottery for the full list. Moreover, given our use of many tasks with only one chosen for payment, a second possible concern is that participants are not treating each task in isolation, but rather are assembling their preferred compound lottery for all tasks combined. Like most of the prior experimental literature—including the literature that estimates probability weighting from within-subject measures of certainty equivalents—we proceed assuming that neither of these is an issue.

[24] Our pre-analysis plan specified five values of M and H, but our implementation code had a small error and only implemented four of the five values in each case.

[25] Online Appendix B.5 provides an alternative way to visualize, within the context of our specific experimental tasks, the bias in paired choice tasks and how paired valuation tasks are immune to that bias.

Table 2: Summary of Binary Choices

p	(i)		(ii)
	Panel A. m-Choices		
0.1	100% chance of $M \in \{\$1, \$3, \$5, \$8\}$	or	10% chance of \$30
0.2	100% chance of $M \in \{\$1, \$4, \$7, \$10\}$	or	20% chance of \$30
0.5	100% chance of $M \in \{\$5, \$8, \$11, \$14\}$	or	50% chance of \$30
0.8	100% chance of $M \in \{\$8, \$12, \$16, \$20\}$	or	80% chance of \$30
0.9	100% chance of $M \in \{\$10, \$14, \$18, \$22\}$	or	90% chance of \$30
	Panel B. h-Choices		
0.1	100% chance of \$3	or	10% chance of $H \in \{\$30, \$25, \$20, \$13\}$
0.2	100% chance of \$6	or	20% chance of $H \in \{\$35, \$30, \$25, \$20\}$
0.5	100% chance of \$15	or	50% chance of $H \in \{\$45, \$40, \$35, \$30\}$
0.8	100% chance of \$24	or	80% chance of $H \in \{\$52, \$45, \$38, \$33\}$
0.9	100% chance of \$27	or	90% chance of $H \in \{\$54, \$47, \$40, \$35\}$

Note: Summary of all possible AB variants of the m- and h-choices. A participant faces one binary choice from each row, where we randomly draw a value of M for each row in panel A and a value of H for each row in panel B. The CD variant of each row keeps the same M and H values but scales all probabilities down by the participant's common ratio r.

3.3 Additional Design Details

Before beginning stage 1 of the experiment, participants complete an unincentivized attention check and quiz about the payment mechanism. After stages 1 and 2 of the experiment, participants complete two incentivized comprehension checks to gauge their understanding of the multiple-price-list format and the binary-choice tasks. The first comprehension check tests whether individuals can correctly fill out a price list given a specified indifference value. The second comprehension check tests whether participants can correctly answer a binary-choice question when given another person's responses to a price list. Online Appendix Figures C.11 and C.12 provide example screenshots of these comprehension checks.[26]

To break up the tasks and reduce fatigue, we present participants with an unincentivized visual puzzle after every fifth question in both stages of the experiment. Online Appendix Figure C.13 provides an example.

3.4 Recruiting

We recruited 900 participants through Prolific who had at least a high school education, were between the ages of 18 and 30, were living in the United States or Western Europe, and had a high approval rating on Prolific (see Online Appendix Table D.1 for summary statistics about

[26]For each comprehension check, roughly 85 percent of participants answer correctly. While our analysis uses the full sample, restricting the sample to those who answer both comprehension checks correctly does not materially change our results.

participants). We focused on this sample for comparison to prior common-ratio studies, the bulk of which have used undergraduate samples in the United States and Western Europe. We recruited an equal number of male and female participants.[27] The experiment took place in August 2021.

Participants received a \$5 payment upon completion. We also randomly selected one in five participants to receive an additional bonus payment based on their decisions in the study. Each of the 42 questions (20 valuations, 20 binary choices, and two incentivized comprehension checks) was equally likely to determine the bonus payment amount. If we randomly selected a valuation, then we randomly selected one row of the price list and paid the participant based on the option they selected in that row. If we randomly selected a binary choice, then we paid the participant based on the option they selected. If we randomly selected a comprehension check, then we paid the participant \$5 if they answered correctly. The experiment took 27 minutes to complete on average, and participants earned an average total payment of \$6.51.[28]

4 Analysis of Paired Valuation Tasks

Stage 1 of the experiment implements the paired valuation tasks needed to conduct our proposed valuations-based tests for a CRP. Our primary focus is an analysis of the m-valuations; however, we also use the h-valuations as a robustness check and to validate our approach.

4.1 Main Results

Figure 4 provides an initial visualization of our data for both m-valuations (in panel A) and h-valuations (in panel B). This figure is analogous to panel B of Figure 2; each dot denotes the mean valuations for the AB task and the CD task for a fixed (p, r). For paired m-valuation tasks, a CRE corresponds to $m_{CD} > m_{AB}$, and is therefore consistent with observations below the 45-degree line. For paired h-valuation tasks, a CRE corresponds to $h_{AB} > h_{CD}$, and thus, given the change in axes, is again consistent with observations below the 45-degree line. Figure 4 reveals little evidence of a systematic CRE.

More formally, we focus on m-valuations and conduct the two tests developed in Section 2.3 based on Proposition 2. For both, we focus on $\Delta m \equiv m_{CD} - m_{AB}$ at the individual level, and test the null of $\Delta m^* = 0$. Consider first the test based on the mean of Δm, which is valid under Assumption 2a. Columns (2) and (3) of Table 3 present the mean value of Δm along with the p-value for the corresponding means test. Out of the 15 means tests, we reject the null hypothesis of $\Delta m^* = 0$ in eight comparisons at the 5 percent level. All eight rejections indicate an RCRP rather than a CRP. Moreover, even the statistically significant means are relatively small in magnitude. (See Online Appendix Table D.2 for complete summary statistics on the m-valuations.)

[27] We did not preregister a gender-balanced sample. After preregistering, we learned that Prolific had very recently experienced a large increase in young female participants due to a social media trend. To better approximate the typical college population, we recruited 450 men and 450 women.

[28] This was double the \$6.50/hour minimum wage on Prolific.

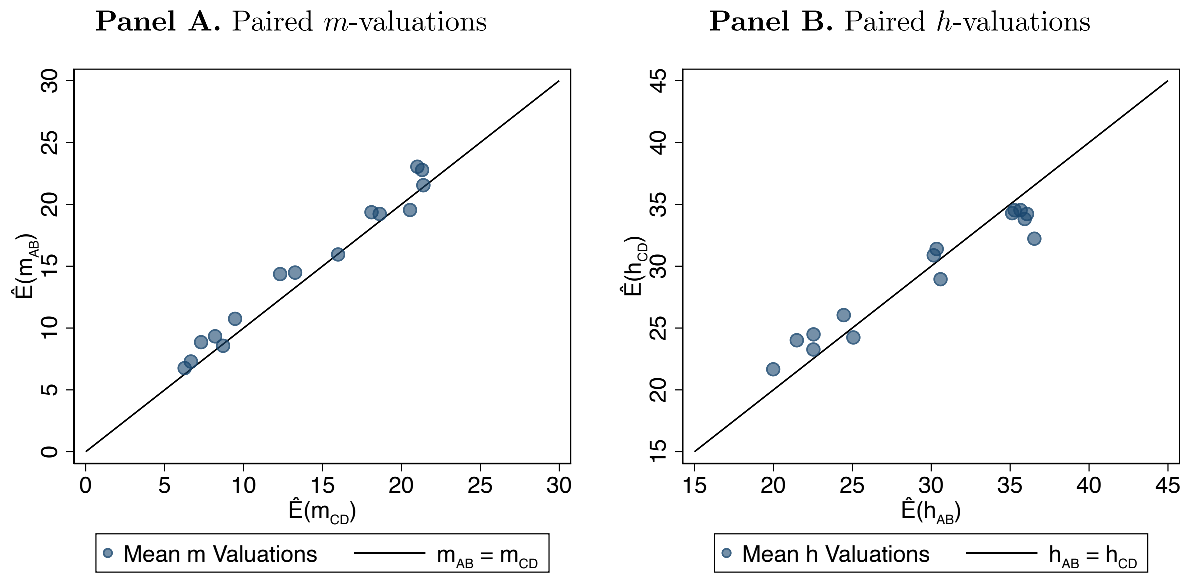

Figure 4: Mean valuations for each of the 15 (p, r) combinations for the paired m-valuation task (panel A) and the paired h-valuation task (panel B). In each panel, points below the 45-degree line are combinations that indicate a CRE, while points above the 45-degree line are combinations that indicate an RCRE.

