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

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

November 21, 2022

Abstract

Without strong assumptions about how noise manifests in choices, we can infer little about whether there exist underlying common ratio preferences (CRP) given existing empirical observations of the common ratio effect (CRE). To solve this inferential challenge, we propose 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, the 2022 Behavioral Economics Annual Meeting, and the 2022 Stanford Institute for Theoretical Economics (Experiment Economics) Conference. 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 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 counterexample to expected utility (EU). The CRE was later popularized by Kahneman and Tversky (1979) under the label "subproportionality," and a large subsequent experimental literature has used paired binary-choice tasks to provide 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 often seen as a 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, 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, little can be inferred from the existence or absence of a CRE in paired choice tasks about whether there exists an underlying *common ratio preference* (CRP).

Motivated by this observation, in this paper we propose a solution to this inference challenge using *paired valuation tasks* instead of *paired choice tasks*. We first demonstrate theoretically that, under the same commonly used assumptions about the structure of noise, paired choice tasks 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 systematic CRP at the aggregate level, albeit with 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. But we further demonstrate how appropriately chosen experimental parameters can generate either a CRE or an RCRE even for individuals without an associated CRP or RCRP.

In Section 2, we develop theoretical results for testing the EU null—or, more generally, the null of no CRP—using both paired choice tasks and paired valuation tasks when both are subject to noise. Importantly, we assume that both are driven by the same underlying preference and noise structure, so the different conclusions are not driven by different assumptions. Moreover, we focus on assumptions about noise that are commonly used in the literature—specifically, we focus on 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 scaled down by a common ratio of 0.2. In contrast, the empirical finding of a CRE is that the aggregate choice frequencies systematically deviate from the EU prediction and toward lotteries A and D—framed as $\Pr(AD) > \Pr(BC)$ or, equivalently, as $\Pr(A) > \Pr(C)$. In other words, individuals appear more risk averse for the AB choice than for the CD choice. Blavatskyy et al. (2022) provide a recent meta-study, and out of 143 paired choice experiments, 78 percent find $\Pr(A) > \Pr(C)$.

While the empirical finding of a CRE in a paired choice task is often interpreted as a rejection of EU, this inference is invalid if there exists noise that has a differential impact on the two choices. For instance, consider a person whose underlying EU preferences favor A and C, but who is influenced by noise that is more impactful for the CD choice. The existence of noise will make this person choose both A and C with a probability less than one. However, this effect will be larger for the CD choice because the noise is more impactful for this choice, resulting in a prediction of $\Pr(A) > \Pr(C)$. In other words, this person would exhibit a CRE even though they have no underlying CRP. Moreover, there is a natural reason to think that there might be differential noise of this form: Because lotteries C and D are scaled-down versions of lotteries A and B, the utility difference between lotteries C and D is smaller than that between A and B, and thus noise is likely to play a larger role in the CD decision. More generally, a researcher trying to draw inference from the difference between the observed choice proportions, $\Pr(A)$ and $\Pr(C)$, cannot reliably disentangle whether that difference is due to preferences or differential noise.

Now consider instead the two valuations below, which together form a *paired valuation task*:

***AB* Valuation**:	m_{AB} such that a 100 percent chance of m_{AB}	$\sim$	50 percent chance of \$30
***CD* Valuation**:	m_{CD} such that a 20 percent chance of m_{CD}	$\sim$	10 percent chance of \$30

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

While the use of valuations is quite 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 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 conduct a simple test of whether the mean of Δm is equal to 0. Alternatively, even if elicited valuations are biased measures of underlying values—e.g., due to utility curvature—as long as the noise is symmetric around its median, then we can instead use a sign test to assess whether there are equal proportions of positive and negative instances of Δm. Importantly, both tests are robust to noise having a differential impact across the AB and CD tasks.

It is worth highlighting that our point is not that valuations are a better instrument than choices in general. When studying single decisions, choices and valuations both provide information on preferences, and in principle both could be unbiased. Our point is that, even if both are unbiased when studying single decisions, a difference emerges when comparing *pairs* of decisions: Specifically, situations in which there is differential noise across decisions creates bias when comparing pairs of choices, but not when comparing pairs of valuations.

In Section 3, we discuss 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 valuations for 15 combinations of (p, r). In stage 2 of the experiment, we present 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.[1]

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. On one dimension, 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. On a second dimension, our m_{AB} valuations yield data that are consistent with models of probability weighting and thus inconsistent with EU. Specifically, our m_{AB} elicitations are equivalent to the tasks that researchers commonly use to estimate probability weighting functions, and they yield an inverse-S-shaped probability weighting

[1]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.

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.

There are two key predictions that 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. Second, the impact of differential noise depends on the distance between the randomly chosen amount M and the stage 1 average indifference point $\bar{m} = (m_{AB} + m_{CD})/2$; we refer to $M - \bar{m}$ as the *distance to indifference*. A sufficiently large positive distance to indifference means M is large enough that preferences favor 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.[2]

When we take these predictions to the data, we find strong support for them at both the individual level and the 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 stage 1 valuations are capturing underlying preferences. Our finding that the distance to indifference has similarly strong predictive power for 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. Prior experiments cannot assess this possibility since they do not have a measure of distance to indifference. However, more than 75 percent of the prior paired choice experiments reviewed by Blavatskyy et al. (2022) have $\Pr(A) > 1/2$, which is indicative of positive distances to indifference. Under the differential noise issue that we document, these are precisely the studies that are likely to yield a CRE even if there were no underlying CRP.

[2]Of course, if the choice noise were more impactful for the AB choice, we would expect to see the reverse pattern. However, in Section 5 we show that our data support noise being more impactful for the CD choice, hence why we often focus on that case in the text.

[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.

Our analysis is related to several strands of prior research. First, previous work has highlighted how EU with additive utility noise can generate a CRE in paired choice tasks (Ballinger and Wilcox, 1997; Loomes, 2005; Hey, 2005; Wilcox, 2008; Blavatskyy, 2007, 2010; Bhatia and Loomes, 2017). We complement these efforts by characterizing the complete set of observed choice probabilities in paired choice tasks that are consistent with EU when there is choice noise and preference heterogeneity. This endeavor also relates to the literature that proposes different approaches to accounting for noise when interpreting experimental data more broadly (Harless and Camerer, 1994; Hey and Orme, 1994; Ballinger and Wilcox, 1997; Loomes and Sugden, 1998; Stott, 2006). Finally, we note that a small number of papers 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. Nonetheless, there are some interesting connections to our findings, which we discuss in our concluding Section 6.

The key idea of this paper applies far beyond the domain of the common ratio effect. The literature is filled with attempts to test "effects" or "axioms" using paired choice tasks, implicitly assuming away any differential noise. As such, they are subject to the same critique as we raise here, and hypothesis testing using paired valuation tasks may be a constructive solution.[4]

2 Underlying Theory and Proposed Tests

In this section, we develop theoretical results for testing the null of expected utility (EU)—or, more generally, the null of no CRP—using paired choice tasks versus paired valuation tasks when both are subject to noise.

2.1 Paired Choices and Paired Valuations

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

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

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 multiplying 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 (in their Section 2.3) 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 outcome 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 points also hold for a non-zero

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

$$\begin{aligned} EU(A) - EU(B) &> 0 \quad \Leftrightarrow \quad u(M) - pu(H) > 0 \quad \text{and} \\ EU(C) - EU(D) &> 0 \quad \Leftrightarrow \quad r\left[u(M) - pu(H)\right] > 0. \end{aligned}$$

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 that are systematically in the direction of choosing lotteries A and D. 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] 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:

$$\begin{aligned} &\textbf{\textit{AB}} \textbf{ Valuation Task}: \quad \text{elicit an } m_{AB} \in [0, H] \text{ such that} \quad (m_{AB}, 1) \sim (H, p) \\ &\textbf{\textit{CD}} \textbf{ Valuation Task}: \quad \text{elicit an } m_{CD} \in [0, H] \text{ such that} \quad (m_{CD}, r) \sim (H, pr). \end{aligned}$$

For paired valuation tasks, a finding of $\Delta m \equiv m_{CD} - m_{AB} > 0$ reflects a CRE because it implies that a paired choice task that offers any M larger than m_{AB} but smaller than m_{CD} would yield a CRE. More intuitively, a CRE involves people acting more risk seeking (or less risk averse) when probabilities are scaled down by a common ratio. 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).

