De Gustibus and Disputes about Reference Dependence*

Pol Campos-Mercade Lorenz Goette Thomas Graeber
Alexandre Kellogg Charles Sprenger

January 30, 2026

Abstract

Existing tests of reference-dependent preferences assume universal loss aversion. This paper examines the implications of heterogeneity in gain-loss attitudes for such tests. In experiments on labor supply and exchange behavior, we first measure gain-loss attitudes and then study a canonical treatment effect that distinguishes different models of reference dependence. We document substantial heterogeneity in gain-loss attitudes and evidence against universal loss aversion. Moreover, we find heterogeneous treatment effects over gain-loss attitudes consistent with formulations of expectations-based reference points. Assuming homogeneous preferences would lead to different and potentially incorrect conclusions in these tests. Our findings provide foundational support for reference points derived from expectations and help reconcile inconsistencies in prior empirical exercises.

JEL classification: D81, D84, D12, D03
Keywords: Reference-Dependent Preferences, Rational Expectations, Personal Equilibrium, Real Effort, Expectations-Based Reference Points

*The title alludes to the classic *De Gustibus Non Est Disputandum* conjecture in economics, stating that "tastes neither change capriciously nor differ importantly between people" (Stigler and Becker, 1977). Some of the experimental evidence reported in this paper (on exchange behavior) was included in an earlier paper titled "Heterogeneity of Gain-Loss Attitudes and Expectations-Based Reference Points" `https://papers.ssrn.com/sol3/papers.cfm?abstractid=3170670`. The research described in this article was approved by the Ethics Committee of the Economics Department at the University of Bonn and UC San Diego. The experiments were pre-registered at the AEA RCT Registry (AEARCTR-0007277 and AEARCTR-0003124). Campos-Mercade: Department of Economics, Lund University; pol.campos@nek.lu.se. Goette: Department of Economics, National University of Singapore; ecslfg@nus.edu.sg. Graeber: Harvard Business School, Harvard University; tgraeber@hbs.edu. Kellogg: Department of Economics, University of California, San Diego; alexkellogg@ucsd.edu. Sprenger: California Institute of Technology; sprenger@caltech.edu.

1 Introduction

Models of reference-dependent preferences are regarded as a major advance in behavioral economics, rationalizing a range of observations at odds with the canonical model of expected utility over final wealth (Kahneman, Knetsch and Thaler, 1990; Camerer, Babcock, Loewenstein and Thaler, 1997; Odean, 1998; Rabin, 2000). The predictions of any reference-dependent model hinge on two model components: the reference point governing the location around which gains and losses are encoded; and gain-loss attitudes encapsulating how individuals weigh gains and losses relative to the reference point.

Recent tests of reference-dependent models focus on hypotheses about the first model component, the location of the reference point—distinguishing backward-looking factors such as the status quo posited by Kahneman and Tversky (1979) from the forward-looking expectations-based mechanisms proposed by Bell (1985); Loomes and Sugden (1986); Kőszegi and Rabin (2006, 2007). As the reference point represents a powerful degree of freedom in application, these tests have been valuable for understanding how to discipline reference-dependent models (Abeler, Falk, Goette and Huffman, 2011; Ericson and Fuster, 2011; Smith, 2019; Heffetz and List, 2014; Cerulli-Harms, Goette and Sprenger, 2019; Gneezy, Goette, Sprenger and Zimmermann, 2017). Importantly, the prior empirical exercises have been designed under a specific *homogeneity* assumption on gain-loss attitudes: universal loss aversion, where all individuals weigh losses more severely than commensurate gains. This strikingly strong assumption —a form of the classic *De Gustibus Non Est Disputandum* assumption (Stigler and Becker, 1977)—is an unstated centerpoint of the prior tests. Thus, prior tests have effectively ignored variation in the second model component, gain-loss attitudes.

In this manuscript, we examine the possibility that individuals are *heterogeneous* in their gain-loss attitudes—i.e., individuals are differentially "loss averse", weighing losses more than gains to varying degrees, and some individuals may even be "gain-seeking",

weighing gains more than losses to varying degrees—and explore the implications of this heterogeneity for identifying models of the reference point.[1]

Permitting heterogeneity in gain-loss attitudes in tests of reference-dependent models may be important for two reasons. First, within experimental designs used to identify expectations-based reference dependence (EBRD), different predictions are generated depending on the extent of loss-averse or gain-seeking preference. In the extreme, gain-seeking subjects should react to key experimental treatments in exactly the opposite way as loss-averse subjects. Second, heterogeneity in gain-loss attitudes reflects an empirical realism: a recent literature has noted that even with loss aversion on average, there is substantial variation around the mean with sizable minorities of subjects appearing to be gain-seeking in laboratory experiments. Chapman, Snowberg, Wang and Camerer (2024) evaluate seven prior studies from lottery choice, along with one of the experiments in this manuscript, and report a weighted average of 22% gain-seeking subjects. Furthermore, they find up to 59% of gain-seeking behavior in the general population.[2] This heterogeneity in gain-loss attitudes is stable over time at the individual level, and correlates with important real-life outcomes such as stock-market investment, gambling, and the experience of financial shocks.[3]

[1]The concept of gain-seeking behavior is not merely a theoretical notion; rather, it aligns with observable behavior in specific contexts such as speculative investing and lottery buying, much like the concept of loss aversion.

[2]Chapman et al. (2024) also document substantially fewer gain-seeking individuals in college-aged laboratory samples (approximately one-quarter across their studies). This proportion accords well with the findings we obtain in laboratory samples for labor supply and exchange behavior, where also roughly one-quarter of individuals are classified as gain-seeking.

[3]Additional evidence either documenting or suggestive of heterogeneity in gain-loss attitudes is provided by Mrkva, Johnson, Gächter and Herrmann (2020); Fehr and Kübler (2022); Brown, Imai, Vieider and Camerer (2021); Sprenger (2015); Erev, Ert and Yechiam (2008); Harinck, Van Dijk, Van Beest and Mersmann (2007); Nicolau (2012); Sokol-Hessner, Hsu, Curley, Delgado, Camerer and Phelps (2009); Knetsch and Wong (2009); Chapman, Dean, Ortoleva, Snowberg and Camerer (2017); Abeler et al. (2011); Gill and Prowse (2012); Gneezy et al. (2017). While this body of work documents how differences in gain-loss attitudes correlate with laboratory and real-life outcomes, it does not test whether this heterogeneity modulates treatment effects as predicted by the heterogeneous comparative statics of EBRD models. Closest in spirit to our labor supply study, Gill and Prowse (2012) structurally estimate heterogeneous gain-loss attitudes under the assumption of EBRD to explain choices in a real-effort experiment, but do not subsequently assess heterogeneous treatment effects across different experimental conditions. Their estimates imply that 17% of subjects have gain-seeking preferences.

If individuals are heterogeneous in their gain loss-attitudes and behave in theoretically predicted ways, then prior exercises have aggregated different effects without any way to disentangle heterogeneity in attitudes from the corresponding test of the reference point.[4] The combination of these two issues may explain the inconclusive, and at times contradictory, findings in the study of EBRD models without accounting for heterogeneity.[5]

We implement two pre-registered laboratory experiments with a total of 1524 subjects to investigate the relevance of heterogeneous gain-loss attitudes for testing models of reference-dependent preferences. Our baseline designs and treatment manipulations closely follow existing work on the two main paradigms used to test the EBRD formulation: labor supply (e.g., Abeler et al., 2011; Gneezy et al., 2017) and exchange (e.g., Ericson and Fuster, 2011; Heffetz and List, 2014; Cerulli-Harms et al., 2019). Each experiment consists of two stages. Stage 1 measures each participant's gain-loss attitudes in the specific context of the experiment. Stage 2 tests EBRD by changing subjects' expectations between a Low expectations and a High expectations condition. Under EBRD models, such manipulations change the location of the reference point, and so should change behavior. Under alternative formulations of reference points, no such effects are predicted. Hence, these designs constitute tests of the expectations-based formulation of the reference point.