As discussed in Section 2.3, the means test may be biased if the function Γ in Assumption 1 is nonlinear; hence, we also consider a test based on the sign of Δm, which is appropriate under Assumption 2b. Columns (4)–(6) report the raw frequency data: For each combination of r and p, the table reports the number of participants who exhibit $\Delta m > 0$ (consistent with CRP), $\Delta m < 0$ (consistent with RCRP), and $\Delta m = 0$. Column (7) reports the p-value from the two-sided sign test that we proposed in Section 2.3 (see footnote 18). Out of the 15 sign tests, we find seven significant deviations from the null of equal proportions at the 5 percent level. Six of these are consistent with an RCRP, and there is only one test in which the deviation from equal proportions is in the direction consistent with a CRP. Beyond the formal sign test, we also note that in 14 of 15 cases, the median value of Δm shown in column (8) is zero, indicating a strong central tendency toward $\Delta m = 0$.

That the sign test sometimes rejects the null of equal proportions even though the median of Δm is zero is partly because there are many observations of $\Delta m = 0$. In conducting the sign tests in Table 3, we adopt the conventional approach of ignoring these ties—that is, we exclude all $\Delta m = 0$ observations from our calculations. Including these ties can lead to changes in the p-values; however, there are multiple ways to incorporate ties, and there is no agreement in the literature about which is best (for discussions, see Coakley and Heise, 1996 and Randles, 2001). Here, we discuss two approaches that offer some bounds on our results. First, we could split the ties evenly between $\Delta m > 0$ and $\Delta m < 0$. This approach would increase all p-values and thus make it less likely that we reject the null of equal proportions. Second, we could split the ties between

Table 3: Testing the Null Hypothesis of $\Delta m^* = 0$

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
			Number of Cases				
Probability	Δm (Mean)	Means Test (p-value)	$\Delta m > 0$ (CRP)	$\Delta m = 0$	$\Delta m < 0$ ($RCRP$)	Sign Test (p-value)	Δm (Median)
			Panel A. $r = 0.2$				
0.1	−1.55	0.000	79	75	144†	0.000	0
0.2	−1.29	0.003	80	73	145†	0.000	0
0.5	0.04	0.932	123	60	115	0.650	0
0.8	1.00	0.052	140†	54	104	0.025	0
0.9	−1.47	0.014	127	42	129	0.950	0
			Panel B. $r = 0.4$				
0.1	−0.63	0.152	103	71	129	0.101	0
0.2	−1.14	0.003	97	65	141†	0.005	0
0.5	−1.22	0.007	104	62	137†	0.039	0
0.8	−0.60	0.262	127	41	135	0.665	0
0.9	−0.16	0.782	124	52	127	0.900	0
			Panel C. $r = 0.6$				
0.1	−0.49	0.158	94	90	115	0.166	0
0.2	0.14	0.692	111	84	104	0.682	0
0.5	−2.05	0.000	89	65	145†	0.000	0
0.8	−1.26	0.008	113	57	129	0.335	0
0.9	−2.03	0.000	79	60	160†	0.000	−1

Note: Means test and sign test for paired m-valuations for all 15 combinations of (p, r). Δm denotes the difference between the CD and AB m-valuations. We conduct a two-sided t-test for the difference in means. We also conduct a two-sided sign test, where we exclude all ties (instances of $\Delta m = 0$). A † indicates the larger group when the sign test rejects the null of equal proportions at the 5 percent level.

$\Delta m > 0$ and $\Delta m < 0$ using the same proportions we observe in the non-ties. This approach would decrease all the p-values and thus make it more likely that we reject the null of equal proportions. In Online Appendix Table D.3, we present the sign-test results using both approaches and show that the overall message is almost identical.

Result 1 *We find no evidence of systematic common ratio preferences.*

A possible challenge to Result 1 is that a price-list effect might mask an underlying CRP. Indeed, one of the most common criticisms of using multiple-price lists to elicit valuations is that valuations may get pulled to the center of the list. In principle, asymmetric pull-to-the-center effects could obfuscate an underlying CRP. For instance, let m^o denote the row to which these effects pull valuations. If $m^*_{AB} < m^*_{CD} < m^o$ (e.g., for low values of p) but the pull effect is stronger for the elicited m_{AB} than the m_{CD}, then we might observe $m_{AB} = m_{CD}$ despite there being an underlying CRP.

There are several reasons to believe such list effects do not drive Result 1. First and foremost, we document in Section 5.4 that we do not observe a systematic CRE in our paired choice tasks when we use a broader and more balanced set of parameters relative to the prior literature. In other words, our stage 2 choice data generate a conclusion akin to Result 1, although the impact of noise on paired choice tasks complicates inference. Second, there is no apparent reason for the pull-to-the-center effect to take the asymmetric form necessary to generate Result 1 despite an underlying CRP. In addition, for such asymmetric pull-to-the-center effects to mask an underlying CRP, the asymmetry would need to be in the opposite direction—i.e., stronger for the elicited m_{CD}—for high p, where an underlying CRP would manifest as $m^o < m^*_{AB} < m^*_{CD}$. There is no apparent reason to assume that this is the case. More generally, the strong link between participants' stage 1 valuations and their stage 2 choices that we document in Section 5 is further evidence against this (or some other) list effect obfuscating an underlying CRP.

A second possible challenge to Result 1 is whether the absence of a CRE in our valuation data is due to our choosing experimental parameter values for which models featuring a CRP would predict very small magnitudes for Δm. To assess this issue, we interpret our data within the prospect theory (PT) structure from Example 1. Figure 5 presents the mean m-valuations in a different way from Figure 4(A): For each of the five values of p, blue dots denote the mean m_{AB} valuations and red diamonds denote the mean m_{CD} valuations. The three panels separate results by the three different values for r. We then use the fact that our AB valuation tasks are identical to those that researchers frequently use to estimate a PT probability weighting function. Following that literature, we use participants' five m_{AB} values to estimate a probability weighting function for each value of r. Online Appendix E.1 provides the details of this structural estimation and reports the parameter estimates; the AB valuations that these estimates predict are depicted in Figure 5 by the dashed blue lines. In each case, the estimated probability weighting function takes the familiar inverse-S shape that the prior literature has typically found—e.g., the dashed blue lines all look quite similar to Figure 2 in the Barberis (2013) review of the prospect-theory literature.[29]

We next use the probability weighting functions estimated from the AB valuations to predict the CD valuations. The red dashed-and-dotted lines in Figure 5 depict these predictions. The difference between the two lines reflects the predicted magnitude of CRP, Δm, given the observed AB valuations. Figure 5 reveals that, under PT, we should see substantial positive values of Δm; across the 15 combinations of p and r, the predicted Δm ranges from 3.15 to 8.63, with an average of 6.30.[30] Hence, we observe small Δm values in the data *despite* the fact that PT predicts much

[29] In Online Appendix E.1, we use the functional forms from (Tversky and Kahneman, 1992) and corresponding parameter labels, and the typical non-linear least-squares approach to estimate those parameters. The specific estimates for utility curvature (α) and probability weighting (γ) are: $\alpha = 1.351$ and $\gamma = 0.580$ for $r = 0.2$; $\alpha = 1.179$ and $\gamma = 0.587$ for $r = 0.4$; and $\alpha = 1.112$ and $\gamma = 0.636$ for $r = 0.6$.

[30] Online Appendix Table D.4 reports the complete range of predicted Δm values and confidence intervals for each (p, r) combination.

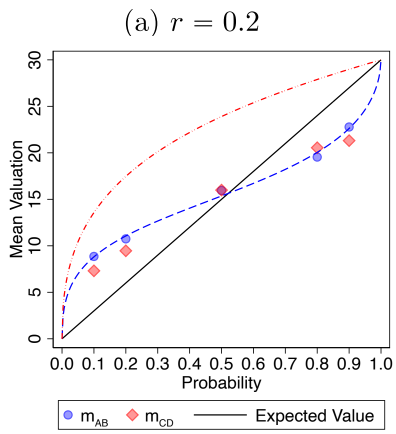

Figure 5: Mean valuations by probability p and common ratio r. The blue dots denote mean AB valuations, and the red diamonds denote mean CD valuations. The dashed blue lines represent the AB valuations predicted by PT with parameters estimated from the AB-valuation data. The dashed-and-dotted red lines denote the CD valuations predicted by PT given the parameter estimates represented by the blue lines.

larger differences.

A further implication of Figure 5 is that our data are inconsistent with models of probability weighting in the tradition of Kahneman and Tversky's (1979, 1992) prospect theory. In such models, a CRP (or subproportionality) is perhaps the central motivating fact behind the structure of probability weighting. Our failure to find a systematic CRP is thus problematic for such models. Indeed, our discussion above highlights how a model in which individuals apply a single probability weighting function to both types of comparisons cannot explain our combined m_{AB} and m_{CD} data.[31]

4.2 Validation and Robustness

Result 1 stands in stark contrast to the widely accepted belief that there exists a systematic CRP. Hence, it is important to validate that our valuation measures reflect underlying preferences, and that our finding of $\Delta m \approx 0$ is not an artifact of our experimental task. We outline several features of our data that alleviate this concern.