When one interprets data from either choices or valuations, it is important to account for noise. Here, we develop a single framework that we use 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 combination of their preferences and noise. The person then makes the choice or states the valuation implied by this realized indifference value.

Without loss of generality, we fix (H, p, r) and focus on behavior as a function of M.[7] Assuming

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 from which X is chosen because it is typically self-explanatory; for example, $\widehat{\Pr}(A)$ is the proportion who choose A from choice set $\{A, B\}$, and $\widehat{\Pr}(AD)$ is the proportion who choose combination AD from the choice set $\{AC, AD, BC, BD\}$.

[7] The framework in the text links directly to the m-tasks in our experiment. We provide an analogous framework in Appendix B.4 where we fix (M, p, r) and focus on behavior as a function of H; that framework links directly to the h-tasks in our experiment.

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
- Prefer $C \equiv (M, r)$ over $D \equiv (H, pr)$ if and only if $M \geqslant m^*_{CD}$.[8]

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 $D \equiv (H, pr)$ 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.

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.

[8] To simplify the exposition, the text assumes that the person prefers and chooses the safer option when indifferent, but this assumption is immaterial.

Assumption 2a is consistent with assumptions that researchers frequently use when analyzing valuations data, where noise is modeled as a disturbance added to an underlying value (in this case the indifference point). Assumption 2b is consistent with assumptions that researchers frequently use when analyzing choice data, where noise is instead modeled 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 $\pi(q)u(x)$.[9] 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 von Neumann–Morgenstern utility function. Under this formulation, the underlying indifference points satisfy

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

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.[10]

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

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

where ϵ_{AB} and ϵ_{CD} reflect additive utility noise.[11] 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 ϵ_{AB} and ϵ_{CD} are both mean-zero normal or logistic distributions.[12] Finally, note that for

[9] Our data include only binary gambles of this type. This equation uses the normalization $u(0) = 0$.

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

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

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

the case of EU with additive utility noise that is i.i.d. across the AB and CD choice tasks, $\varepsilon_{CD} = \varepsilon_{AB}/r$.

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 $\Delta m^* = 0$.[13]

Proposition 1 *(Paired Choice Tasks Yield Biased Tests of $\Delta m^* = 0$):* Consider a person who has $m^*_{AB} = m^*_{CD} \equiv m^*$ and thus $\Delta m^* = 0$. Suppose further that $\varepsilon_{CD} \stackrel{d}{=} k\varepsilon_{AB}$ 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.

To see the intuition behind Proposition 1, consider the case where $M - m^* > 0$ and the median of ε_{AB} is zero and thus $Z = 1/2$. The person's underlying preference is for A and C, but choice noise pulls both $\Pr(A)$ and $\Pr(C)$ from 1 toward 1/2. The magnitude of this effect is increasing in the variance of the choice noise. If $k > 1$, then ε_{CD} has a larger variance than ε_{AB}; thus the choice noise will have a larger impact on the CD choice, and so $\Pr(A) > \Pr(C) > 1/2$. The other cases in Proposition 1 are analogous. The implication is that a person with $\Delta m^* = 0$ could exhibit either a CRE or an RCRE depending on the combination of whether (i) 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 $k = 1$. While this could be the case in principle, there is no reason to presume that it holds. Indeed, for the case of EU with additive utility noise, $k = 1/r > 1$ (see Example 1). Parts (1)(a) and (2)(a) of Proposition 1 therefore imply that EU with additive utility

[13] All proofs appear in Appendix A.

noise will lead to an observed CRE when $M - m^* > 0$ and an observed RCRE when $M - m^* < 0$. This observation is not novel; Ballinger and Wilcox (1997), Loomes (2005), Hey (2005), Wilcox (2008), Blavatskyy (2007, 2010), and Bhatia and Loomes (2017) have all pointed to variations on this theme. Proposition 1 expands on their results by establishing that the result holds for any model that predicts $\Delta m^* = 0$ (i.e., not just EU) and by permitting the variance of the choice noise to be larger for either the AB or the CD choice.

Whereas paired choice tasks yield a biased test, Proposition 2 provides conditions for when unbiased tests are possible using paired valuation tasks.

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 $(\varepsilon', \varepsilon')$,[14] 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 $\widehat{E}(\Delta m) = 0$.[15] The intuition for part (1) is straightforward. 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 means different from zero, 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.

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 Appendix B.1 we show that, for the case of EU with additive i.i.d. utility noise, concave utility implies $E(\Delta m) > 0$, and 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, specifically, testing whether the observed proportions of $\Delta m > 0$ and $\Delta m < 0$ are the same. The intuition behind part (2) is easiest to see when noise is symmetric around zero (i.e., $\varepsilon' = 0$). For this case, 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

[14] 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.

[15] 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.

$\varepsilon_{CD} > -\bar{\varepsilon}$. Averaging over all $\bar{\varepsilon}$, it is therefore equally likely to get $\Delta m > 0$ and $\Delta m < 0$.[16]

Proposition 1 outlines scenarios under which paired choice tasks yield a biased test of the null of $\Delta m^* = 0$, whereas 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 yield a biased test of $\Delta m^* = 0$, whereas paired valuations yield an unbiased test based on the mean of Δm. Under Assumption 2b, paired choice tasks yield a biased test of $\Delta m^* = 0$, whereas paired valuations yield an unbiased test based on the sign of Δm.

To illustrate the implications of Corollary 1, Figure 1 depicts the set of theoretical predictions consistent with everyone having $\Delta m^* = 0$ under Assumption 2a, and also connects these theoretical predictions to data from existing experimental tests of the CRE. The theoretical predictions, which we formally derive in Appendix B.2, permit heterogeneity in preferences (i.e., where some prefer A and C, while others prefer B and D) and in the impact of noise.

Panel A focuses on data from paired choice tasks, where observed behavior in a population is $\widehat{\Pr}(A)$ and $\widehat{\Pr}(C)$. If the impact of noise were the same for both the AB and the CD choices (i.e., $k = 1$ in Proposition 1), 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.[17] Panel B focuses on data from paired valuation tasks, where observed behavior in a population is $\widehat{E}(m_{AB})$ and $\widehat{E}(m_{CD})$. Under Assumption 2a, the set of predicted $(E(m_{AB}), E(m_{CD}))$ 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 identified in the meta-study by Blavatskyy et al. (2022). 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) have this outcome—hence, the widespread perception that there exists a systematic CRE. At the same time, panel A also reveals that almost

[16] 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 $2 * G(\min\{n_+, n_-\}, n_+ + n_-)$.

[17] The logic behind Proposition 1 implies that the squares in the upper right and lower left are possible combinations even with a homogeneous population; heterogeneous preferences are required for the remaining space. 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 Appendix B.2 characterizes the full set.