The EBRD predictions in these two canonical paradigms depend on gain-loss attitudes. Aggregating different effects can lead heterogeneity in gain-loss attitudes to confound the test of EBRD in both settings. Our key experimental innovation is the addition of Stage 1 in order to measure gain-loss attitudes in specific ways that do not interfere with the theoretical predictions and experimental manipulations in each context. These measures allow us to evaluate the extent of heterogeneity in gain-loss attitudes, account for heterogeneity

[4] In Appendices A.3 and B.4 we demonstrate this point concretely. We show that predicted KR treatment effects are not necessarily linear in gain-loss attitudes. Hence, the average treatment effect may not coincide with the treatment effect of the average preference. Indeed, average treatment effects can differ dramatically and even have a different sign from the treatment effect at the average preference.

[5] While early experimental applications showed treatment effects in line with the EBRD formulation of reference points (Abeler et al., 2011; Ericson and Fuster, 2011; Gill and Prowse, 2012), other exercises have shown more limited or contradictory effects (Smith, 2019; Heffetz and List, 2014; Cerulli-Harms et al., 2019; Gneezy et al., 2017).

when testing the EBRD formulation of the reference-point, and examine the heterogeneous treatment effects over gain-loss attitudes predicted by EBRD models. In this sense, we explore whether heterogeneous preferences, *gustus*, can help resolve the outstanding dispute on the nature of reference dependence.

Our two studies generate two consistent sets of results. First, we find substantial heterogeneity in gain-loss attitudes. While subjects in both studies exhibit loss aversion on average, we estimate substantial variation around the mean and sizable minorities of gain-seeking subjects. Under our preferred structural specifications, both studies show around three quarters loss-averse, and one quarter gain-seeking subjects. Furthermore, we document a similar heterogeneity using monetary lottery decisions. The findings from these three different techniques reinforce prior results on heterogeneous gain-loss attitudes in lottery choice (Chapman et al., 2024), and clarify that homogeneous loss aversion would be an incorrect assumption to maintain in tests of reference-dependent models. Our analyses leverage both reduced-form and structural approaches to infer gain-loss attitudes from the data and accommodate potential uncertainty in measurement in various ways.

Second, in each study, gain-loss attitudes from Stage 1 are highly predictive of the treatment effects observed in Stage 2. We document heterogeneous treatment effects over gain-loss attitudes. Higher values of the key loss aversion parameter are associated with larger treatment effects on average. We also document that negative treatment effects are more frequently associated with gain-seeking individuals. In one of the studies, the average treatment effect for gain-seeking individuals is significantly negative. In the other one, contrary to the EBRD predictions, the average treatment effect for gain-seeking individuals—albeit smaller than for loss-averse individuals—remains positive. Overall, we find that differentially loss-averse and gain-seeking subjects respond quite differently to the manipulation of expectations. Without accounting for heterogeneity, we would draw very different conclusions from our studies, finding more limited, or even no, aggregate support for EBRD. However, accounting for it, we find strong evidence for EBRD. This represents the first

experimental test of EBRD accounting for heterogeneous gain-loss attitudes, and the first experimental findings of heterogenous EBRD treatment effects over gain-loss types.

Our empirical results indicate that mixed evidence on EBRD is likely not driven by a failure of the expectations-based formulation of reference points, but rather by a failure of the second component of the joint hypothesis inherent to prior tests: that gain-loss attitudes are homogeneously loss averse. Without accounting for heterogeneous gain-loss attitudes, prior tests suffer from both aggregation and power issues: the average treatment effect need not be the treatment effect of the average individual (which we discuss in detail in Appendices A.3 and B.4), and potentially muted theoretical average effects require larger sample sizes for appropriately-powered experiments. In a simple and reproducible way, we show that several critical predictions of EBRD are recovered once one accounts for heterogeneity in gain-loss attitudes.

While our evidence provides compelling arguments in favor of EBRD, there are also aspects of the data that cannot be fully explained by the model. EBRD posits that both the levels of behavior and the treatment effects in our experiment are exclusively determined by expectations and their influence on reference points. However, in both of our studies we observe quantitative deviations from the model. We discuss these limitations for each of our studies and conclude that the EBRD formulation may overlook some relevant determinants of subjects' behavior beyond expectations. The existing literature puts forward a variety of sources—including status quo-based reference points, attention, anchoring, and cognitive limitations—that may affect behavior. We thus view this evidence as pointing at the potential multiplicity of determinants for behavior, motivating future work that aims to disentangle them.

Above all, this paper highlights the need to account for heterogeneity in gain-loss attitudes in order to use and test models of reference-dependence. Besides tests of expectations-based models, our results also have implications for other applications of gain-loss attitudes, including Rabin's (2000) explanation for risk aversion in the small and in the large, insurance for small losses (Slovic, Fischhoff, Lichtenstein, Corrigan and Combs (1977)), and

preferences for bunched resolution of uncertainty (Kőszegi and Rabin, 2009). The explanations for these phenomena rely on loss aversion. Admitting heterogeneity in gain-loss attitudes will lead to more nuanced predictions in each of these settings. Future work on these phenomena is now equipped with a methodology for investigating and controlling for the influence of heterogeneity in gain-loss attitudes.

The manuscript proceeds as follows. In Section 2, we discuss our two-stage labor supply experiment ($N = 500$), building upon the original designs of Abeler et al. (2011); Gneezy et al. (2017). In Section 3, we discuss our two-stage exchange experiment ($N = 1024$), building on the designs of Ericson and Fuster (2011); Heffetz and List (2014); Cerulli-Harms et al. (2019). Section 4 provides additional discussion and concludes.

2 Labor Supply Experiment

2.1 Experimental Design

The labor supply experiment consists of two stages. In Stage 1, we present subjects with a number of decisions that elicit how much effort they are willing to provide at various piece rates, both fixed and uncertain. The objective is to recover each individual's gain-loss attitudes. In Stage 2, we present subjects with a set of choices that manipulate the implied expectations-based reference point while holding other potential reference points constant, constituting a test of the EBRD formulation.

Stage 1: Measuring Gain-Loss Attitudes. Subjects were informed about the experiment's various parts and the task they would be asked to complete—transcribing a row of blurry Greek text.[6] They went on to complete two practice tasks to familiarize themselves with the process.

[6]The task is borrowed from Augenblick and Rabin (2019).

Next, subjects used a slider to indicate how many of these transcription tasks they were willing to complete at a given piece rate. They were shown the earnings associated with a given number of tasks, as well as an estimate of the corresponding completion time. Each piece rate offering was either fixed, e.g., $w = \$0.20$ per completed task, or uncertain, e.g., a 50% chance of $w_h = \$0.30$ per task and a 50% chance of $w_l = \$0.10$ per task. Subjects made decisions for a total of 30 piece rates, 10 of which were fixed. Each uncertain piece rate was linked to a fixed piece rate with the same mean, i.e., $0.5w_h + 0.5w_l = w$. We rely on these two types of piece rates to identify gain-loss attitudes for each individual accounting for auxiliary parameters such as the shape of their cost function.