First, we reiterate that our m_{AB} valuation tasks are identical to the valuation tasks used in the extensive literature that estimates probability weighting functions, and our data yield probability weighting functions that look very similar to those estimated in the literature (as seen in Figure 5 and discussed in footnote 29). It is therefore reassuring that part of our task aligns so closely with the broader literature on probability weighting, and supports our use of valuation data to test for CRP.

[31] For a more formal test, see Online Appendix E.1, where we estimate a structural model that permits separate probability weighting functions for the AB and CD valuations, and strongly reject the null of there being no difference.

Second, participants' valuations respond sensibly to changes in p and hence the expected value of the lottery: Increases in p lead participants to report higher valuations. We see this pattern even though we present 20 tasks in random order that differ along multiple dimensions: an m- or h-valuation, an AB or CD variant, and the value of p.

Third, we designed our h-valuation tasks for both validation and robustness. In terms of robustness, the h-valuation tasks present a different choice structure to participants, yet permit analogous means and sign tests. Table 4 reports the results for the h-valuation tasks, and they are similar to the results for the m-valuation tasks in Table 3 (see Online Appendix Table D.5 for complete summary statistics on the h-valuations). While the mean Δh is often significantly different from zero, and the magnitudes are slightly larger than those for the m-valuation tasks, there are roughly equal numbers of positive and negative instances of Δh. Out of the 15 sign tests, we find nine significant deviations from the null of equal proportions at the 5 percent level: Five are consistent with an RCRP, and four are consistent with a CRP.[32] Finally, the median Δh is zero in 12 of the 15 cases. Hence, Table 4 reaffirms Result 1: There is no evidence of a systematic CRP in valuations.[33]

We also use the h-valuation data to validate the m-valuation data. As described in Section 2.5, the two valuations approximately measure the same preference, and thus the proportional risk premium m_z/H for an m-valuation should be strongly correlated with the proportional risk premium M/h_z for the corresponding h-valuation. In Online Appendix Table D.7, we report the rank correlations between m_{AB}/H and M/h_{AB} and the rank correlations between m_{CD}/H and M/h_{CD} for each of the 15 combinations of (p, r). All 30 rank correlations are significantly positive at the 5 percent level, and the average is 0.28. Overall, these positive correlations confirm that our valuations capture meaningful information about underlying preferences and are not merely driven by some artifact of our experimental task.[34]

4.3 Heterogeneous Preferences and Noise in Stage 1 Data

Our aggregate tests provide no evidence of a systematic CRP. However, we also observe significant variation in participants' stage 1 responses. This variation could merely reflect the impact of choice noise; indeed, the premise of our analysis is that choice noise can generate idiosyncratic variation in responses around underlying values. However, this variation could also be due to heterogeneity in participants' underlying preferences. There are two relevant forms of heterogeneity in preferences

[32]Online Appendix Table D.6 shows that the message is almost identical no matter how we treat observations of $\Delta h = 0$.

[33]Castillo and Eil (2014) propose and test a theory of status quo bias in which holding fixed the safe option in a paired valuation task (as in our h valuations) leads to an RCRP, while holding fixed the risky option (as in our m-valuations) yields a CRP. While they find limited support in their data, the consistent pattern that we find across our m- and h-valuations runs counter to their theory.

[34]Here and in subsequent sections, we provide interpretations of correlations in behavior across different tasks. These interpretations implicitly assume that any correlation in noise draws across tasks is limited; the interpretations would be different if there were sizeable correlations in noise draws across tasks.

Table 4: Testing the Null Hypothesis of $\Delta h^* = 0$

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
			Number of Cases				
Probability	Δh (Mean)	Mean Test (p-value)	$\Delta h > 0$ (CRP)	$\Delta h = 0$	$\Delta h < 0$ ($RCRP$)	Sign Test (p-value)	Δh (Median)
			Panel A. $r = 0.2$				
0.1	−1.67	0.006	100	60	138†	0.016	0
0.2	−1.94	0.001	94	53	151†	0.000	−1
0.5	1.64	0.001	136†	81	81	0.000	0
0.8	4.31	0.000	174†	45	79	0.000	3
0.9	2.11	0.000	143†	64	91	0.001	0
			Panel B. $r = 0.4$				
0.1	−2.53	0.000	82	59	162†	0.000	−1
0.2	−1.59	0.002	92	65	146†	0.001	0
0.5	−1.05	0.036	101	70	132†	0.049	0
0.8	1.84	0.002	148†	47	108	0.015	0
0.9	1.13	0.055	138	47	118	0.235	0
			Panel C. $r = 0.6$				
0.1	−0.73	0.192	100	71	128	0.074	0
0.2	0.83	0.108	131	65	103	0.077	0
0.5	−0.72	0.130	93	85	121	0.065	0
0.8	0.84	0.146	136	47	116	0.231	0
0.9	0.76	0.173	126	54	119	0.702	0

Note: Means test and sign test for paired h-valuations for all 15 combinations of (p, r). Δh denotes the difference between the AB and CD h-valuations. We conduct a two-sided t-test for the difference in means. We also conduct a two-sided sign test, where we exclude all ties (instances of $\Delta h = 0$). A † indicates the larger group when the sign test rejects the null of equal proportions at the 5 percent level.

that we investigate. First, there is heterogeneity in the degree of risk aversion, as reflected in the levels of participants' m^*_{AB} and m^*_{CD}. Second, there is heterogeneity in the degree to which people have an underlying CRP or RCRP, as reflected by heterogeneity in Δm^*.

We begin by assessing the degree of heterogeneity in risk aversion. For each value of p, our stage 1 data contain two valuations that reflect an individual's risk aversion: m_{AB} and m_{CD}. As a benchmark, risk neutrality would lead a person to state $m_{AB} = m_{CD} = pH$. Hence, we define $\bar{m} \equiv (m_{AB} + m_{CD})/2$ and then use $pH - \bar{m}$ as a measure of a person's risk aversion. With this measure, positive values reflect risk aversion and negative values reflect risk seeking.

Panel A of Figure 6 presents the distribution of the observed $pH - \bar{m}$ across all 15 combinations of (p, r). There is substantial variation in the magnitude of $pH - \bar{m}$. While some of this variation is surely due to noise, there is also evidence that some of it is due to heterogeneity in participants' underlying $\bar{m}^*$. First, in Section 4.2, we highlighted the strong rank correlations between participants' m_x/H and M/h_x for the same (p, r), which indicate the existence of heterogeneity in underlying preferences. Second, panel A of Online Appendix Table D.8 documents substantial rank correlations of $pH - \bar{m}$ across different values of p: Across the 30 possible combinations of (p, p', r), all rank correlations are positive, ranging from 0.06 to 0.61, with an average of 0.31.[35]

Panel A. Risk Premia **Panel B.** Value Differences

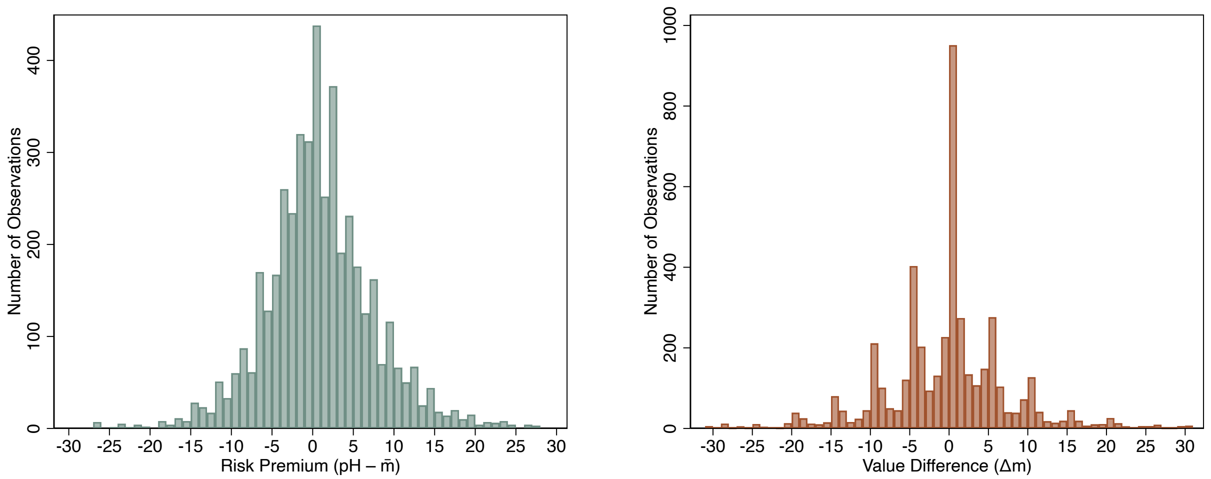

Figure 6: Distributions of risk premia (panel A) and value differences (panel B) in m-valuations. The risk premium is $pH - \bar{m}$ where $\bar{m} = (m_{AB} + m_{CD})/2$ and $H = 30$. Positive risk premia indicate risk aversion, and negative risk premia indicate risk seeking. The value difference is $\Delta m = m_{CD} - m_{AB}$. Positive value differences indicate CRP, and negative value differences indicate RCRP. For each panel, the data include all 15 combinations of (p, r).