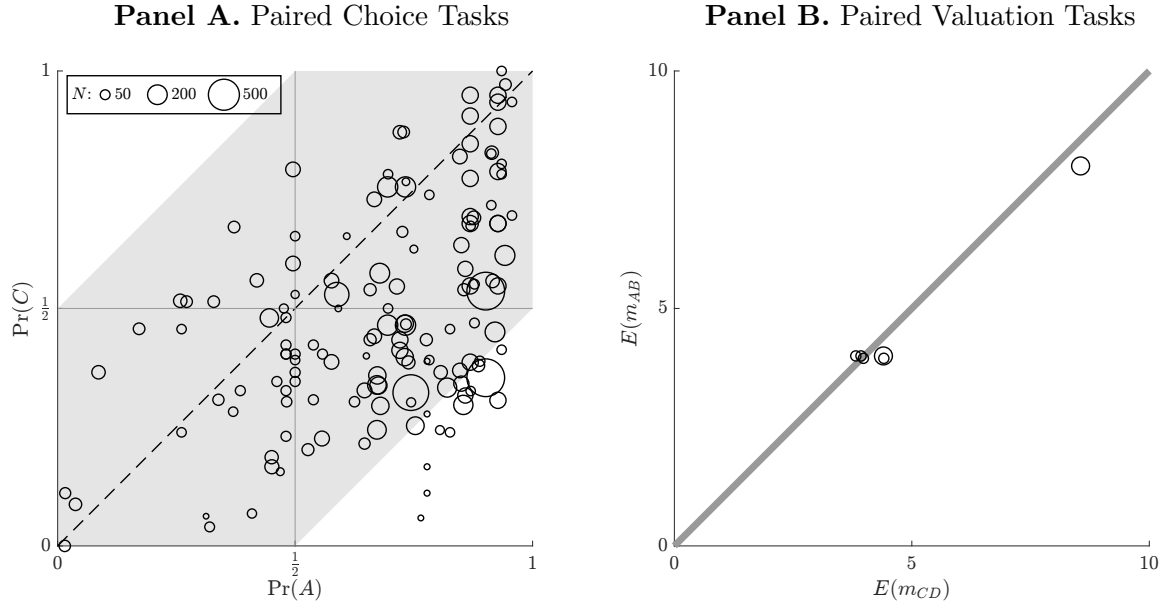

Figure 1: Predictions and observations for paired choice tasks (panel A) and paired valuation tasks (panel B). In each panel, points below the 45-degree line are combinations that indicate a CRE, while points above the 45-degree line are combinations that indicate an RCRE. The shaded grey region in panel 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 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. (2022) scaled by the number of observations; panel B depicts six similarly scaled CRE paired-valuation experiments identified by us (see Appendix B.3 for details).

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.

Panel A is consistent with the approach in much of the CRE literature that compares $\widehat{\Pr}(A)$ versus $\widehat{\Pr}(C)$ (or, equivalently, that compares $\widehat{\Pr}(AD)$ versus $\widehat{\Pr}(BC)$). However, one could instead study whether there is an excess share of AD (or BC) combinations. For instance, an additional implication of Proposition 1 is that, when the noise has a median of zero and thus $Z = 1/2$, we could have $\Pr(AD)$ as high as 50 percent, but no larger.[18] Among the experiments depicted in panel A, this prediction is rarely violated: Of the 143 experiments, only 20 have 50 percent or more AD combinations, including all 14 that fall outside of the shaded region in panel A. An analogous prediction holds for $\Pr(BC)$; and no experiments have 50 percent or more BC combinations.

The black circles in panel B of Figure 1 depict observed $(\widehat{E}(m_{AB}), \widehat{E}(m_{CD}))$ combinations from six CRE paired-valuation experiments that we identified (see Appendix B.3 for details about the

[18] Part 1 of Proposition 1 implies that anyone with $M - m^* > 0$ must have $\Pr(A)$ and $\Pr(C)$ between 1/2 and 1. The latter implies that $\Pr(D)$ could be as large as 1/2 but no larger, and thus $\Pr(AD)$ cannot be larger than 1/2. Similarly, part (2) implies that anyone with $M - m^* < 0$ must have $\Pr(A)$ and $\Pr(C)$ between 0 and 1/2. The former implies that $\Pr(A)$ could be as large as 1/2 but no larger, and thus again $\Pr(AD)$ cannot be larger than 1/2.

data). In terms of economic significance, panel B demonstrates that the data from paired valuation tasks do not fall far from the 45-degree line (the figure does not show statistical significance). Moreover, comparing panel B to panel A reveals how the literature has tested for a CRE using almost exclusively paired choice tasks.

The theoretical predictions in Figure 1 are the limits imposed by Assumption 2a and $\Delta m^* = 0$ on arbitrarily large data sets. For paired valuation tasks, the tight prediction stems from the fact that noise is merely a disturbance that can matter for individual valuations but which must average out in sufficiently large samples. For paired choice tasks, in contrast, even with infinite data, the empirical choice proportions $(\widehat{\Pr}(A), \widehat{\Pr}(C))$ are not restricted in the same way. In particular, the difference between $\widehat{\Pr}(A)$ and $\widehat{\Pr}(C)$ is a function of the relative noise terms, and larger samples merely yield more precise estimates of the impact of differential noise.

This section has highlighted one key implication of our framework: Paired choice tasks yield biased tests of $\Delta m^* = 0$, whereas paired valuation tasks can yield unbiased tests. However, the proposed framework also implies a strong connection between paired valuation tasks and paired choice tasks because each is driven by the same underlying preferences, m^*_{AB} and m^*_{CD}. Motivated by this connection, our experiment collects data from the same participants on both paired valuation tasks and linked paired choice tasks. In Section 5, we derive and test predictions for when we should observe a CRE versus an RCRE in paired choice tasks as a function of one's valuations and the experimental parameters chosen by the researcher.[19]

2.2 Robustness: h-Valuation Tasks

Our main analysis focuses on the m-valuation tasks described in Section 2.1 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 $m \in [0, H]$, and this domain is the same for all p and r. Second, the AB variants of the m-valuation task 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 is typically observed in the literature.

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

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

Appendix B.4 provides a full development of the theory underlying h-valuation tasks that is analogous to the theory developed in Section 2.1 for m-valuation tasks. Here, we highlight

[19] Some researchers have raised the question whether valuations or choices provide better insight into people's underlying preferences (see in particular Freeman and Mayraz (2019) and Freeman et al. (2019)). In Section 5, we use the connection between the same participants' valuations and choices to shed insight on this issue.

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.[20] 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.[21] 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 entire experiment. In stage 1, participants complete 10 paired valuation tasks for a total of 20 valuations. In stage 2, participants complete 10 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 2 provides a high-level overview of the experiment timeline.

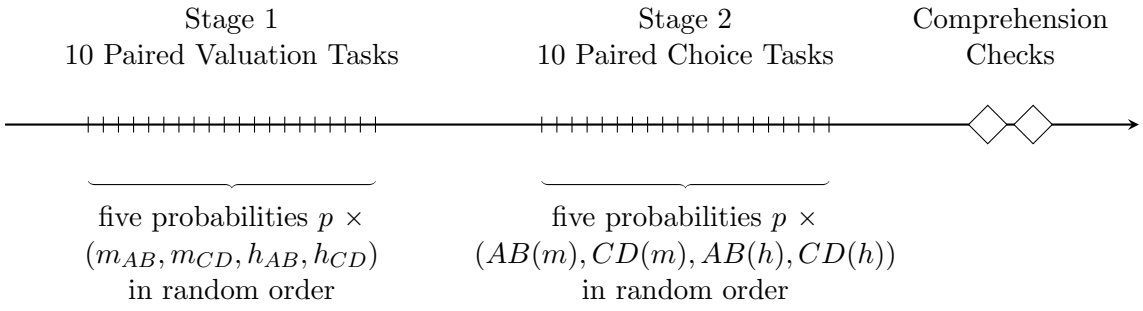

Figure 2: Experiment Timeline

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 high 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

[20] 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.

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

is a 100 percent chance of M that starts with $M = \$0$ and increases by \$1 per row. The paired CD variant has the same structure, except both probabilities are reduced by the participant's common ratio r. For each variant, the average value of M at the switching rows is taken as a 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 value of the middle 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, the average value of H at the switching rows is taken as a 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. 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 $(0.1, 0.2, 0.5, 0.8, 0.9)$ $\times$ two variants (AB and CD) $\times$ 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.