On each decision screen, subjects made choices for five different piece rates. On a given decision screen, all offered piece rates were fixed, or all were uncertain. Subjects completed a total of six decision screens which appeared in random order. Similar to Augenblick and Rabin (2019), we selected (expected) piece rates between \$0.05/task and \$0.3/task (an hourly wage rate between approximately \$4.00 and \$26.00, according to subjects' average time of completion).

Stage 2: Experimental Manipulation of Expectations. After completing the Stage 1 choices, we informed subjects that they would make two additional effort decisions with slightly different earnings structures. In these additional decisions, subjects were informed that there would be a 50% chance of a per task piece rate of \$0.20, a p% chance that a fixed payment \$20 would be paid regardless of the number of completed tasks, and a q% chance that a fixed payment \$0 would be paid regardless of the number of completed tasks.[7] Subjects chose a number of tasks to complete in two conditions: Condition Low, where $p = 0.05$ and $q = 0.45$; and Condition High where $p = 0.45$ and $q = 0.05$. Each subject made both decisions in different screens, which were displayed in random order.

In both conditions subjects received a piece rate with 50% chance. With complementary chance, their earnings were unrelated to the number of tasks completed, and were either

[7]These instructions remained purposefully vague about the amounts of money involved as well as any variation over the two choices because our aim was to obtain within-individual comparisons.

Low or High in expectation across the two conditions.[8] Within EBRD models, the Low and High conditions induce different expectations of earnings and so induce different reference points. This, in turn, leads to different willingness to work across the two conditions. In the neoclassical model and in models with exogenous reference points, this manipulation should have no effect on optimal choice.

Lottery Elicitation, Incentives, and Questionnaire. Following the real-effort decisions, subjects evaluated two risky lotteries using Multiple Price Lists (MPLs), a common elicitation technique to measure gain-loss attitudes in the monetary lottery domain. Subjects made a total of 42 monetary lottery choices in two probability equivalent tasks (following Sprenger, 2015) in which we held fixed a sure payoff of \$5 [\$0] and offered the lottery $(p, \$10; 0)$ [or $(p, \$3; -\$3.5)$] with p ranging from 0% to 100% in increments of 5% as the alternative.[9]

Both the labor supply and lottery choices were incentivized. The experimental earnings were based on one of the 32 effort choices or the 42 monetary lottery choices, with each choice having the same chance of being randomly selected to be the *decision-that-counts*. Regardless of which decision or how many tasks were selected, each subject had to complete a minimum of 10 transcriptions. If the decision-that-counts was one of the monetary lottery choices, the computer generated a random number and determined the outcome of the lottery, and the subjects received their payment upon completion of the mandatory tasks and an ensuing survey. If one of the effort decisions was selected for payment, subjects first completed the mandatory 10 tasks and then the additional number they indicated in

[8]This structure allows us to study both within-subject treatment effects by comparing a given subject's answers across conditions and between-subject treatment effects by restricting the sample to only the first condition subjects saw. We pre-registered predictions about within-subject treatment effects in order to maximize statistical power. Appendix Table A5 provides the between-subject results for comparison. While the estimates are noisier, the results are qualitatively similar regardless of the method of analysis.

[9]Assuming subjects have monotonic preferences over money—e.g., they prefer \$5 for sure to a 0% chance of \$10 and prefer a 100% chance of \$10 to \$5 for sure—the p at which they switch from preferring one option to another informs us about their gain-loss attitudes. Within our elicitation, a single switch point was enforced for all subjects.

that decision; if the relevant wage was stochastic, uncertainty in wages was not resolved until after they had completed all of the additional tasks.[10]

After all the tasks were completed, subjects were presented with a series of Raven's matrices (Raven and Raven, 2003) to obtain a measure of cognitive skill, followed by a demographic survey (gender, major, age, parental income, and risk attitudes).

Procedures and Pre-Registration. Our sample for the labor supply experiment consists of 500 subjects recruited through the UC San Diego Economics Laboratory. The experiment was pre-registered at the AEA RCT Registry (Campos-Mercade, Goette, Kellogg and Sprenger (2021), AEARCTR-0007277) and conducted between April and July 2021. On average, subjects earned $15.5. The experiment was implemented in *oTree* (Chen, Schonger and Wickens, 2016). A full set of decision screenshots is provided in Appendix C.

2.2 Identifying Gain-Loss Attitudes and Heterogeneous Theoretical Predictions

We derive theoretical predictions of the seminal Kőszegi and Rabin (2006, 2007) EBRD model in the labor supply context.[11] A reader familiar with the Kőszegi and Rabin (2006, 2007) model may wish to skip this section and proceed directly to the predictions spelled out at the end of Section 2.2.2 or the results presented in Section 2.3. We assume that individual i's utility function is represented by

$$u_i(w, e|r_w, r_e) = m(we) - c_i(e) + \mu_i(m(we) - m(r_w)) + \mu_i(c_i(e) - c_i(r_e)).$$

[10] All subjects had been informed of this procedure in the instructions.

[11] Throughout, our theoretical analysis will use the Kőszegi and Rabin (2006, 2007) formulation. An earlier literature also provided formulations of reference dependence grounded in rational expectations, but without the equilibrium concepts we use to analyze behavior (Bell, 1985; Loomes and Sugden, 1986).

The first component of utility, $m(we) - c_i(e)$, is standard consumption utility obtained from working e tasks and earning we. Consumption utility is complemented with a reference-dependent, psychological component of utility, for which the utility from realized earnings $m(we)$ is compared to the utility of reference-point earnings $m(r_w)$ under a piece-wise linear gain-loss function μ_i, where

$$\mu_i(z) = \begin{cases} \eta z & z \geq 0 \\ \eta \lambda_i z & z < 0 \end{cases} .$$

Intuitively, if an outcome falls short of the reference point by a difference of z, this leads to a reduction of utility by $\eta\lambda_i$ times this difference. An outcome that exceeds the reference point increases utility by η times the difference, where $\eta > 0$. Thus, λ_i represents individual gain-loss attitude and can either exhibit loss-aversion where losses are felt more severely than commensurate gains, $\lambda_i > 1$, or gain-seeking where gains are felt more severely than commensurate losses, $\lambda_i < 1$. If $\lambda_i = 1$, the individual is considered "loss-neutral". Throughout the analysis, we assume that $m(we) = we$ and constant for all individuals, that $c_i(e)$ is an increasing, at least twice-differentiable, strictly convex function, $c_i''(e) > 0$, and normalize $\eta = 1$ for all individuals.

Kőszegi and Rabin (2006, 2007) propose that agents hold the entire distribution of expected outcomes as their referent. Each potential realization is compared to each potential reference point and weighted by the relevant densities. In order to close the model, Kőszegi and Rabin (2006, 2007) equip it with the rational expectations Choice-Acclimating Personal Equilibrium (CPE) concept. Intuitively, a choice is a CPE if the agent's expected utility from this choice given their expectation of this choice as the referent exceeds the expected utility of any alternative choice given the expectation of that alternative choice as the referent. We consider the CPE identification (and estimation) of gain-loss attitudes in Stage 1 of our experimental design, and the CPE comparative statics in Stage 2 of our experimental design.