We next assess the degree of heterogeneity in underlying CRP versus RCRP. Panel B of Figure 6 presents the distribution of the observed Δm across all 15 combinations of (p, r). Again, there is substantial variation, and while some of it is undoubtedly due to noise, there is also evidence that some of it is due to heterogeneity in participants' underlying Δm^*. First, analogous to our assessment of the rank correlations between m_x/H and M/h_x for the same (p, r), we study rank correlations between the differences $m_{CD}/H - m_{AB}/H$ and $M/h_{CD} - M/h_{AB}$. Online Appendix Table D.9 reports this rank correlation for all 15 combinations of (p, r). Twelve are significantly positive at the 5 percent level, and the average across all 15 rank correlations is 0.15. Second, panel B of Online Appendix Table D.8 documents substantial rank correlations of Δm across different values of p: Across the 30 possible combinations of (p, p', r), ten are significantly positive at the 5 percent level, and none are significantly negative. The average across all 30 rank correlations is 0.11.[36]

[35] These correlations are weakest when comparing $pH - \bar{m}$ for a high p (0.8 or 0.9) with that for a low p (0.1 or 0.2), which is perhaps not surprising given that some models can predict a negative correlation for such comparisons. For instance, under PT with heterogeneity in the extent of probability weighting, people with more pronounced probability weighting (smaller γ) would have a smaller (more negative) $pH - \bar{m}$ for low p and a larger (more positive) $pH - \bar{m}$ for high p.

[36] Dean and Ortoleva (2019) use paired valuation tasks to obtain a continuous measure of the CRE. They find that

Hence, while we fail to find a systematic CRP in the aggregate, the variation in participants' stage 1 valuations reflects a combination of idiosyncratic noise and systematic heterogeneity in underlying preferences. Heterogeneity in both the levels of risk attitudes and underlying CRP or RCRP drives the positive correlations observed between measures. Choice noise attenuates these correlations toward zero, even for values that should measure the same construct. Recognizing both heterogeneity in preferences and a substantial role for noise will have important implications for our stage 2 paired choice tasks. We turn to these connections next.

5 Analysis of Paired Choice Tasks

This section analyzes empirical results from our paired choice tasks, making explicit the connections between the valuations elicited in stage 1 and the corresponding choices made in stage 2. Doing so allows us to assess whether there is differential noise across the AB and CD choices and to reconcile our main finding of no systematic CRP in paired valuation tasks with the vast literature that finds a CRE in paired choice tasks.

5.1 Connections Between Paired Valuations and Paired Choices

Under Assumption 1 from Section 2.3, the same underlying preferences drive behavior for both paired valuation tasks and paired choice tasks, and thus there should be a strong connection between the two. To illustrate, consider the special case of Assumption 2a where a person with underlying indifference values (m^*_{AB}, m^*_{CD}) would choose A over B when $M \geqslant m^*_{AB} + \varepsilon_{AB}$ and would choose C over D when $M \geqslant m^*_{CD} + \varepsilon_{CD}$, where $\varepsilon_{CD} \overset{d}{=} k\varepsilon_{AB}$. For this case, the probability of making CRE choices (A and D) minus the probability of making RCRE choices (B and C) is

$$\begin{aligned} CRE - RCRE \equiv \Pr(A) - \Pr(C) &= \Pr\left(\varepsilon_{AB} < (M - m^*_{AB})\right) - \Pr\left(\varepsilon_{CD} < (M - m^*_{CD})\right) \\ &= \Pr\left(\varepsilon_{AB} < (M - m^*_{AB})\right) - \Pr\left(\varepsilon_{AB} < \frac{1}{k}(M - m^*_{CD})\right). \end{aligned}$$

Defining $\Psi \equiv (M - m^*_{CD})/k$, and substituting $\bar{m}^* \equiv (m^*_{AB} + m^*_{CD})/2$ and $\Delta m^* \equiv m^*_{CD} - m^*_{AB}$, we can rewrite this as:

$$CRE - RCRE = \Pr\left(\varepsilon_{AB} < \Psi + 0.5\left(1 + \frac{1}{k}\right)\Delta m^* + \left(1 - \frac{1}{k}\right)(M - \bar{m}^*)\right) - \Pr\left(\varepsilon_{AB} < \Psi\right). \quad (3)$$

This formulation links the extent of $CRE - RCRE$ in a paired choice task to two terms: (i) a *scaled value difference* term $0.5(1 + 1/k)\Delta m^*$, and (ii) a *scaled distance to indifference* term $(1 - 1/k)(M - \bar{m}^*)$. To illustrate the implications of equation (3), Figure 7 presents the predicted $CRE - RCRE$ as a function of the value difference term (panel A) and the distance to indifference

their valuations-based measure of the CRE is strongly correlated with other risk-preference-related behaviors, which parallels our finding of reliable heterogeneity in underlying CRP versus RCRP.

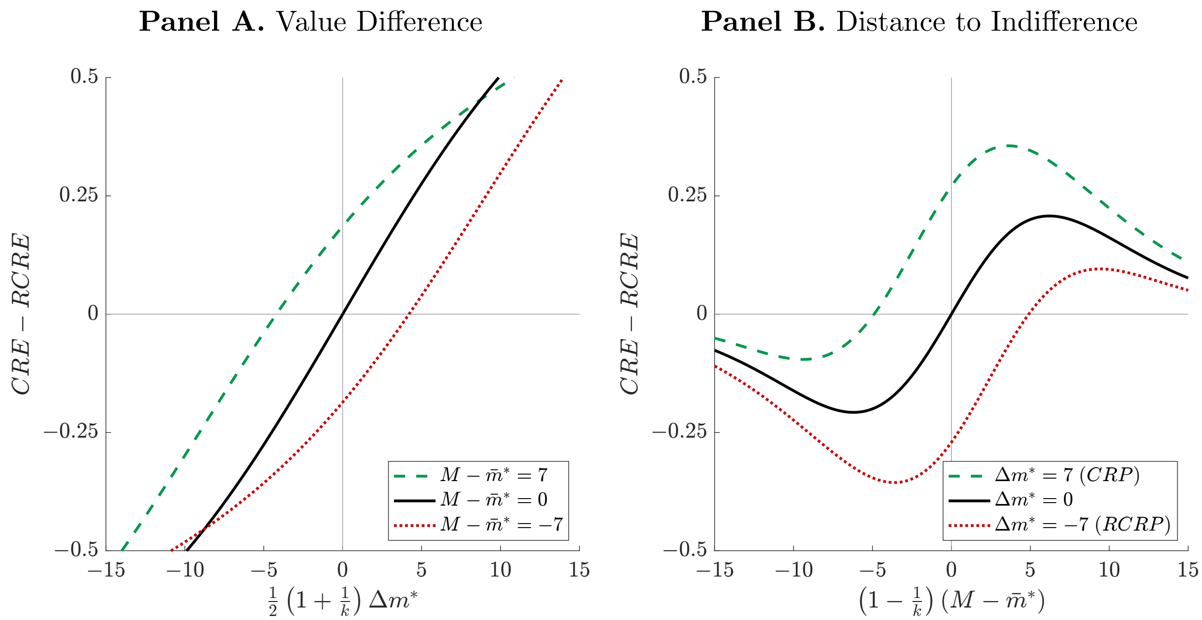

Figure 7: Predicted $CRE - RCRE$ as a function of the value difference $0.5(1 + 1/k)\Delta m^*$ (panel A) or distance to indifference $(1-1/k)(M-\bar{m}^*)$ (panel B). Panel A depicts predicted $CRE-RCRE$ as a function of value difference for three cases: (i) a positive distance to indifference (green dashed), (ii) a negative distance to indifference (red dotted), and (iii) a zero distance to indifference (black solid). Panel B depicts predicted $CRE - RCRE$ as a function of distance to indifference for three cases: (i) a positive (CRP) value difference (green dashed), (ii) a negative ($RCRP$) value difference (red dotted), and (iii) a zero value difference (black solid). Both panels assume $\varepsilon_{AB} \sim N(0, 7^2)$ and $k = 2.5$.

term (panel B) when ε_{AB} is distributed $N(0, 7^2)$ and $k = 2.5$, so that the noise is more impactful for the CD choice.