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 high 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 middle 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.[22]

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

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

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

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.

check tests whether participants can correctly answer a binary-choice question when given another person's responses to a price list. Appendix Figures C.11 and C.12 provide example screenshots of these comprehension checks.[23]

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. 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 Appendix Table D.1 for summary statistics about 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 men and women participants.[24] 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

[23]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.

[24]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 as a result of a social media trend. To better approximate the typical college population, we chose to recruit 450 men and 450 women.

the 42 questions (20 valuations, 20 binary choices, and two incentivized comprehension checks) was equally likely to determine the amount of the bonus payment. 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 on average 27 minutes to complete, and participants earned an average total payment of \$6.51.[25]

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 3 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 1; specifically, 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 implies $m^*_{CD} > m^*_{AB}$, and thus is reflected by observations below the 45-degree line. For paired h-valuation tasks, a CRE implies $h^*_{AB} > h^*_{CD}$, and thus, given the change in axes, is again reflected by observations below the 45-degree line. Figure 3 reveals very little evidence of a systematic CRE.

More formally, we focus on m-valuations and conduct the two tests developed in Section 2.1 based on Proposition 2.[26] 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. 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 Appendix Table D.2 for complete summary statistics on the m-valuations.)

As discussed in Section 2.1, the means test may be biased if the function Γ in Assumption 1 is nonlinear; hence, we also consider the test based on the sign of Δm. Columns (4)–(6) report the raw frequency data: For each combination of r and p, the table reports the number of participants who have $\Delta m > 0$ (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 2.1 (see footnote 16). Out of the 15 sign tests, we find seven significant deviations from the null of equal proportions

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

[26] We consider formal tests for the h-valuations in Section 4.2.

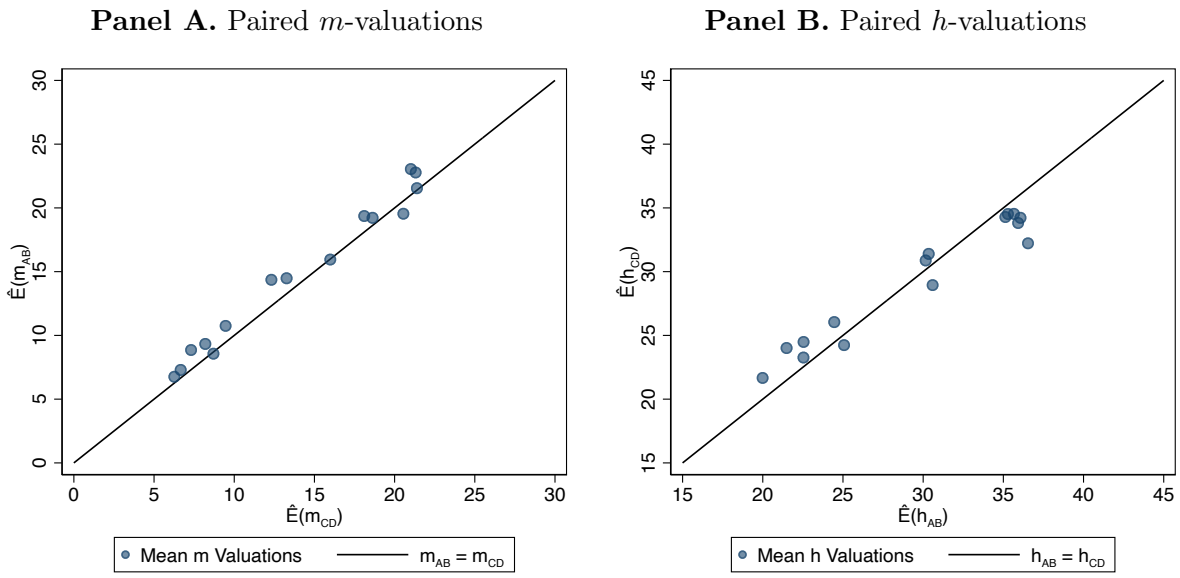

Figure 3: Mean valuations for each of the 15 (p, r) combinations for the paired m-valuation task (panel A) and the paired h-valuation task (panel B). In each panel, points below the 45-degree line are combinations that indicate a CRE, while points above the 45-degree line are combinations that indicate an RCRE.

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 predicted by 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 due to the fact that 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 $\Delta m > 0$ 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 Appendix Table D.3, we present the sign-test results using both of these approaches and show that the overall message is almost identical.

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

A possible issue is whether the absence of a CRE in our valuation data is due to theory predicting

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. $r = 0.2$				
$p = 0.1$	-1.55	0.000	79	75	144†	0.000	0
$p = 0.2$	-1.29	0.003	80	73	145†	0.000	0
$p = 0.5$	0.04	0.932	123	60	115	0.650	0
$p = 0.8$	1.00	0.052	140†	54	104	0.025	0
$p = 0.9$	-1.47	0.014	127	42	129	0.950	0
			Panel B. $r = 0.4$				
$p = 0.1$	-0.63	0.152	103	71	129	0.101	0
$p = 0.2$	-1.14	0.003	97	65	141†	0.005	0
$p = 0.5$	-1.22	0.007	104	62	137†	0.039	0
$p = 0.8$	-0.60	0.262	127	41	135	0.665	0
$p = 0.9$	-0.16	0.782	124	52	127	0.900	0
			Panel C. $r = 0.6$				
$p = 0.1$	-0.49	0.158	94	90	115	0.166	0
$p = 0.2$	0.14	0.692	111	84	104	0.682	0
$p = 0.5$	-2.05	0.000	89	65	145†	0.000	0
$p = 0.8$	-1.26	0.008	113	57	129	0.335	0
$p = 0.9$	-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.

small magnitudes of Δm for our experimental lotteries. To assess this issue, we interpret our data within the prospect theory (PT) structure from Example 1. Figure 4 presents the mean m-valuations in a different way from Figure 3(a): For each of the five values of p, the blue dots denote the mean m_{AB} valuations and the red diamonds denote the mean m_{CD} valuations, where 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 the valuation tasks frequently used 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. Appendix E provides the details of this structural estimation and reports the parameter estimates; the AB valuations predicted by these estimates are depicted in Figure 4 by the dashed blue lines. In each case, the estimated probability weighting function takes the familiar inverse-S shape that is typically found in the literature.[27]

[27]We use the functional forms from (Tversky and Kahneman, 1992) and corresponding parameter labels. The specific estimates for utility curvature (α) and probability weighting (γ) are: $\alpha = 1.351$ and $\gamma = 0.580$ for $r = 0.2$;

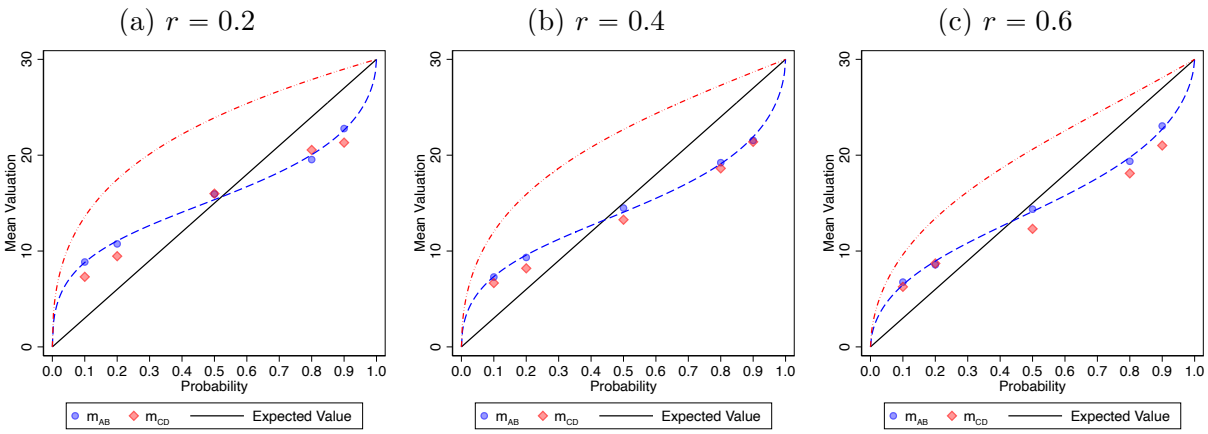

Figure 4: Mean valuations by probability p and common ratio r. The blue dots denote mean AB valuations, and the red diamonds denote mean CD valuations. The dashed blue lines represent the AB valuations predicted by PT with parameters estimated from the AB-valuation data. The dashed-and-dotted red lines denote the CD valuations predicted by PT given the parameter estimates represented by the blue lines.