CPE's requirement of rational expectations implies that reference points immediately adapt to the choices someone makes. While there is evidence that reference points adapt rapidly in experiments (Buffat and Senn, 2015; Song, 2016), it may be natural to consider reference points that are partially set as CPE expectations and partially at exogenous levels. As noted above, models with exclusively exogenous reference points predict null effects for the primary comparative statics that have been used to test EBRD models, including the one we implement for labor supply. In Appendix A.2.1 we demonstrate this point and also consider intermediate cases where reference points are partially endogenous (i.e., CPE) and partially exogenous, which leads to more muted treatment effects in Stage 2. We also explore how partially exogenous reference points affect the identification and estimation of gain-loss attitudes from Stage 1. Notably, in the presence of exogenous reference points, the identification of gain-loss attitudes from labor supply responses to fixed versus stochastic wages is substantially more challenging and could subject our estimation approach to misspecification error. The combination of this misspecification in Stage 1 and limited differences in Stage 2 work against finding reliable individual differences in gain-loss attitudes that are predictive of heterogeneous treatment effects. In Appendix A.2.2 we conduct a simulation exercise varying the relevance of exogenous reference points and assessing the impact on our analysis. Interestingly, the quantitative patterns, including the fraction of gain-seeking individuals, are nearly unchanged as long as there is a substantial expectation-based component of the reference point. Furthermore, the qualitative patterns that we observe in the data are also revealed in the simulations even when reference points have a large exogenous component.

2.2.1 Stage 1 Estimates of Gain-Loss Attitudes

In this subsection we develop our approach for empirically identifying a structural as well as a reduced-form measure of individual gain-loss attitudes.

Consider an uncertain piece rate condition in Stage 1, $(0.5, w_l; 0.5, w_h)$, $w_h > w_l$. The individual chooses effort, e_i, knowing that with 50% chance they will earn either $e_i \times w_l$ or

$e_i \times w_h$. The associated CPE utility for such an effort choice, e_i, is

$$0.5w_l e_i + 0.5w_h e_i - 0.25(\lambda_i - 1)(w_h e_i - w_l e_i) - c_i(e_i).$$

If the individual faces a fixed piece rate, w, then CPE utility reduces to

$$u(we_i|we_i) = we_i - c_i(e_i).$$

In choosing a functional form for the cost of effort, our pre-registered analysis follows Augenblick and Rabin (2019) by assuming $c_i(e_i) = \frac{1}{\alpha_i \gamma_i}(e_i + 10)^{\gamma_i}$ with $\gamma_i > 1$, where 10 represents the required minimum number of tasks that all subjects must complete.[12] Note that this formulation permits individual variation in γ_i and α_i, the parameters of the cost function.

The optimal effort choice, e_i^*, in these two cases thus satisfies the marginal condition

$$\frac{1}{\alpha_i}(e_i^* + 10)^{\gamma_i - 1} = \overline{w} - 0.25(\lambda_i - 1)\Delta w, \tag{1}$$

where $\overline{w}$ is the average wage, such that $\overline{w} = 0.5w_l + 0.5w_h$ for uncertain piece rates, and $\overline{w} = w$ for fixed piece rates; and Δw is the spread in the wage, such that $\Delta w = w_h - w_l$ for uncertain piece rates, and $\Delta w = 0$ for fixed piece rates.

This provides an intuitive formulation for identifying gain-loss attitudes from the sensitivity of behavior to wage spreads, Δw. Loss neutral individuals with $\lambda_i = 1$ make their effort decisions only as a function of the average wage, $\overline{w}$, and, thus, choices are invariant to Δw. Loss-averse individuals with $\lambda_i > 1$ lower their effort in response to increases in

[12] As Augenblick and Rabin (2019) point out: "The parameter α is necessary and represents the exchange rate between effort and the payment amount. If instead $c_i(e_i) = \frac{1}{\gamma_i}(e+10)^{\gamma_i}$, a requirement such as linear marginal costs (which necessitates $\gamma_i = 2$), would also imply that the marginal cost of e_i tasks is exactly e_i monetary units, regardless of the task type or the payment currency." (P. 955)

Δw, all else equal. Conversely, gain-seeking individuals with $\lambda_i < 1$ will increase their effort in response to increases in Δw, all else equal.[13]

This intuition on identification motivates a simple methodology for estimation of gain-loss attitudes. Specifically, taking logs of equation (1), one obtains

$$log(e_i^* + 10) = \frac{log(\alpha_i)}{\gamma_i - 1} + \frac{1}{\gamma_i - 1} log(\overline{w}) + \frac{1}{\gamma_i - 1} log\left[1 + 0.25(1 - \lambda_i)\frac{\Delta w}{\overline{w}}\right]. \qquad (2)$$

Noting that for small values of $0.25(1-\lambda_i)\frac{\Delta w}{\overline{w}}$, the first-order approximation $log\left[1 + 0.25(1 - \lambda_i)\frac{\Delta w}{\overline{w}}\right] \approx 0.25(1 - \lambda_i)\frac{\Delta w}{\overline{w}}$ holds, one can write

$$log(e_i^* + 10) \approx k_i + g_i log(\overline{w}) - l_i \frac{\Delta w}{\overline{w}}, \qquad (3)$$

where

$$k_i = \frac{log(\alpha_i)}{\gamma_i - 1}, \quad g_i = \frac{1}{\gamma_i - 1}, \quad \text{and } l_i = \frac{0.25(\lambda_i - 1)}{\gamma_i - 1}.$$

This formulation is linear in the experimentally-varied parameters $log(\overline{w})$ and $\frac{\Delta w}{\overline{w}}$. Moreover, $l_i \geq 0$ if and only if $\lambda_i \geq 1$, such that l_i provides a sufficient statistic for whether an

[13] Our formulation assumes that utility of money, $m(\cdot)$, is linear. If individuals had diminishing marginal utility of money, one would expect a potential deviation between $e_{i,U}^*$ in the uncertain condition and $e_{i,F}^*$ in the fixed wage condition even if $\lambda = 1$. Indeed, if $m(\cdot)$ were concave, the optimal responses with $\lambda = 1$ would be calculated from marginal conditions

$$m'(we_{i,F}^*)w = \frac{1}{\alpha_i}(e_{i,F}^* + 10)^{\gamma_i - 1}$$

and

$$0.5m'(w_l e_{i,U}^*)w_l + 0.5m'(w_h e_{i,U}^*)w_h = \frac{1}{\alpha_i}(e_{i,U}^* + 10)^{\gamma_i - 1}.$$

These two values will differ to the extent that marginal utility changes over the range $[w_l * e, w_h * e]$. For values of e around 40 tasks and a range of $w_h - w_l \approx 0.1 - 0.2$ this corresponds to a \$4-8 range. Changes in marginal utility over such ranges would have to be dramatic to deliver perceptible effects on behavior and would deliver calibrational implausibilities at larger stakes. Moreover, if one were to attribute differences between $e_{i,U}^*$ and $e_{i,F}^*$ to changes in marginal utility, one would predict null effects (and no heterogeneity) in Stage 2 of our design.

individual is loss-averse or gain-seeking.[14] We will refer to l_i as the reduced form and to λ_i as the structural measure of gain-loss attitudes.