The value difference captures the strength of a person's underlying CRP (when $\Delta m^* > 0$) or RCRP (when $\Delta m^* < 0$). The impact of the value difference on $CRE - RCRE$ seen in panel A therefore reflects the impact of underlying preferences on choice-based measures of the CRE. Indeed, in taking empirically observed instances of $CRE - RCRE > 0$ in paired choice tasks as evidence of an underlying CRP, the prior literature has effectively assumed that this effect is the dominant effect. Panel A highlights that the relationship between $CRE - RCRE$ and the value difference is straightforward: The larger the value difference, the larger $CRE - RCRE$.

The distance to indifference has its most natural interpretation under EU, when $m^*_{AB} = m^*_{CD} \equiv m^*$ and thus $\bar{m}^* = m^*$. Under EU, the magnitude of the distance to indifference captures the strength of a person's underlying EU preference for A and C (when $M - m^* > 0$) or for B and D (when $M - m^* < 0$). The solid line in panel B illustrates the predictions highlighted in Section 2.2: When EU preferences favor A and C, noise generates a prediction of $CRE - RCRE > 0$, and when EU preferences favor B and D, noise generates a prediction of $CRE - RCRE < 0$. The dashed and dotted lines in panel B illustrate that the same pattern holds even for people with $\Delta m^* \neq 0$, where equation 3 establishes that $M-\bar{m}^*$ is the natural analogue for $M-m^*$. Hence, an individual with an underlying RCRP may exhibit a CRE if the offered M implies a large and positive distance to indifference, and an individual with an underlying CRP may exhibit an RCRE if the offered M implies a large and negative distance to indifference.[37]

The scaling factors in equation (3) depend on k, which captures the extent of differential noise across AB versus CD choices. When $k=1$, there is no differential noise impact, and the scaled distance to indifference term disappears—that is, the person will exhibit a CRE if and only if they have an underlying CRP. In contrast, when there is differential noise across the AB and CD choices ($k \neq 1$), the scaled distance to indifference will impact whether people exhibit a CRE or an RCRE. Our empirical analysis in Sections 5.2 and 5.3 will provide clear evidence that the distance to indifference has a positive impact on $CRE-RCRE$, implying that $k>1$ (as depicted in Figure 7). In other words, we find evidence that noise has a larger impact on the CD choices than on the AB choices.

In the following subsections, we test these predictions relating the value difference and the distance to indifference to $CRE-RCRE$. We emphasize three points as we transition from theoretical predictions to empirical results. First, we have framed the predictions above in terms of a person's underlying value difference Δm^* and distance to indifference $\bar{m}^*$, both of which are unobserved. In our empirical analysis, we replace these quantities with their empirical counterparts Δm and $\bar{m}$, recognizing that there is measurement error and attempting to correct for it when possible.

Second, our empirical analysis combines data for the three different values of r to increase the power of our statistical tests. However, r may impact the extent of differential noise (i.e., k)—e.g., EU with additive utility noise implies $k=1/r$. Moreover, a change in k would change the slopes in Figure 7, where a k closer to one would yield that Figure 7(A) is steeper while Figure 7(B) is flatter (in the range around $M-\bar{m}^*=0$). To account for this, and motivated by the EU case, our empirical analysis uses $0.5(1+r)\Delta m$ for the scaled value difference and $(1-r)(M-\bar{m})$ for the scaled distance to indifference.[38]

Finally, when we connect people's stage 1 valuations to their stage 2 choices, we assume that both reflect the same underlying preferences as stated in Assumption 1. However, we do not impose that the error distributions for valuations are the same as those for choices. Our analysis considers stage 1 valuations as noisy measures of underlying preferences, which motivates an instrumental variables approach. However, this approach does not require consistency between the noise in

[37] Online Appendix B.6 explores the impact of distance to indifference in the absence of choice noise, that is, when all variation in the data is due to heterogeneity in preferences. Two key predictions are: (i) If the population distribution of Δm^* is symmetric, then the distance to indifference has no impact on $CRE-RCRE$; and (ii) otherwise, the distance to indifference can have an impact on $CRE-RCRE$, but that impact must be symmetric around a zero distance to indifference. Both of these predictions are inconsistent with our data.

[38] Online Appendix B.4 derives analogous corrected regressors for h-tasks. Basing these corrections on the EU case is not perfect, as we have already documented that our data are inconsistent with that case. However, using that case as motivation, instead of choosing a correction in a more ad hoc way, imposes some discipline on our analysis. In Online Appendix B.7, we assess this approach and show that it works as intended.

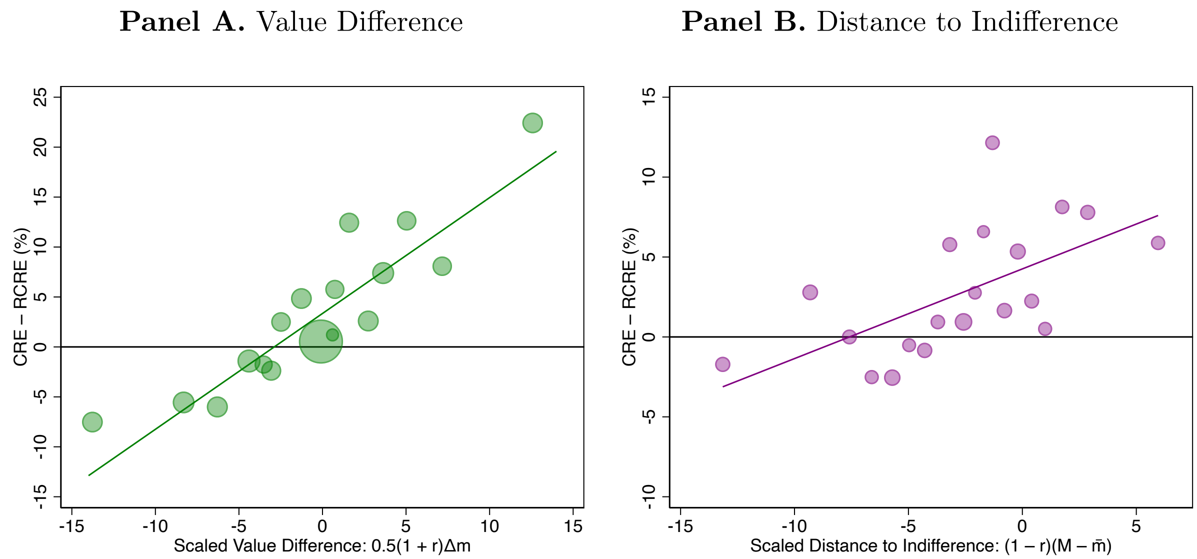

Figure 8: Stage 2 individual-level $CRE-RCRE$ as a function of individual-level stage 1 value difference (panel A) and distance to indifference (panel B). For each, we divide the stage 1 individual-level data into 20 equally sized bins, and the dots in each graph denote the average $CRE-RCRE$ for a given bin (we join four value difference bins together at zero due to the large number of observations at this point). The solid lines represent the best-fitting linear prediction for $CRE-RCRE$ as a function of the value difference (panel A) or the distance to indifference (panel B).

valuations and noise in the choices.

5.2 Individual-Level Stage 2 Data

We first analyze stage 2 data at the individual level. As in Section 4, we focus on the paired m-choice tasks and use the paired h-choice tasks as a robustness check. There are 60 different paired m-choice tasks, which reflect different combinations of p, r, and M. Each participant faces five of these, one for each value of p.

Consider a graphical analysis of stage 2 behavior analogous to the theoretical predictions in Figure 7. Figure 8 plots the relationship between observed $CRE-RCRE$ and the scaled value difference (panel A) and the scaled distance to indifference (panel B). We create 20 equally sized bins of the value difference and the distance to indifference, respectively, and report the average $CRE-RCRE$ within each bin.

Panel A of Figure 8 shows a tight connection between individuals' stage 1 value differences and their stage 2 choices: Consistent with value differences capturing people's underlying CRP or RCRP, individuals with more positive value differences are more likely to exhibit a CRE, and those with more negative value differences are more likely to exhibit an RCRE. Panel B shows that there is also a clear relationship between an individual's distance to indifference and whether we observe a CRE or an RCRE. As predicted by models with noise that is more impactful for the CD choice, individuals with greater distances to indifference are more likely to exhibit a CRE relative to an RCRE.