We next use the probability weighting functions estimated from the AB valuations to make a prediction for the CD valuations. These predictions are depicted by the red dashed-and-dotted lines in Figure 4. Figure 4 reveals that, under PT, we should see substantial positive values of Δm given the observed AB valuations. Across the 15 combinations of p and r, the predicted Δm ranges from 3.15 to 8.63, with an average of 6.30.[28] Hence, we observe small Δm values in the data *despite* the fact that PT predicts much larger differences.

A final implication of Figure 4 is that our data cannot be explained by applying a single probability weighting function to both types of comparisons.[29] Hence, we have a combination of two findings: (i) We see the typical inverse-S shaped probability weighting function for AB valuations, and (ii) we see CD valuations that are roughly the same as the AB valuations. We emphasize that, while (ii) is consistent with EU, (i) is inconsistent with EU. Hence, the pattern of valuations we observe is inconsistent with both EU and PT. We provide a further discussion of this point in Section 6.

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

$\alpha = 1.179$ and $\gamma = 0.587$ for $r = 0.4$; and $\alpha = 1.112$ and $\gamma = 0.636$ for $r = 0.6$.

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

[29] For a more formal test, see Appendix E, 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.

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 large literature that estimates probability weighting functions, and our data yield probability weighting functions that look very similar to those estimated in the literature. 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.

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 despite the fact that we present 20 tasks in random order that differ along multiple dimensions: whether it is an m- or h-valuation, whether it is 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 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 a little 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 of these are consistent with an RCRP, and four are consistent with a CRP.[30] Finally, the median Δh is exactly zero in 12 of the 15 cases. Hence, Table 4 reaffirms Result 1: There is no evidence of a systematic CRP in valuations.

We also use the h-valuation data to validate the m-valuation data. As described in Section 2.2, the two valuations approximately measure the same preference, and thus the proportional risk premium m_x^*/H for an m-valuation should be strongly correlated with the proportional risk premium M/h_x^* for the corresponding h-valuation. In 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 confusion or heuristic responses.

4.3 Heterogeneous Preferences and Noise in Stage 1 Data

Our aggregate tests provide no indication of a systematic CRP. But 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

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

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. $r = 0.2$				
$p = 0.1$	-1.67	0.006	100	60	138†	0.016	0
$p = 0.2$	-1.94	0.001	94	53	151†	0.000	-1
$p = 0.5$	1.64	0.001	136†	81	81	0.000	0
$p = 0.8$	4.31	0.000	174†	45	79	0.000	3
$p = 0.9$	2.11	0.000	143†	64	91	0.001	0
			Panel B. $r = 0.4$				
$p = 0.1$	-2.53	0.000	82	59	162†	0.000	-1
$p = 0.2$	-1.59	0.002	92	65	146†	0.001	0
$p = 0.5$	-1.05	0.036	101	70	132†	0.049	0
$p = 0.8$	1.84	0.002	148†	47	108	0.015	0
$p = 0.9$	1.13	0.055	138	47	118	0.235	0
			Panel C. $r = 0.6$				
$p = 0.1$	-0.73	0.192	100	71	128	0.074	0
$p = 0.2$	0.83	0.108	131	65	103	0.077	0
$p = 0.5$	-0.72	0.130	93	85	121	0.065	0
$p = 0.8$	0.84	0.146	136	47	116	0.231	0
$p = 0.9$	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.

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 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}.[31] 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 5 presents the distribution of the observed $pH - \bar{m}$ across all 15 combinations

[31] We use "risk aversion" here to mean an observed aversion to risk, without reference to any model. See O'Donoghue and Somerville (2018) for a discussion of how many different models can explain observed risk aversion.

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 clearly indicate the existence of heterogeneity in underlying preferences. Second, panel A of 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.[32]

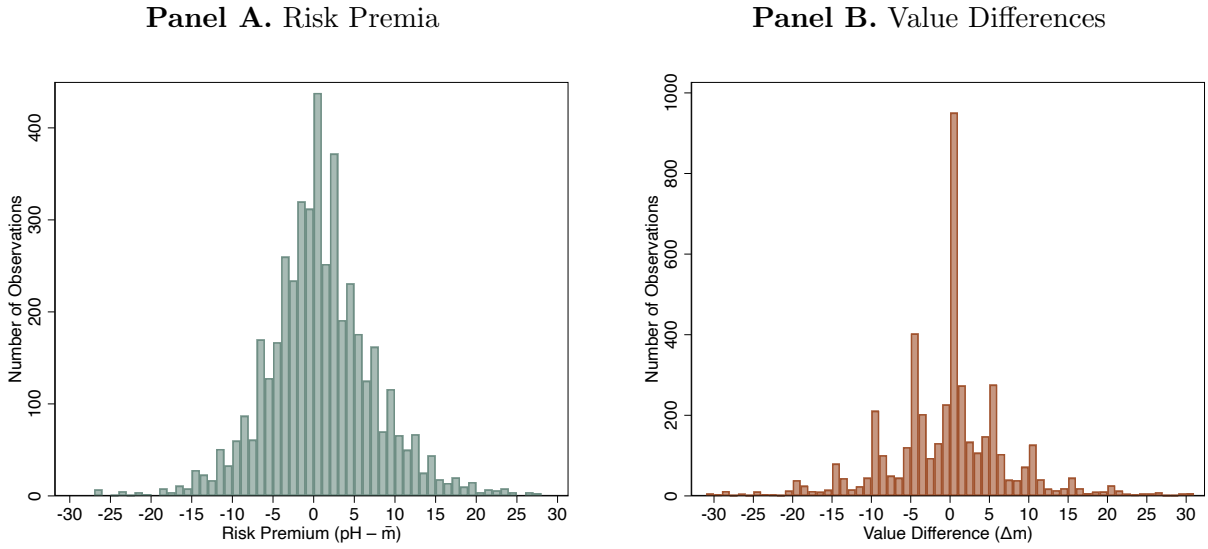

Figure 5: Distributions of risk premia (panel A) and value differences (panel B) in m-valuations. The risk premium is $pH - \bar{m}$ where $\bar{m} = (m_{AB} + m_{CD})/2$ and $H = 30$. Positive risk premia indicate risk aversion, and negative risk premia indicate risk seeking. The value difference is $\Delta m = m_{CD} - m_{AB}$. 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 5 presents the distribution of the observed Δm across all 15 combinations of (p, r). There is again substantial variation, and while some of it is surely 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 $M/h_{CD} - M/h_{AB}$. 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 Appendix Table D.8 documents substantial rank correlations of Δm across different values of p: Across the

[32]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.

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.