Assuming equation (3) is satisfied with equality subject to a mean zero independent disturbance term, this formulation can be estimated with linear least squares techniques. The corresponding regression estimate for $\hat{l}_i$ captures the response of labor supply to wage uncertainty, and maps closely to a quantitative estimate for λ_i. Indeed, one can give this reduced form estimate a structural interpretation by considering the value $1 + 4 \cdot (\frac{\hat{l}_i}{\hat{g}_i}) \equiv \hat{\lambda}_i$. We topcode estimates of $\hat{\lambda}_i$ from the linear procedure at 3. We do so for two reasons. First, $\lambda_i < 3$ ensures that marginal benefits of effort are strictly positive for all stochastic wages implemented in our study, and thus desired effort levels are positive.[15] Second, the KR model shows limitations for values of $\lambda_i > 3$, generating violations of first-order stochastic dominance.[16] In addition to this issue regarding extreme levels of loss aversion, because values of $\hat{\lambda}_i < 0$ are difficult to interpret within the model, we bottomcode such values at zero.[17]

2.2.2 Heterogeneous Effects of Stage 2 Low vs. High Conditions

In the following we develop our empirical approach for identifying the individual-level treatment effect of our Stage 2 manipulation of expectations.

[14] To see this, note that for $x \in (-1, \infty)$, $log(1 + x) \geq 0 \iff x \geq 0$, with $x = 0.25(1 - \lambda_i)\frac{\Delta w}{w}$.

[15] Equation (1) establishes the marginal benefits of effort as $MB(e) \equiv \overline{w} - 0.25(\lambda_i - 1)\Delta w$. Note that

$$MB(e) > 0 \iff 1 - 0.25(\lambda_i - 1)\frac{\Delta w}{\overline{w}} > 0.$$

The highest value in our labor supply study is $\frac{\Delta w}{\overline{w}} = 2$, requiring $1 - 0.5(\lambda_i - 1) > 0$ or $\lambda_i < 3$. Note as well that as a consequence of negative marginal benefits values of $\lambda_i \geq 3$ would deliver undefined values of $log\left[1 + 0.25(1 - \lambda_i)\frac{\Delta w}{\overline{w}}\right]$ for the case of $\frac{\Delta w}{\overline{w}} = 2$.

[16] When $\lambda_i > 3$, the decision maker prefers zero with certainty over a 50% chance of winning a positive positive amount and 50% chance of remaining with zero. We thank an anonymous referee for this particular illustration of dominance violations.

[17] In order to provide a structural estimate of $\hat{\lambda}_i$ without relying on the linearity approximation of equation (3), one could simply estimate the partially linear regression equation implied by equation (2) via non-linear least squares. We conduct these estimates and compare them to our estimated values of $\hat{\lambda}_i$. The correlation between the values of $\hat{\lambda}_i$ from linear and non-linear procedures is 0.97 for the 443 subjects for whom both are estimable.

We consider how individuals behave when offered an earnings structure $(p, X; q, Y; 0.5, w)$ where $X > Y$; that is, individuals have a 50% chance of earning a piece-rate, w, per unit of effort, a p% chance of earning $\$X$ regardless of effort, and a $q = (50 - p)$% chance of earning $\$Y$ regardless of effort. Following the development of Gneezy et al. (2017), we study the effects of an increase in p when $Y \leq we_i^* \leq X$.[18] In Appendix A.1 we derive the CPE choice, e_i^*, in this case satisfying marginal condition

$$0.5w\left[1 + (p - q)(\lambda_i - 1)\right] = c_i'(e_i^*), \tag{4}$$

and the effect of increasing the probability of the high outcome, p, while keeping $p+q = 0.5$ as

$$\frac{\partial e_i^*}{\partial p}|_{p+q=0.5} = \frac{(\lambda_i - 1)w}{c_i''(e_i^*)}.$$

As the outside possibility unrelated to effort, $(p, X; q, Y)$, increases in expectation, individuals should change their level of effort. The change in effort is governed by λ_i, with $\frac{\partial e_i^*}{\partial p}|_{p+q=0.5}$ increasing in λ_i provided strictly convex costs, $c_i''(e) > 0$. This effect contrasts with that of alternative models of the reference point, where $\frac{\partial e_i^*}{\partial p}|_{p+q=0.5} = 0$. Moreover, the direction of the response is also governed by λ_i with

$$\begin{aligned} \lambda_i > 1 &\implies \frac{\partial e_i^*}{\partial p}|_{p+q=0.5} > 0 \\ \lambda_i < 1 &\implies \frac{\partial e_i^*}{\partial p}|_{p+q=0.5} < 0. \end{aligned}$$

In our implementation we set $X = \$20$, $Y = \$0$, $w = 0.20$, and vary p from 0.05 in the (L)ow condition to 0.45 in the (H)igh condition. Under the assumed functional form $c_i(e_i) = \frac{1}{\alpha_i \gamma_i}(e_i + 10)^{\gamma_i}$, where 10 represents the required tasks, these conditions are

[18] For all other rank cases, there is no predicted treatment effect (see Appendix A.1 for details).

associated with solutions

$$\begin{aligned} e_{i,L}^{*} + 10 &= (\alpha_i 0.10\,[1 - 0.4(\lambda_i - 1)])^{\frac{1}{\gamma_i - 1}} \\ e_{i,H}^{*} + 10 &= (\alpha_i 0.10\,[1 + 0.4(\lambda_i - 1)])^{\frac{1}{\gamma_i - 1}}, \end{aligned}$$

such that the theoretical treatment effect can be expressed in percentage terms as the log difference in effort across the two conditions:

$$TE_i^{*} \equiv log(e_{i,H}^{*} + 10) - log(e_{i,L}^{*} + 10) = \frac{1}{\gamma_i - 1} log\left[\frac{1 + 0.4(\lambda_i - 1)}{1 - 0.4(\lambda_i - 1)}\right]$$

Our second stage focuses on measuring $e_{i,H}$ and $e_{i,L}$ for each subject, thus delivering an empirical analog for this theoretical treatment effect, TE_i^{*}. Similarly, our first stage provides both reduced form and structural measures of the key behavioral parameter λ_i: $\hat{l}_i$, and $\hat{\lambda}_i$. The theoretical formulation above thus leads to the following empirical predictions.

Prediction 1. The empirical treatment effect, TE_i, in the labor supply experiment increases in loss aversion ($\hat{\lambda}_i$).

Prediction 2. The empirical treatment effect, TE_i, in the labor supply experiment is positive for loss-averse individuals ($\hat{\lambda}_i > 1$).

Prediction 3. The empirical treatment effect, TE_i, in the labor supply experiment is negative for gain-seeking individuals ($\hat{\lambda}_i < 1$).

2.3 Results From The Labor Supply Experiment

2.3.1 Stage 1: The Distribution of Gain-Loss Attitudes in Labor Supply

In Stage 1, our 500 subjects each make 30 effort choices, 10 for fixed piece rates and 20 for uncertain piece rates. In Appendix Table A3, we present the mean, median, and interquar-

tile range for each choice. Overall, subjects exhibit increasing labor supply, being willing to complete more tasks for greater fixed piece rates. Importantly, subjects are willing to complete fewer tasks under uncertain piece rates relative to fixed rates of equal mean. Within the context of our KR analysis, this implies loss aversion on average. Appendix Table A3 also documents substantial heterogeneity. At every piece rate, whether fixed or uncertain, the interquartile range covers a wide portion of the choice space. This, in turn, suggests substantial heterogeneity in both costs and gain-loss attitudes.