To provide a quantitative assessment of these relationships, Table 5 reports the results from linear regressions using the outcome variable $CRE - RCRE \in \{-1, 0, 1\}$ at the individual level. The regressions control for the values of p and r associated with the stage 2 paired choice task and individual characteristics. The regressors of interest are the scaled value difference and the scaled distance to indifference. Columns (1) and (2) include each regressor by itself, while column (3) includes both together. The coefficient estimates are quite stable across columns. They imply that a \$10 increase in the scaled value difference implies an 11 percentage-point increase in $CRE-RCRE$. At the same time, a \$10 increase in the scaled distance to indifference implies an eight percentage-point increase in $CRE-RCRE$. Moreover, these numbers imply that a person with a scaled value difference of $-\$7$ (i.e., a strong RCRP) would exhibit a CRE if their scaled distance to indifference were larger than \$10. These results show that distance to indifference, through its interaction with choice noise, plays an important role in whether people exhibit a CRE or an RCRE in their stage 2 behavior.[39]

Table 5: Predicting Individual-Level $CRE - RCRE$

	(1)	(2)	(3)	(4)
	Outcome: $CRE - RCRE \in \{-1, 0, 1\}$			
	OLS	OLS	OLS	2SLS
Scaled Value Difference:				
$\frac{1+r}{2}\Delta m$	1.12		1.14	6.98
	(0.14)		(0.14)	(1.07)
Scaled Distance to Indifference:				
$(1-r)(M-\bar{m})$		0.78	0.82	0.83
		(0.19)	(0.19)	(0.40)
Outcome Mean	2.64	2.64	2.64	2.64
Individuals	900	900	900	900
Observations	4,500	4,500	4,500	4,500

Note: OLS regressions using individual-level m-task data with dependent variable $CRE-RCRE \in \{-1, 0, 1\}$. Specifications include p and r fixed effects, as well as controls for gender, education, age, language, student status, employment, and the number of previous Prolific approvals. All numbers reported in percentage points; individual-cluster-robust standard errors in parentheses. For column (4), instruments are $(1-r)p\bar{h}$, $0.5(1+r)p\Delta h$, and $(1-r)M$.

Since our stage 1 values for m_{AB}, m_{CD}, and Δm reflect a combination of preference and noise, the coefficients in Table 5 might be attenuated by measurement error. To account for this, column (4) of Table 5 pursues an instrumental-variables approach using the h-valuation tasks and the

[39] The magnitude of distance to indifference also predicts stage 2 decision timing. When measured distance to indifference $|M - m_{XY}| = 0$, participants take around 7 seconds to complete their stage 2 choice. For every \$10 increase in $|M - m_{XY}|$, stage 2 choices occur around 1.4 seconds ($\approx$ 20 percent) faster. These results are consistent with the interpretation that easier choices occur faster, as suggested by some neuroscience models.

exogenous value $(1-r)M$ as instruments. The coefficient on the scaled distance to indifference is unaffected, which is not surprising given that a large share of the variation in this variable is due to our random variation in M. In contrast, the coefficient on the scaled value difference is substantially larger, which again is not surprising given that the variation in this variable is entirely endogenous and thus strongly influenced by choice noise. While it is difficult to compare the magnitudes of the two effects under the IV specification, the latter result provides further support for our conclusion that the stage 1 valuations include a substantial preference component, especially when using information across both the m- and h-valuation tasks.[40]

5.3 Experiment-Level Stage 2 Data

We next analyze stage 2 data at the experiment level—that is, we analyze the share of CRE choices minus the share of RCRE choices among participants who saw the same pair of choice tasks. As described in Section 5.2, there are 60 combinations of (M, p, r) linked to stage 1 m-valuations, which yields 60 experiments. To expand the number of experiment-level observations, we also include the 60 analogous experiments linked to stage 1 h-valuations—that is, for 60 combinations of (H, p, r). Each participant completed ten of these 120 experiments, and each experiment had between 57 and 101 observations with an average sample size across experiments of 75. We present the summary data for all 120 experiments in Online Appendix Tables D.11 and D.12.

Our goal is to demonstrate that we can predict which experiments will likely yield a CRE or an RCRE in choices based on our linked stage 1 data on valuations. Specifically, we calculate for each experiment (i) the average scaled value difference ($0.5(1+r)\Delta m$ for m-choice tasks or $0.5(1+r)p\Delta h$ for h-choice tasks) and (ii) the average scaled distance to indifference ($(1-r)(M-\bar{m})$ or $(1-r)p(\bar{h}-H)$).[41] We then investigate whether these averages predict $CRE - RCRE$ at the experiment level.

Figure 9 provides visual evidence and is the experiment-level analogue to the individual-level analysis in Figure 8. Panel A presents the relationship between $CRE - RCRE$ and the scaled value difference, where the variables on both axes are residualized by the distance to indifference. Panel B presents the relationship between $CRE - RCRE$ and the scaled distance to indifference, where the variables on both axes are residualized by the value difference. In both cases, each point is one experiment, and the size of each point is proportional to the sample size of each experiment. Similar to the individual-level data, we see that both variables predict whether we observe a CRE or an RCRE.

In Table 6, we re-conduct the analysis of Table 5 at the experiment level. The dependent variable is now the continuous variable $CRE - RCRE$ for each experiment, and the regressions all control for the values of p and r along with whether the experiment involves m or h choices.

[40] As a robustness check, we can perform an analogous individual-level analysis of the links between the stage 1 h-valuations and the stage 2 h-choices. Online Appendix Table D.10 is the h-task analogue to Table 5 and yields similar conclusions.

[41] See Online Appendix B.4 for a discussion the terms we use for h-choice tasks.

Panel A. Value Difference　　　　**Panel B.** Distance to Indifference

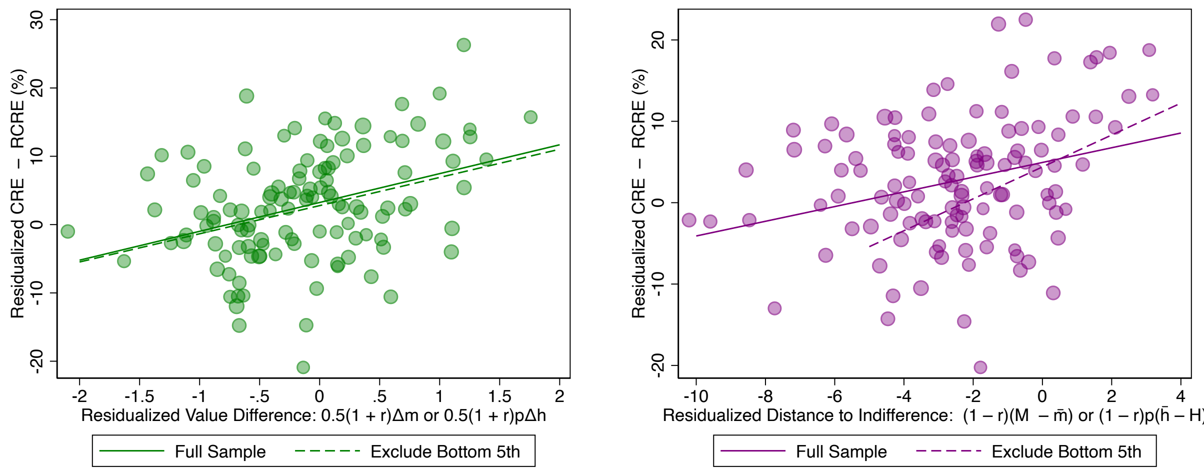

Figure 9: Stage 2 experiment-level $CRE - RCRE$ as a function of experiment-level stage 1 value differences (panel A) and distance to indifference (panel B). Each dot denotes an experiment, and the size is proportional to the number of observations in that experiment. Panel A observations are residualized by the average distance to indifference; panel B observations are residualized by the average value difference. The solid and dashed lines represent the best fitting OLS estimates from columns (3) and (4) of Table 6; the latter excludes experiments with values in the bottom quintile of the distance to indifference variable.

The regressors of interest are the average scaled value difference and the average scaled distance to indifference. Columns (1) and (2) include each by themselves, while column (3) includes both together; once again, the coefficients are quite stable across columns.

Table 6: Predicting Experiment-Level $CRE - RCRE$

	(1)	(2)	(3)	(4)
	Outcome: $CRE - RCRE \in [-100, 100]$			
Scaled Value Difference:				
$\frac{1+r}{2}\Delta m$ or $p\frac{1+r}{2}\Delta h$	3.94		4.22	4.13
	(1.06)		(1.02)	(1.15)
Scaled Distance to Indifference:				
$(1-r)(M-\bar{m})$ or $p(1-r)(\bar{h}-H)$		0.80	0.90	1.96
		(0.30)	(0.28)	(0.53)
Mean CRE − RCRE	2.69	2.69	2.69	1.92
Data Exclusion: Distance to Indifference	No	No	No	Bottom Quintile
Number of Experiments	120	120	120	96

Note: OLS regressions using experiment-level data with dependent variable $CRE - RCRE \in [-100, 100]$. Specifications include p and r fixed effects, as well as task (m versus h) fixed effects. All numbers reported in percentage points. For column (4), data restricted to experiments with an average distance-to-indifference value in the top four quintiles.