Hence, while we fail to find a systematic CRP in the aggregate, the variation in participants' stage 1 valuations reflects a combination of both idiosyncratic choice 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

In this section, we analyze 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

In the theoretical framework from Section 2.1, Assumption 1 states that 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 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

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

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:

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

This formulation links the extent of $CRE - RCRE$ to two terms: (i) a *scaled value difference* term $0.5(1 + 1/k)\Delta m^*$, and (ii) a *scaled distance to indifference* term $(1 - 1/k)(M - \bar{m}^*)$. The scaling factors depend on k, which captures the extent of differential noise across AB versus CD choices. When $k = 1$, there is no differential noise issue, and the scaled distance to indifference term disappears—that is, the person will exhibit a CRE if and only if they have an underlying

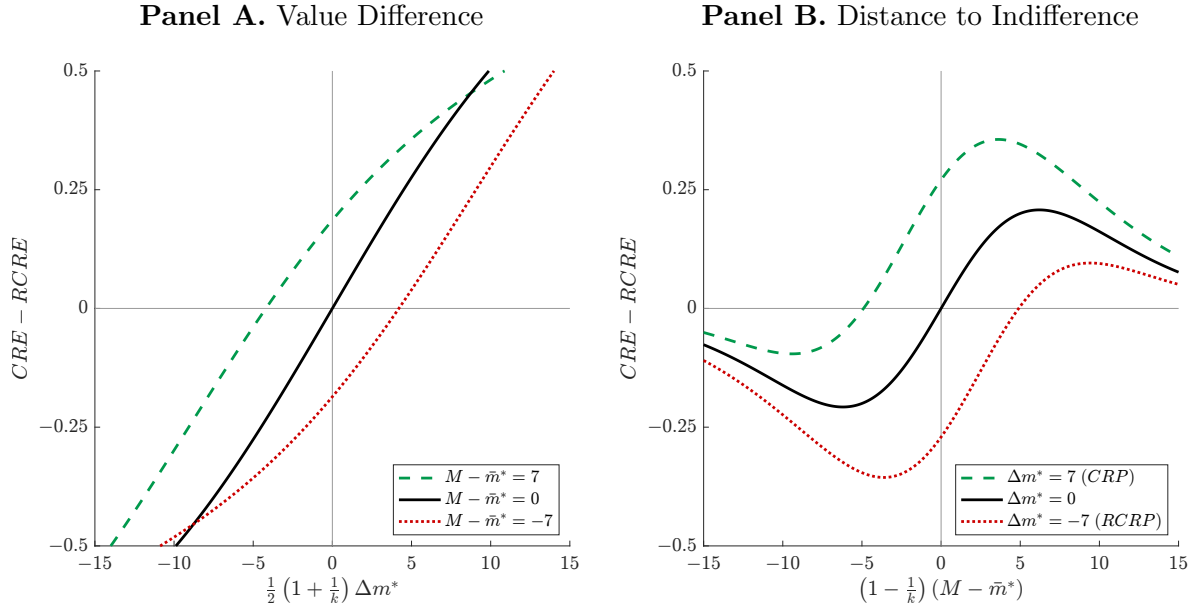

Figure 6: Predicted $CRE-RCRE$ as a function of the value difference $0.5(1+1/k)\Delta m^*$ (panel A) or distance to indifference $(1-1/k)(M-\bar{m}^*)$ (panel B). Panel A depicts predicted $CRE-RCRE$ as a function of value difference for three cases: (i) a positive distance to indifference (green dashed), (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 $k = 2.5$.

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 $k > 1$. In other words, we find evidence that noise has a larger impact on the CD choices than on the AB choices. This pattern matches the qualitative pattern we would expect under an assumption of EU with i.i.d. additive utility noise, where $k = 1/r$ (see Example 1).

Using equation (1), Figure 6 presents the predicted $CRE-RCRE$ as a function of the scaled value difference (panel A) and the scaled distance to indifference (panel B) when $k = 2.5$ and ε_{AB} is distributed $N(0, 7^2)$. Panel A considers variation in the value difference holding the distance to indifference fixed at three different levels. It highlights that the relationship between predicted $CRE-RCRE$ and the value difference is straightforward: The larger the value difference, the larger is $CRE-RCRE$. This relationship simply reflects the direct impact of preferences: The stronger the underlying CRP (or RCRP) that an individual has, the more likely they are to show a CRE (or an RCRE).

Panel B considers variation in the distance to indifference when the noise is more impactful for the CD choice ($k > 1$), holding the value difference fixed at three different levels. Panel B

illustrates how a person with $\Delta m^* = 0$ can exhibit either a CRE or an RCRE, where the direction and magnitude depend on how close M is to the person's indifference point $\bar{m}^*$.[33] Panel B further illustrates that the same qualitative pattern holds even for people with $\Delta m^* \neq 0$—indeed, an individual with an underlying RCRP may exhibit a CRE if the offered M implies a large and positive distance to indifference, while an individual with an underlying CRP may exhibit an RCRE if the offered M implies a large and negative distance to indifference.[34]

In the following subsections, we test these predictions, but we emphasize three points as we make the transition from theoretical predictions to analyzing data. First, the predictions above are framed in terms of a person's underlying Δm^* and $\bar{m}^*$, both of which are unobserved. In our empirical analysis, we replace these with their empirical counterparts Δm and $\bar{m}$, recognizing that there is measurement error and attempting to correct for it when we can.

Second, our empirical analysis combines data for the three different values of r to increase the power of our statistical tests. However, r ought to 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 6, where a k closer to one would yield that Figure 6(A) is steeper while Figure 6(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.[35]

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 must be the same as those for choices. Our analysis considers stage 1 valuations as noisy measures of underlying preferences, which motivates an instrumental variables approach; but this approach does not require any consistency between the noise in valuations and the errors in choices.

5.2 Individual-Level Stage 2 Data

We first analyze stage 2 data at the individual level. Much 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.

[33] Appendix B.5 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.

[34] At extreme levels of distance to indifference, $CRE - RCRE$ goes to zero as the preference component becomes so strong that it dominates the role of choice noise.

[35] This approach is not perfect, as we have already documented that our data are inconsistent with EU with additive utility noise. However, using that case as motivation, as opposed to choosing a correction in a more ad hoc way, imposes some discipline on our analysis. See Appendix B.4 for more discussion of this correction, including the analogue for the h-tasks.

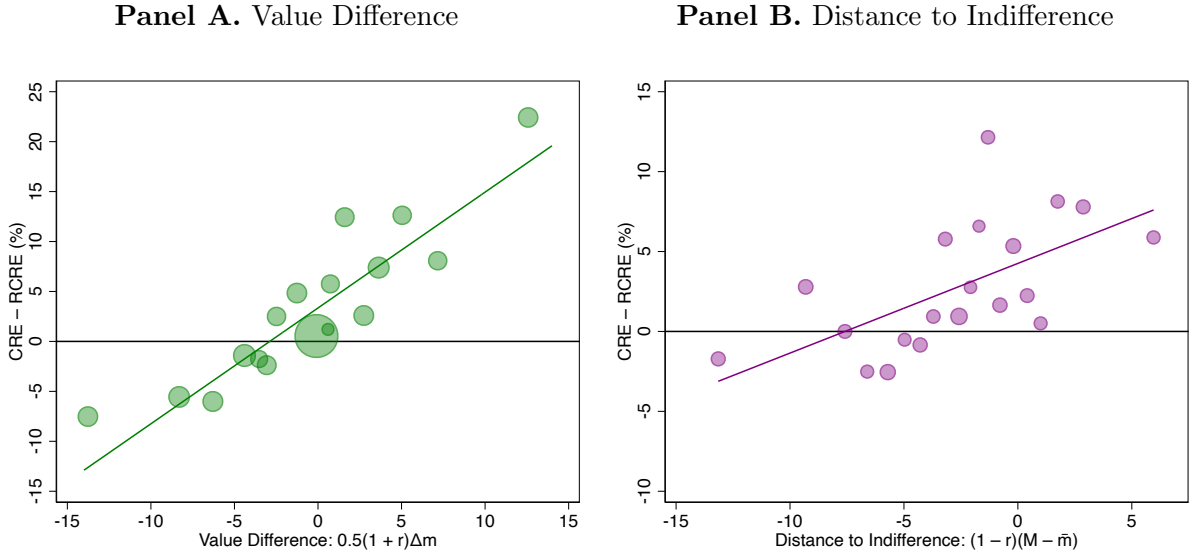

Figure 7: Stage 2 $CRE - RCRE$ as a function of individual-level stage 1 value difference (panel A) and distance to indifference (panel B). For each, the stage 1 individual-level data are collected into 20 equally sized bins, and the dots in each graph denote the average $CRE - RCRE$ for a given bin (four value difference bins are joined together at zero). 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).