In order to evaluate the extent of heterogeneity, we estimate the linear regression implied by equation (3). Because this formulation is identical for all subjects with individual values of k_i, g_i, and l_i, we fit the standard random coefficients model of Swamy (1970), which delivers individual estimates of each parameter.[19] Of our 500 subjects, 5.4% (27 subjects) have zero variation in e_i across their 30 effort choices and so no estimates can be obtained. Additionally, 4% (20 subjects) have estimated values of $\hat{g}_i \leq 0$ implying non-convex costs. Removing these observations that cannot be estimated or are prima-facie inconsistent with our pre-registered theory removes a total of 9.4% (47 subjects) of observations, leaving a final sample of 453 subjects.

Panel A of Figure 1 plots the distribution of the reduced form loss aversion measure, $\hat{l}_i$, for our 453 observations. The average estimate of $\hat{l}_i$ is 0.090, while the average estimate of $\hat{g}_i$ is 0.520, and the average estimate of $\hat{k}_i$ is 4.54. Of the 453 subjects, 70.6% exhibit $\hat{l}_i > 0$, indicating loss aversion, while 29.4% exhibit $\hat{l}_i < 0$, indicating gain seeking. Panel B provides the mapping between the reduced form $\hat{l}_i$ and the structural $\hat{\lambda}_i$. The raw correlation between the two values is $\rho = 0.85$ ($p < 0.01$). Additionally, as the theory requires, $\hat{l}_i > 0$ is perfectly diagnostic for $\hat{\lambda}_i > 1$, and so the taxonomy of loss-averse and gain-seeking is identical: 70.6% exhibit $\hat{\lambda}_i > 1$, indicating loss aversion, while 29.4% exhibit $\hat{\lambda}_i < 1$, indicating gain seeking.[20] Panel C plots the distribution of the structural

[19] Our implementation makes use of the random coefficients xtrc command in Stata, recovering individual estimates as the best linear predictors for each individual post estimation and standard errors through the estimated variance covariance matrix of best linear predictors.

[20] Our methodology also produces estimates for the standard error associated with each $\hat{l}_i$. Thus we can also study the proportion of $\hat{l}_i$ that are statistically significantly higher or lower than zero at the 5% level.

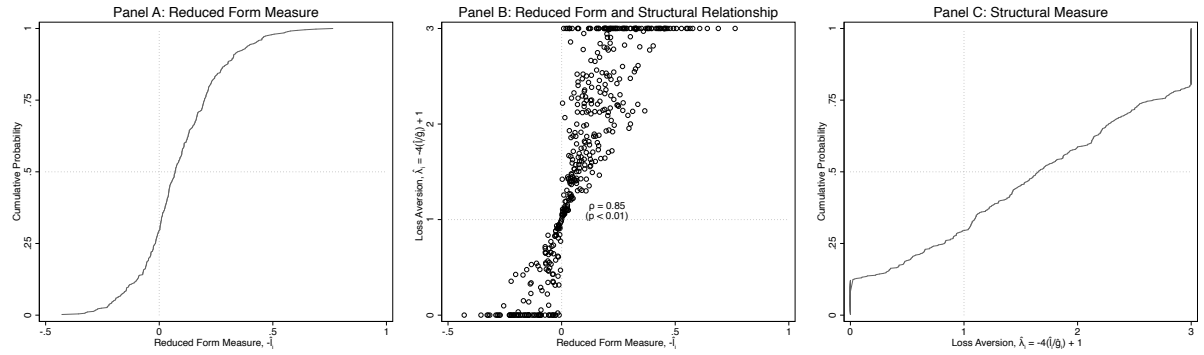

Figure 1: Stage 1: Gain-loss attitudes in the labor supply experiment

Notes: Panel (a) and (c) show CDFs of the reduced form and structural measures of gain-loss attitudes, respectively. Panel (b) displays the relationship between the two measures ($\rho = 0.85,\ p < 0.01$)

measure, $\hat{\lambda}_i$, with an average value of 1.65 and a median value of 1.66. Roughly 12.1% of $\hat{\lambda}_i$ estimates are censored at zero and 19.7% are censored at three, despite the smoothness of the reduced form distribution, $\hat{l}_i$. Individuals with very little or very much sensitivity to the average wage, $\overline{w}$, yield estimates of $\hat{g}_i$ that are either very large or close to zero, respectively, and correspondingly extreme measures of $\hat{\lambda}_i = 1 + 4 \cdot (\frac{\hat{l}_i}{\hat{g}_i})$.

Our findings of substantial heterogeneity in gain-loss attitudes in the labor supply setting echo findings from lottery choice with similar samples. Chapman et al. (2024)'s analysis of prior data indicates median values of $\hat{\lambda}$ between 1.5 and 2.5 and an average of 22% gain-seeking subjects. Having reproduced these heterogeneities, we now turn to the second stage of our design and the examination of EBRD treatment effects.

2.3.2 Stage 2: Heterogeneous Treatment Effects of Low vs. High

We now examine whether individual gain-loss attitudes (estimated from Stage 1 choices) are predictive of individual-level treatment effects (estimated from Stage 2 effort choices).[21]

We find significant positive estimates for 178 subjects (39%), significant negative estimates for 48 subjects (11%), and cannot statistically distinguish the estimates from zero for 227 subjects (50%).

[21]Our main (pre-registered) analysis exploits the within feature of the experiment, leveraging each subject's answers to both Condition Low and Condition High. Appendix Table A5 provides between-subjects analysis using either each subject's first or second choice. As expected, the estimates are noisier.

Analyses of Prediction 1. Figure 2 illustrates the relationship between $\hat{\lambda}_i$ from Stage 1 and treatment effects from Stage 2. We construct fifteen equally sized bins of $\hat{\lambda}_i$ and calculate the average behavior in each bin. Panel A provides an analysis corresponding to Prediction 1, plotting the relationship between $\hat{\lambda}_i$ and individual treatment effects, TE_i. Individuals with greater values of $\hat{\lambda}_i$ have systematically larger treatment effects, consistent with Prediction 1. The raw correlation between TE_i and $\hat{\lambda}_i$ is $\rho = 0.18$ ($p < 0.01$).

Column (1) of Table 1 provides corresponding regression analyses for Prediction 1, controlling for additional factors. The log effort level $log(e_i+10)$ is regressed on an indicator for Condition High, providing an estimate of TE_i. Without accounting for gain-loss attitudes, Condition High is associated with a treatment effect of approximately 0.26 (individual clustered s.e. = 0.03). This corresponds to a roughly 26% increase in effort in Condition High relative to Condition Low, reproducing the findings of Abeler et al. (2011). Importantly, however, the value $R^2 = 0.03$ indicates that much of the variation in behavior is not accounted for in this aggregate analysis.

Panel A of Figure 2 indicates that the aggregate analysis in column (1) of Table 1 masks substantial heterogeneity in treatment effects. In columns (2) and (3), we provide estimates of heterogeneous treatment effects. We interact the indicator for Condition High with the reduced form measure of loss aversion, $\hat{l}_i$, and the structural measure, $\hat{\lambda}_i$, respectively. We additionally control for the other estimated parameters, $\hat{g}_i$ and $\hat{k}_i$ along with their interactions with Condition High. In both columns (2) and (3), we find that within Condition Low, there is a substantial negative correlation between gain-loss attitudes and effort levels: more loss-averse individuals state lower effort levels in Condition Low. This finding is consistent with theoretical predictions laid out in section 2.2.2 in the formula for $e^*_{i,L}$. In both columns (2) and (3), we document a sizable and significant degree of heterogeneity in treatment effects depending on loss aversion. More loss aversion is associated with greater values of TE_i, consistent with Prediction 1. Importantly, when accounting for heterogeneous treatment effects over gain-loss attitudes, along with the additional param-

Heterogeneous treatment effects are more pronounced when examining each subject's second choice, but they move in the predicted direction when examining only first choices as well.