Across the 120 experiments, the average value difference has a range of about \$4. The estimates suggest that this range corresponds to a roughly 16 percentage-point change in the aggregate $CRE - RCRE$. Across the 120 experiments, the average distance to indifference has a range of about \$15. This range corresponds to a roughly 14 percentage-point change in $CRE - RCRE$. Hence, as in our individual-level analysis, we see that, in practice, the distance to indifference has an impact similar in magnitude to that of the value difference. Thus the distance to indifference—which is largely determined by the experimenter's choice of parameter values—plays an essential role in determining whether an experiment will have a CRE or an RCRE.

Panel B of Figure 9 reveals that our experimental variation in M and H leads to some large negative values of the average distance to indifference. Given that Figure 7 implies that the relationship between $CRE - RCRE$ and distance to indifference could invert for sufficiently large magnitudes, it is natural to exclude observations with extreme magnitudes. In column (4) of Table 6, we re-run the regression after dropping the bottom 20 percent of observations in terms of distance to indifference. The effect of distance to indifference on $CRE - RCRE$ within this sub-sample more than doubles. Hence, our numbers above perhaps understate the importance of distance to indifference in generating observed CRE or an RCRE.

Result 2 *Both value difference and distance to indifference significantly predict whether standard paired choice tasks will reveal a common ratio effect.*

5.4 Stage 2 Paired Choices Relative to the Prior Literature

In Section 2.4, we highlighted how the prior literature has used a limited set of experimental parameters. In contrast, we use a broader and more balanced set of experimental parameters in our stage 2 paired-choice experiments. We now investigate the implications of doing so, further highlighting the inferential challenges inherent to paired-choice data.

In fact, our stage 2 paired-choice data yield the same broad conclusion as our stage 1 paired-valuations data: There is little evidence of a systematic CRP. Looking across all 9000 observations in our data at the individual level, we observe $\widehat{\Pr}(A) = 49.14$ percent and $\widehat{\Pr}(C) = 46.46$ percent.[42] Hence, we observe that $CRE - RCRE = 2.7$ percent, which is substantially smaller than the (sample-weighted) 22.0 percent observed across all prior studies. At the experiment level, if we follow Blavatskyy et al. (2023) and use a Conlisk z-test with a 5 percent significance level, then, among our 120 experiments, 16.7 percent find a CRE while 9.2 percent find an RCRE, which is substantially more balanced than the 57.3 percent and 9.1 percent among prior experiments. Online Appendix Figure D.1 provides a visualization of the latter point by depicting the analogue of Figure 2(A) for our 120 experiments. Hence, based on our stage 2 paired-choice data, at best one might conclude that there is evidence for a mild underlying CRP.[43]

[42] The observed proportions of the four combinations are $\widehat{\Pr}(BD) = 38.0$ percent, $\widehat{\Pr}(AC) = 33.6$ percent, $\widehat{\Pr}(AD) = 15.6$ percent, and $\widehat{\Pr}(BC) = 12.9$ percent.

[43] Our stage 2 choice data also yield the same broad conclusion as our stage 1 valuations data when interpreted

Moreover, we can document that the different patterns in our experiments relative to prior experiments derive from our use of a broader and more balanced set of experimental parameters (see Online Appendix B.8 for full details). Specifically, we develop a measure of whether an experiment is more representative of prior studies or more representative of our study based on the experimenter-chosen values for p, r, and $M/(pH)$. To do so, we first create a combined data set of 263 observations consisting of our own and prior experiments. We then regress an indicator for an experiment coming from a prior study on p, r, and $M/(pH)$. Finally, we use the estimated coefficients from this regression to generate for each experiment a predicted likelihood that it comes from a prior study. Importantly, this predicted likelihood depends on only an experiment's experimenter-chosen values for p, r, and $M/(pH)$, and is independent of the experiment's observed realization for $CRE - RCRE$.

We next compare experiments based on whether they are more representative of prior studies (predicted likelihood larger than 0.50) or more representative of our study (predicted likelihood smaller than 0.50). Among the 143 prior experiments, 112 are more representative of prior studies and have a (sample-weighted) average $CRE - RCRE$ of 24.7 percent, while the other 31 have an average of 4.5 percent. Among our 120 experiments, 40 are more representative of prior studies and have an average $CRE - RCRE$ of 8.4 percent, while the other 80 have an average of -0.1 percent. In other words, when we (or prior studies) use experimental parameters that are more representative of prior studies, we find more CRE; in contrast, when we (or prior studies) use experimental parameters that are less representative of prior studies, we find much less CRE.

While our stage 2 paired-choice data are suggestive that, at best, there might be a mild underlying CRP, it is not obvious how to turn this into a rigorous test. This point highlights the key limitation of using paired-choice data to test for an underlying CRP: The underlying null hypothesis of no CRP implies that we should see a CRE for some parameters, an RCRE for other parameters, and neither for yet other parameters. Given these predictions for paired-choice data, it is difficult to create a test for an underlying CRP without making structural assumptions about both preferences and noise. In contrast, with paired valuations, our two simple tests rely on relatively weak assumptions only about the nature of the noise.

6 Discussion

Our paper has two main contributions. First, methodologically, we demonstrate the limitations of using paired choice tasks to make inferences about preferences and illustrate how paired valuations can overcome these limitations. Second, empirically, our paired valuation data yield no evidence of a systematic CRP but are also inconsistent with EU. We conclude by discussing the implications of our paper for future research.

in terms of probability weighting: The probability weighting function estimated from the AB choice data takes the familiar inverse-S shape, but our combined AB and CD choice data are inconsistent with a stable probability weighting function. See Online Appendix E.2 for details.

Methodologically, our paper argues that valuation data may be better suited than choice data to reveal features of underlying preferences in the presence of noise. We do not claim valuations are always superior, even when working with paired decisions. For example, when estimating a fully specified structural model of underlying preferences and noise, there is no inherent advantage to using valuation versus choice data (beyond the fact that observing one valuation is effectively equivalent to observing many choices). The benefits of valuation data emerge in model-free tests of features of preferences that are akin to axiom testing.

Our methodological point relates to, but is distinct from, contemporary work that highlights the usefulness of valuation data. Carrera et al. (2022) point out how noise can bias choice data when studying *single decisions* and how valuations may be immune to that bias. Specifically, they investigate whether people exhibit a preference for commitment in the context of intertemporal choice. For a binary choice of whether to take up a commitment contract, they note how mean-zero error on underlying valuations can distort the take-up rate relative to what noise-free valuations would imply. They then point out that if one instead elicits valuations directly, the same mean-zero error does not create a bias. Our contribution is distinct in that it revolves around how noise creates bias when comparing *pairs of decisions*. Proposition 1 establishes that, if there is differential noise across the two decisions, then a systematic bias emerges even in populations with homogeneous preferences. Moreover, unobserved heterogeneity in underlying preferences exacerbates the inference problem. At the same time, Proposition 2 establishes that using valuations can be a useful solution under both mean-zero valuation error as in Carrera et al. (2022) (i.e., under Assumption 2a) and mean-zero utility error (i.e., under Assumption 2b).

We further highlight that our methodological contribution applies to many other domains. Canonical tests for a large number of behavioral-economic phenomena use paired choice tasks: dynamic inconsistency and present bias (comparing choices with different time stamps), loss aversion (comparing choices between mixed and non-mixed gambles), endowment effects (comparing choices of owners and non-owners), decoy effects and compromise effects (comparing choices for two-option versus three-option choice sets), ambiguity attitudes, rank dependence, and common consequence effects (comparing choices between particular sets of acts or lotteries). All of these paradigms are potentially subject to the central critique embodied in our main theoretical result and are similarly suited to the use of paired valuations to overcome the underlying inference problem.

Our core empirical contribution presents an important challenge for the literature that seeks to understand risk attitudes. Like prior research, we find clear evidence against EU as an accurate descriptive model of behavior, thus supporting the need for alternative models. However, many prominent non-EU models are built around a systematic CRP as a motivating feature of preferences—indeed, admission of a CRP is often axiomatic. Perhaps most prominent are models of probability weighting in the tradition of prospect theory. Based on their evidence, Kahneman and Tversky (1979) take a global CRP to be a property of preferences and design their model of probability weighting to accommodate this property. Later, Prelec (1998) takes an axiomatic approach to probability weighting; his key axiom of "subproportionality" is equivalent to assuming a global CRP.[44] The models of disappointment aversion in Bell (1985) and Gul (1991) are also motivated by evidence on the CRE, as is the model of cautious expected utility in Cerreia-Vioglio et al. (2015). The latter two papers pursue axiomatic approaches that modify the standard EU axioms to permit a CRP.