Consider first the aggregate shares of the four possible choice patterns across all individual observations. We see 1,984 (44.1 percent) instances of combination BD, 1,353 (30.1 percent) instances of combination AC, 641 (14.2 percent) instances of combination AD (CRE), and 522 (11.6 percent) instances of combination BC (RCRE). Across all 4,500 observations, we observe that $CRE - RCRE = 2.6$ percent. This small overall difference is an initial indication that when the value of M encompasses both positive and negative distances to indifference, there can be little evidence of a systematic CRE.

We next consider a graphical analysis of stage 2 behavior analogous to the predicted behavior in Figure 6. Figure 7 plots the relationship between observed $CRE - RCRE$ and both 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 7 shows that individuals' stage 1 value differences are tightly linked to stage 2 behavior: 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, as well as 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. Across the three columns, the coefficient estimates are quite stable. They imply that a \$10 increase in the scaled value difference—which is roughly 1/3 the variation we see in panel A of Figure 7—is associated with an 11 percentage-point increase in $CRE - RCRE$. At the same time, a \$10 increase in the scaled distance to indifference—which is roughly 1/2 the variation we see in panel B of Figure 7—is associated with an 8 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 clearly show that distance to indifference, through its interaction with choice noise, plays an important role in whether people exhibit a CRE versus an RCRE in their stage 2 behavior.[36]

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

	(1)	(2)	(3)	(4)
	Outcome: $CRE - RCRE \in \{-1, 0, 1\}$			
	OLS	OLS	OLS	2SLS
Value Difference				
$\frac{1+r}{2}\Delta m$	1.12		1.14	6.98
	(0.14)		(0.14)	(1.07)
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 $(1-r)p\bar{h}$, $0.5p(1+r)\Delta h$, 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 due to measurement error. To account for this, column (4) of Table 5 pursues an instrumental-variables approach using the h-valuation tasks as

[36]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.

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 of 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.[37]

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 a set of 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 the average sample size across experiments of 75. The data for all 120 experiments are presented in Appendix Tables D.11 and D.12.

This experiment-level analysis is more in line with the approach in prior experimental CRE studies, and indeed our observation numbers are comparable to those in existing CRE experiments with paired choice tasks. If we were to use a two-sided test at the 5 percent level, as is typically done in the prior literature, we would conclude that among our 120 experiments, there are 20 with a significant CRE and 11 with a significant RCRE. Our goal, however, is to demonstrate that we can predict, based on our stage 1 data, which experiments are likely to yield a CRE or an RCRE.

Specifically, we calculate for each experiment (i) the average scaled value difference ($0.5(1+r)\Delta m$ for m-choice tasks or $0.5p(1+r)\Delta h$ for h-choice tasks) and (ii) the average scaled distance to indifference ($(1-r)(M-\bar{m})$ or $p(1-r)(\bar{h}-H)$).[38] We then investigate whether these averages predict $CRE - RCRE$ at the experiment level.

Figure 8 provides graphical evidence and is the experiment-level analog to the individual-level analysis in Figure 7. 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. Much as for the individual-level data, we see that both variables predict whether we observe a CRE or

[37] 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. Appendix Table D.10 is the h-task analog to Table 5 and yields similar conclusions.

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

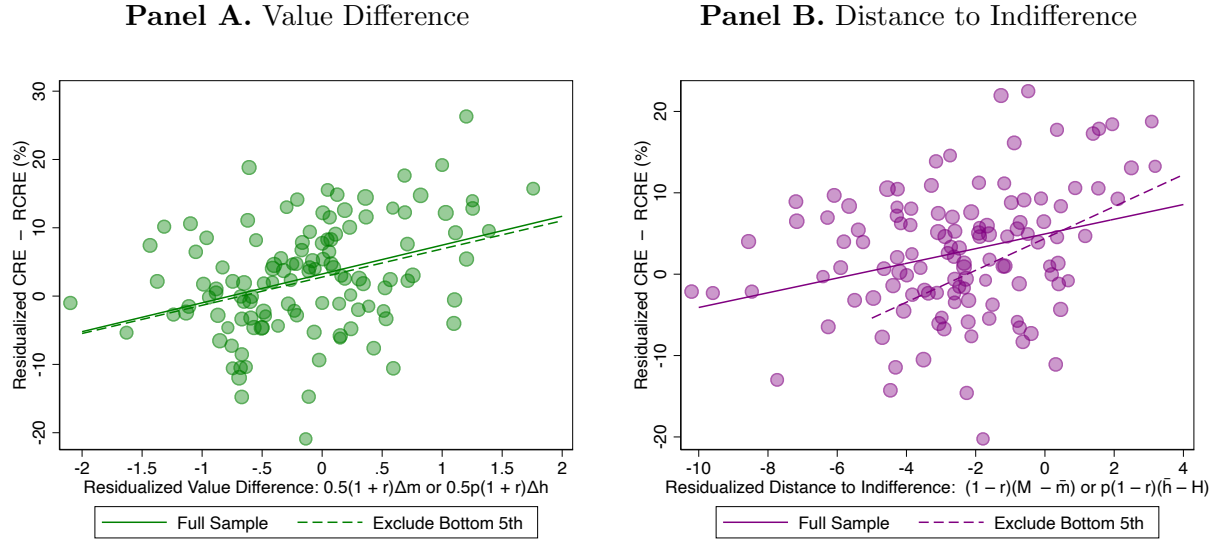

Figure 8: Stage 2 $CRE - RCRE$ as a function of experiment-level stage 1 value differences (panel A) and distance to indifference (panel B). Each dot denotes an experiment, and the size is proportional to the number of observations in that experiment. Panel A observations are residualized by the average 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.

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

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, much 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, and thus it plays an important role in determining whether an experiment will have a CRE or an RCRE.

Panel B of Figure 8 reveals that our experimental variation in M and H leads to some large negative values of the average distance to indifference. Given that Figure 6 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

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

	(1)	(2)	(3)	(4)
	Outcome: $CRE - RCRE \in [-100, 100]$			
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)
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.

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.*

Our analysis of stage 2 data clarifies several key points. First, the data support our hypothesis that differential noise makes inference based on paired choice tasks problematic. In particular, whether one finds a CRE or an RCRE in paired choice tasks depends on whether the experimenter's choice of design parameters—p, r, M, and H—induce a positive or negative distance to indifference. Prior experiments using paired choice tasks do not collect valuations and thus do not have a measure of distance to indifference. However, a striking feature of Figure 1(A) is that more than 75 percent of prior paired choice experiments yield $\widehat{\Pr}(A) > 1/2$, which is indicative of positive distances to indifference. Under the differential noise issue that we document, these are precisely the studies that are more likely to yield a CRE even if there were no underlying CRP.

An analysis of the relationship between $\widehat{\Pr}(A)$ and distance to indifference in our experiments further supports this point. In contrast to prior studies, we deploy a range of parameters across our 120 paired choice experiments that induce a wide range of distances to indifference. In Appendix Figure D.1, we produce the analogue of Figure 1(A) for our 120 experiments. Relative to Figure 1(A), we observe many more observations with $\widehat{\Pr}(A) < 1/2$. Moreover, distance to indifference is

highly predictive of $\widehat{\Pr}(A)$—indeed, every experiment with an average positive distance to indifference exhibits $\widehat{\Pr}(A) > 1/2$, and roughly 2/3 of experiments with an average negative distance to indifference have $\widehat{\Pr}(A) < 1/2$.