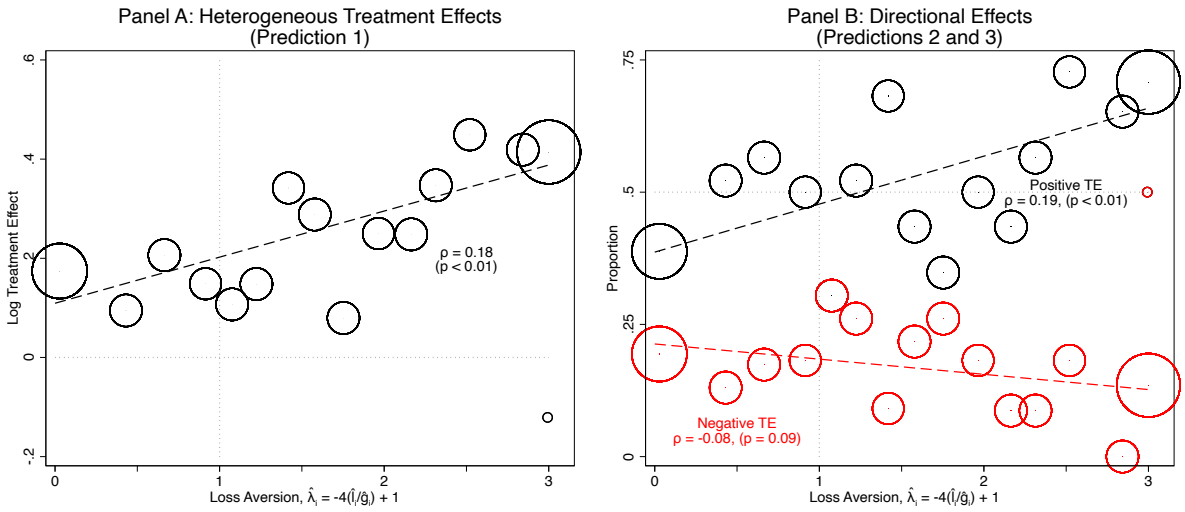

Figure 2: Stage 2: Heterogeneous treatment effects in the labor supply experiment

Notes: Panel A shows the relationship between $\hat{\lambda}_i$ and individual treatment effects, TE_i, in fifteen equally sized bins of $\hat{\lambda}_i$. Panel B plots the relationship between $\hat{\lambda}_i$ and the empirical probability of having either a positive (black markers) or negative (red markers) value of TE_i.

eters $\hat{g}_i$ and $\hat{k}_i$, a substantially greater proportion of behavior is explained; the R^2 values increase by more than a factor of 10.[22]

In addition to the analytical standard errors presented in columns (2) and (3) of Table 1, we also present bootstrap analyses to account for the potential issue of using the values $\hat{l}_i$, $\hat{\lambda}_i$, $\hat{g}_i$, and $\hat{k}_i$ generated from a prior estimation procedure as regressors. This classic 'generated regressor problem' (Murphy and Topel, 2002) could intuitively lead to flawed inference as it treats preference parameters that should be recognized as quantitatively imprecise as ideal data. To overcome this issue, we bootstrap the entirety of Stage 1 estimation and the evaluation of heterogeneity in Stage 2 treatment effects.[23] The resulting

[22]Much of this additional explanatory power derives from the levels of $\hat{g}_i$ and $\hat{k}_i$. Interestingly, consistent with the formula for TE_i, $\hat{g}_i$ (representing the convexity of the cost function, γ_i) is correlated with treatment effects, whereas $\hat{k}_i$ (capturing the level of α_i) is not. For completeness, Appendix Table A4 provides the full table of estimates for Table 1 including those for $\hat{g}_i$ and $\hat{k}_i$ and corresponding interactions with Condition High.

[23]In each iteration of the bootstrap we follow this procedure: 1) sample with replacement to arrive at a data set of the same size as the original (a required 30 observations per subject for 453 subjects); 2) conduct the estimation to arrive at individual values of $\hat{l}_i$, $\hat{\lambda}_i$, $\hat{g}_i$, and $\hat{k}_i$; 3) run the linear regressions associated with Table 1, columns (2) and (3); 4) record coefficients. In each iteration of the bootstrap

Table 1: Heterogeneous treatment effects in the labor supply experiment

Dependent Variable:	Panel A: Prediction 1 *log* $(e+10)$			Panel B: Predictions 2 and 3 *Sign of Treatment Effect*	
	(1)	(2)	(3)	(4)	(5)
Condition High	0.26	0.06 [0.05]	-0.10 [-0.11]		
	(0.03)	(0.15) [(0.15)]	(0.15) [(0.15)]		
Gain-loss attitude: Reduced form ($\hat{l}_i$)		-1.22 [-1.12]		1.55 [1.23]	
		(0.18) [(0.18)]		(0.53) [(0.49)]	
Condition High × Reduced form ($\hat{l}_i$)		0.64 [0.53]			
		(0.19) [(0.18)]			
Gain-loss attitude: Structural ($\hat{\lambda}_i$)			-0.19 [-0.17]		0.33 [0.25]
			(0.03) [(0.03)]		(0.09) [(0.09)]
Condition High × Structural ($\hat{\lambda}_i$)			0.10 [0.08]		
			(0.03) [(0.03)]		
Constant (Condition Low)	3.50	0.92 [0.99]	1.23 [1.32]		
	(0.03)	(0.18) [(0.17)]	(0.17) [(0.17)]		
Controlling for $\hat{g}_i$ and $\hat{k}_i$	No	Yes	Yes	Yes	Yes
R-Squared	0.03	0.40	0.39		
# Individuals	453	453	453	453	453
H_0: Zero TE (High-Low)	$F_{1,452} = 102.66$	$F_{1,452} = 0.17$	$F_{1,452} = 0.48$		
	$(p < 0.01)$	$(p = 0.68)$	$(p = 0.49)$		
H_0: Gain-Loss ⊥ Effort in Low		$F_{1,452} = 44.01$	$F_{1,452} = 49.14$		
		$(p < 0.01)$	$(p < 0.01)$		
H_0: Gain-Loss ⊥ TE		$F_{1,452} = 11.27$	$F_{1,452} = 15.23$	$\chi^2(1) = 8.65$	$\chi^2(1) = 12.61$
		$(p < 0.01)$	$(p < 0.01)$	$(p < 0.01)$	$(p < 0.01)$

Notes: Panel A: Ordinary least squares regression explaining each subject's effort choice. Each subject provides two observations: one with their effort in Condition Low, and one with their effort in Condition High. Panel B: Ordered logit regression for sign of treatment effect. Each subject provides one observation based on the difference between Condition High and Condition Low. Clustered standard errors at the individual level in parentheses. The values in brackets represent bootstrapped estimates based on 500 iterations, where gain-loss attitudes are re-estimated and the regression is re-run in each iteration. The bootstrapping involves resampling both across individuals and within individuals from the 30 choices used to calculate loss aversion. Since some observations must be excluded in certain iterations (as noted in footnote 24), we also report the average coefficient across all bootstrapped values. Each regression also controls for values of $\hat{g}_i$, $\hat{k}_i$, and interactions of each with Condition High. Null hypotheses tested for 1) zero treatment effect (Condition High coefficient= 0); 2) no relationship between gain-loss attitudes and behavior in Condition Low behavior ($\hat{\lambda}_i$ or $\hat{l}_i = 0$); 4) constant treatment effect over gain-loss attitudes (Condition High·$\hat{\lambda}_i$ or Condition High·$\hat{l}_i = 0$). F-statistics, χ^2-statistics and two-sided p-values reported.

average bootstrap coefficient and its standard deviation are presented in brackets in Table 1, columns (2) and (3). The conclusions drawn are identical to those derived from the standard regression analysis. Recognizing the potential for uncertainty in estimated preference parameters is an important factor when conducting exercises of this form, but the conclusions drawn in this setting are not altered.

we may have $\hat{g}_i \leq 0$ or insufficient response variation for some subjects. In such cases, these subjects' observations are dropped. The average bootstrap has observations from 286 subjects (some of whom are sampled more than once).