Our core empirical finding of no aggregate CRP calls for a reassessment of these models. Because our m_{AB} valuations are consistent with the standard S-shaped probability weighting function found in the literature, it might be fruitful for future work to understand how to retain the appealing psychology of probability-weighting models without requiring a built-in systematic CRP. Moreover, while we find no systematic CRP on average, we do find that some individuals seem to exhibit a reliable CRP while others seem to exhibit a reliable RCRP. Hence, it may be important to develop models that permit heterogeneity in whether people exhibit a CRP versus an RCRP.

More generally, our paper calls for more data to develop a solid empirical foundation upon which to build descriptively accurate models of risk attitudes. It is critical to assess the robustness of the findings obtained here (while still being responsive to the central identification problems discussed in this manuscript), and to further explore the space of parameters for common-ratio problems. Moreover, future work should investigate a richer space of decision problems. For instance, in ongoing work (McGranaghan et al., 2023), we use a valuations approach to gather data on both common-ratio problems and common-consequence problems over a much wider range of parameters than previously studied for either problem. Interestingly, in some circumstances, a CRP emerges. More importantly, we highlight the natural connection that exists between the two types of problems, and studying connected common-ratio and common-consequence problems reveals novel information on the shape of risk preferences. As more studies of this type emerge, the literature will be better positioned to develop more descriptively accurate models of risk attitudes.

References

Agranov, M. and Ortoleva, P. (forthcoming). Ranges of randomization. *Review of Economics and Statistics.*

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'école Américaine. *Econometrica*, 21(4):503–546.

Alós-Ferrer, C., Fehr, E., and Netzer, N. (2021). Time will tell: Recovering preferences when choices are noisy. *Journal of Political Economy*, 129(6):1828–1877.

Ballinger, T. P. and Wilcox, N. T. (1997). Decisions, error and heterogeneity. *The Economic Journal*, 107(443):1090–1105.

[44]In our notation, Kahneman and Tversky (1979) and Prelec (1998) both assume that, for any $H > M > 0$ and $0 < p < q \leqslant 1$, $(H,p) \sim (M,q)$ implies $(H,rp) \succ (M,rq)$ for any $r \in (0,1)$. Note that this axiom covers a broader set of decisions than we study in our experiments since we restrict attention to the case with $q = 1$.

Barberis, N. C. (2013). Thirty years of prospect theory in economics: A review and assessment. *Journal of Economic Perspectives*, 27(1):173–196.

Bell, D. E. (1985). Disappointment in decision making under uncertainty. *Operations Research*, 33(1):1–27.

Bernheim, B. D. and Sprenger, C. D. (2020). On the empirical validity of cumulative prospect theory: Experimental evidence of rank-independent probability weighting. *Econometrica*, 88(4):1363–1409.

Bhatia, S. and Loomes, G. (2017). Noisy preferences in risky choice: A cautionary note. *Psychological Review*, 124(5):678–687.

Blavatskyy, P., Panchenko, V., and Ortmann, A. (2023). How common is the common-ratio effect? *Experimental Economics*, 26(2):253–272.

Blavatskyy, P. R. (2007). Stochastic expected utility theory. *Journal of Risk and Uncertainty*, 34(3):259–286.

Blavatskyy, P. R. (2010). Reverse common ratio effect. *Journal of Risk and Uncertainty*, 40(3):219–241.

Bordalo, P., Gennaioli, N., and Shleifer, A. (2012). Salience theory of choice under risk. *The Quarterly Journal of Economics*, 127(3):1243–1285.

Brown, A. L. and Healy, P. J. (2018). Separated decisions. *European Economic Review*, 101:20–34.

Bruhin, A., Fehr-Duda, H., and Epper, T. (2010). Risk and rationality: Uncovering heterogeneity in probability distortion. *Econometrica*, 78(4):1375–1412.

Butler, D. and Loomes, G. (2011). Imprecision as an account of violations of independence and betweenness. *Journal of Economic Behavior & Organization*, 80(3):511–522.

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

Carrera, M., Royer, H., Stehr, M., Sydnor, J., and Taubinsky, D. (2022). Who chooses commitment? evidence and welfare implications. *Review of Economic Studies*, 89(3):1205–1244.

Castillo, M. and Eil, D. (2014). Taring the multiple price list: Imperceptive preferences and the reversing of the common ratio effect. *Working Paper*.

Cerreia-Vioglio, S., Dillenberger, D., and Ortoleva, P. (2015). Cautious expected utility and the certainty effect. *Econometrica*, 83(2):693–728.

Chapman, J., Dean, M., Ortoleva, P., Snowberg, E., and Camerer, C. (2022). Econographics. *Journal of Political Economy Microeconomics*.

Coakley, C. W. and Heise, M. A. (1996). Versions of the sign test in the presence of ties. *Biometrics*, 52(4):1242–1251.

Dean, M. and Ortoleva, P. (2019). The empirical relationship between nonstandard economic behaviors. *Proceedings of the National Academy of Sciences*, 116(33):16262–16267.

Enke, B. and Graeber, T. (forthcoming). Cognitive uncertainty. *Quarterly Journal of Economics*.

Enke, B., Graeber, T., and Oprea, R. (2023). Complexity and time. *Working Paper*.

Freeman, D. J., Halevy, Y., and Kneeland, T. (2019). Eliciting risk preferences using choice lists. *Quantitative Economics*, 10(1):217–237.

Freeman, D. J. and Mayraz, G. (2019). Why choice lists increase risk taking. *Experimental Economics*, 22(1):131–154.

Frydman, C. and Jin, L. J. (2023). On the source and instability of probability weighting. *Working Paper*.

Gillen, B., Snowberg, E., and Yariv, L. (2019). Experimenting with measurement error: Techniques with applications to the caltech cohort study. *Journal of Political Economy*, 127(4):1826–1863.

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

Harless, D. W. and Camerer, C. F. (1994). The predictive utility of generalized expected utility theories. *Econometrica*, 62(6):1251–1289.

Hey, J. D. (2005). Why we should not be silent about noise. *Experimental Economics*, 8(4):325–345.

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

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

Khaw, M. W., Li, Z., and Woodford, M. (2021). Cognitive imprecision and small-stakes risk aversion. *Review of Economic Studies*, 88(4):1979–2013.

Loomes, G. (2005). Modelling the stochastic component of behaviour in experiments: Some issues for the interpretation of data. *Experimental Economics*, 8(4):301–323.

Loomes, G. and Sugden, R. (1982). Regret theory: An alternative theory of rational choice under uncertainty. *The Economic Journal*, 92(368):805–824.

Loomes, G. and Sugden, R. (1998). Testing different stochastic specifications of risky choice. *Economica*, 65(260):581–598.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In Zarembka, P., editor, *Frontiers in Econometrics*, pages 105–142. Academic Press, New York.

McFadden, D. (1981). Econometric models of probabilistic choice behavior. In Manski, C. F. and McFadden, D., editors, *Structural Analysis of Discrete Data and Econometric Applications*, chapter 5, pages 198–272. MIT Press, Cambridge, MA.

McGranaghan, C., Nielsen, K., O'Donoghue, T., Somerville, J., and Sprenger, C. D. (2023). Connecting common ratio and common consequence preferences. *Work in Progress*.

Oprea, R. (2022). Simplicity equivalents. *Working Paper*.

Prelec, D. (1998). The probability weighting function. *Econometrica*, 66(3):497–527.

Quiggin, J. (1982). A theory of anticipated utility. *Journal of Economic Behavior and Organization*, 3(4):323–343.

Randles, R. H. (2001). On neutral responses (zeros) in the sign test and ties in the Wilcoxon–Mann–Whitney test. *The American Statistician*, 55(2):96–101.

Schneider, M. and Shor, M. (2017). The common ratio effect in choice, pricing, and happiness tasks. *Journal of Behavioral Decision Making*, 30(4):976–986.

Stott, H. P. (2006). Cumulative prospect theory's functional menagerie. *Journal of Risk and Uncertainty*, 32(2):101–130.

Tversky, A. and Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323.

Wilcox, N. T. (2008). Stochastic models for binary discrete choice under risk: A critical primer and econometric comparison. In Cox, J. C. and Harrison, G. W., editors, *Risk Aversion in Experiments (Research in Experimental Economics, Vol. 12)*, pages 197–292. Emerald Group Publishing Limited.

Wu, G. and Gonzalez, R. (1996). Curvature of the probability weighting function. *Management Science*, 42(12):1676–1690.