A second, more exploratory point, was not expected: Our analysis suggests that, in samples of 75 participants, the unobserved selection of CRP versus RCRP types might also play a role in whether an experiment yields a CRE or an RCRE. Given our finding in stage 1 that individual value differences are centered at zero, it might seem reasonable to assume that a sample of 75 would typically generate an average value difference of zero. In contrast to this intuition, however, the range of average value differences observed in our experiments is large enough to have a substantial impact on whether we observe a CRE or RCRE in typical experimental sample sizes.

Third, the tight connection between our stage 1 valuations and our stage 2 choices at both the individual and the experiment level further validates the preference content in our stage 1 valuations. Related to this, however, we raise one final point. A feature of our data is that participants were more likely to choose the safer option (A or C) in stage 2 than their stage 1 valuations imply. A similar discrepancy between behavior in price lists (as in our valuation task) and behavior in binary choices (as in our choice tasks) has been noted by some prior authors (Beattie and Loomes, 1997; Harrison and Swarthout, 2014; Brown and Healy, 2018; Freeman et al., 2019; Freeman and Mayraz, 2019). Some have interpreted this discrepancy as suggesting that price lists may not deliver reliable measures of true preferences, ostensibly captured in binary choices (Brown and Healy, 2018; Freeman et al., 2019). Others have argued that binary choices, themselves, may yield unreliable preference measures (Freeman and Mayraz, 2019).[39] The systematic connections that we document between stage 1 valuations and stage 2 choices suggests that neither conclusion is warranted. If binary choices were truly unbiased reliable measures of CRP, we would not have the ability to systematically influence the direction of CRE versus RCRE via exogenous variation in distance to indifference. That we show the impact of differential noise on the existence or absence of CRE in choices clarifies that choice-based measures are likely not a gold standard for identifying such EU deviations in preferences. Of course, if choices, themselves, had no useful preference content, then we would not find systematic relationships between CRE in choices and valuations controlling for distance to indifference. That we show tight connections between common ratio behavior across the two decision environments clarifies that choice tasks contain relevant preference information, although valuation tasks yield richer and more accurate measures of any underlying CRP.

6 Discussion

In this paper, we document problems with the standard approach to testing for a CRP, propose and implement an alternative approach based on paired valuation tasks, and find no evidence of a

[39]The view that valuation tasks may be substantially harder to comprehend and therefore lead to less reliable responses is consistent with some earlier views on eliciting indifference (see, e.g., Hey and Di Cagno, 1990; Stott, 2006).

systematic CRP. Hence, we believe that it should no longer be seen as an accepted fact that most people have a CRP, and being able to explain a CRP should no longer be seen as a litmus test for new decision models.

A small set of papers have also used paired valuation tasks in the context of the CRE (Castillo and Eil, 2014; Schneider and Shor, 2017; Dean and Ortoleva, 2019; Freeman et al., 2019). These papers address different research questions, and none of them focus on the role of choice noise in interpreting typical CRE data from paired choice tasks. Nonetheless, there are some interesting connections to our findings.

Dean and Ortoleva (2019) study the correlations between 11 different behavioral patterns related to risk and time preferences. One of these is the CRE, and they use paired valuation tasks to obtain a more continuous measure of the CRE. They find that their measure of the CRE is strongly correlated with other risk-preference-related behaviors, thus suggesting that these preferences reflect underlying individual types. This finding clearly parallels our finding of heterogeneity in underlying CRP versus RCRP.[40]

Schneider and Shor (2017) and Freeman et al. (2019) investigate the robustness of the CRE to different response modes, though for different reasons. The goal of Schneider and Shor (2017) is to compare three different process models that predict different patterns across paired choice tasks, paired pricing tasks (a form of paired valuation tasks), and paired happiness-rating tasks. The goal of Freeman et al. (2019) is to assess whether valuation tasks reveal underlying preferences, where they assume that isolated choice tasks are an accurate measure of underlying preferences. In each paper, the paired valuation task seems to indicate little CRP, much as in our data. Neither paper accounts for choice noise in comparing participants' responses across response modes.[41]

Castillo and Eil (2014) propose a theory of status quo bias in which holding fixed the safe option in a paired valuation task leads to an RCRP, while holding fixed the risky option yields a CRP. They conduct a between-subjects test of this prediction and do find some limited support. Our h- and m-valuation tasks hold fixed the safe and risky options, respectively, and thus offer a broader test of the their theory. Our data involve a much larger sample size and a much broader set of (p, r) combinations, and we observe a fairly consistent pattern across our m- and h-valuations. Additionally, it is challenging to interpret our valuation data as being driven by status quo bias given that our paired valuation measures subsequently predict paired choice tasks that do not have an obvious status quo option.

Our results also have implications for models of probability weighting. On the one hand, our AB valuations yield exactly what we expected given the extensive literature that estimates probability weighting functions using certainty equivalents for binary lotteries. But our CD valuations are

[40]Dean and Ortoleva (2019) conduct two h-valuation tasks with $p = 0.8$ and $r = 0.25$, with $M = \$4$ and $M = \$8$, and they find a mild CRP for both. These tasks are closest to our h-valuation task with $p = 0.8$ and $r = 0.2$, which happens to yield the largest CRP in all of our data.

[41]Both papers consider a single paired valuation task and compare it to a single paired choice task, whereas we implement 30 of each. The lotteries selected by these papers use amounts and probabilities inspired by Problems 3 and 4 in Kahneman and Tversky (1979).

wholly inconsistent with a stable probability weighting function and cast doubt on existing models. One possible reaction is to develop a modified model of probability weighting. For instance, an admittedly post hoc way to explain our data is to assume that people manipulate probabilities prior to applying a weighting function. For example, when presented with a choice between (M, r) and (H, pr), if a person first computes the ratio $pr/r = p$ and only then transforms that ratio into decision weights, then we would get exactly our observed pattern $m_{CD} = m_{AB}$ while still generating an inverse-S-shaped probability weighting function for the AB tasks.

An alternative reaction is to develop a new foundation for the behaviors that have commonly been attributed to probability weighting. For instance, Butler and Loomes (2011) and Enke and Graeber (2021) both suggest that forms of cognitive imprecision might be the source of these behaviors. However, each paper limits attention to a subset of such behaviors: Butler and Loomes (2011) study the implications of imprecision for a few specific violations of independence and betweenness, while Enke and Graeber (2021) study the implications of imprecision only for the structure of certainty equivalents for binary gambles. Neither approach in their current form has clear implications for our CD valuations. Hence, our analysis can be seen as setting a challenge for this approach by providing a new empirical pattern to be explained and, more broadly, by highlighting the need to develop models that can be taken to many types of data.

References

Agranov, M. and Ortoleva, P. (2021). Ranges of randomization. *Working Paper*.

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.

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

Beattie, J. and Loomes, G. (1997). The impact of incentives upon risky choice experiments. *Journal of Risk and Uncertainty*, 14(2):155–168.

Bernheim, B. D. and Sprenger, C. (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. (2022). How common is the common-ratio effect? *Experimental Economics*.

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.

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. (2021). Cognitive uncertainty. *Available at SSRN 3982035*.

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.

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.

Harrison, G. W. and Swarthout, J. T. (2014). Experimental payment protocols and the bipolar behaviorist. *Theory and Decision*, 77(3):423–438.

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

Hey, J. D. and Di Cagno, D. (1990). Circles and triangles: An experimental estimation of indifference lines in the Marschak-Machina triangle. *Journal of Behavioral Decision Making*, 3(4):279–306.

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.

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.

O'Donoghue, T. and Somerville, J. (2018). Modeling risk aversion in economics. *Journal of Economic Perspectives*, 32(2):91–114.

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.