Analyses of Predictions 2 and 3. Panel B of Figure 2 and columns 4 and 5 of Table 1 provide analyses associated with Predictions 2 and 3: that loss averse individuals will be more likely to have positive treatment effects and gain-seeking individuals will be more likely to have negative treatment effects. Panel B plots the relationship between $\hat{\lambda}_i$ and the empirical probability of having either a positive (black markers) or negative (red markers) value of TE_i. Individuals with $\hat{\lambda}_i > 1$ are systematically more likely than those with $\hat{\lambda}_i < 1$ to exhibit a positive TE_i. The correlation between $\hat{\lambda}_i$ and positive TE_i is $\rho = 0.19$ ($p < 0.01$). In contrast, individuals with $\hat{\lambda}_i < 1$ are somewhat more likely than those with $\hat{\lambda}_i > 1$ to exhibit negative TE_i. The raw correlation between $\hat{\lambda}_i$ and negative TE_i is $\rho = -0.08$ ($p = 0.09$). Using ordered logit regressions, columns (4) and (5) of Table 1 show that both $\hat{l}_i$ and $\hat{\lambda}_i$ are highly predictive of the sign of TE_i (i.e. $\{-1, 0, 1\}$) in both standard and bootstrapped analyses. These findings are directionally consistent with theoretical Predictions 2 and 3.

Limitations. One critical observation to note in Panel B of Figure 2 is that even for $\hat{\lambda}_i < 1$, positive treatment effects are more likely than negative treatment effects. Similarly, in Panel A of Figure 2 the average treatment effects for individuals with $\hat{\lambda}_i < 1$ is slightly positive. These facts are inconsistent with Prediction 3 and the EBRD formulation, which predicts negative treatment effects for such individuals. More broadly, we can assess the consistency of the data with precise theoretical predictions. To do so we calculate the value of $\hat{TE}_i = TE^*(\hat{\lambda}_i)$ for each individual and compare it to its empirical counterpart. Though the two measures are significantly correlated ($\rho = 0.25,\ p < 0.01$), the empirical treatment effects for gain-seeking individuals exceed the theoretical predictions, and the empirical treatment effects for loss-averse individuals fall short of theoretical predictions. Overall, individuals are simply not as sensitive to the difference between Conditions High and Low as their gain-loss estimates would theoretically imply. One possibility is that this lack of sensitivity is driven by the presence of noise: all of our constructs are subject to measurement error, which can attenuate the estimated relationship between them. However, the EBRD formulation of the reference point may also be substantively incomplete,

so that our formulation misses some determinants of subjects' behavior, such as issues related to attention, saliency, and the potential for slower adjustment of reference points.[24] We interpret our findings as showing that while expectations-based reference points are quantitatively important drivers of effort choices, there are likely additional determinants of behavior in our study.

Gain-loss Attitudes Across Domains. Prior work has documented linkages between gain-loss attitudes measured with and without risk, coupling measures of small-stakes risk aversion with exchange behavior in standard endowment effect experiments (see, e.g., Gächter, Johnson and Herrmann, 2022; Dean and Ortoleva, 2015). This work documents sizeable correlations between different measures, ranging from 0.3 to 0.6.

Appendix Figure A4 provides the distribution of gain-loss attitudes calculated using CPE from subjects' lottery choices. The mean and median λ are 1.48 and 1.42, respectively. As in the labor supply setting, we find substantial heterogeneity in gain-loss attitudes across subjects. We classify a sizable minority of 28 percent as gain-seeking. We find that gain-loss attitudes estimated from lottery choices are correlated with the structural estimates of gain-loss attitudes based on labor supply decisions, but not dramatically so (Pearson's $r = 0.091,\ p = 0.03$; Spearman's $\rho = 0.084,\ p = 0.075$). And, we find that our lottery measure of gain-loss attitudes has no predictive power for treatment effects in Stage 2. These findings suggest that though heterogeneity is similar across domains, gain-loss attitudes at the individual level are potentially more domain-specific than generally appreciated.

[24] We would like to thank an anonymous referee for this observation.

3 Exchange Experiment

3.1 Experimental Design

The basic structure of the exchange experiment closely follows that of the labor supply experiment. Stage 1 serves to elicit gain-loss attitudes at the individual level. Stage 2 features a manipulation of expectations adapted to the exchange setting.

Stage 1: Measuring Gain-Loss Attitudes. At the beginning of the experiment, participants saw equally-sized pictures and descriptions of two objects. They were then randomly assigned a private cubicle in which they found one of the two objects. We informed them that the object in front of them was in their possession.[25] After three minutes allotted for inspection of the object, we asked subjects three questions. First, for each object subjects were asked "How much do you like this object?" with a Likert response scale ranging from 0="Not at all" to 8="Very much". Second, for each object they were asked "How much would you want to have this product?" using the same response scale. Third, they were asked "If you had to choose one of the objects, which one would you prefer to keep?". These three unincentivized preference statements are the raw data from which our estimates of gain-loss attitudes are constructed.[26]

After subjects provided their preference statements, the experimenter randomly selected half of all subjects in the session based on a draw from a lotto drum that was clearly visible to all subjects. The experimenter replaced the endowed good with the alternative good for each of the selected subjects. This random replacement of Stage 1 objects was conducted to provide subjects with an experience of probabilistic exchange and to generate exogenous

[25] Crucially, we did not say that they "own" the object, and we asked them to not remove the packaging yet.

[26] Our design decision to use unincentivized preference statements for estimating Stage 1 gain-loss attitudes was motivated by a desire to focus on just a single experimental choice in Stage 2. Analytically this avoids subjects considering their suite of experimental choices in both stages as their CPE strategy. One may worry about the reliability of unincentivized preference statements. However, given the predictive power of these preference statements for predicting choices over other objects, these worries are allayed by the data.

variation in the objects obtained in Stage 1. We informed subjects at the end of Stage 1 that they now own the object in their possession.

Stage 2: Experimental Manipulation of Expectations. The procedures in Stage 2 were purposefully similar to those in Stage 1. In a separate room, subjects saw pictures and descriptions of two objects—different from those used in Stage 1. Upon returning to their private cubicle they would find one of the two new objects, which we again assigned randomly. We study two between-subjects conditions, with randomization at the session level.[27] In both conditions, subjects decide whether they would like to retain their assigned object or exchange it. The two conditions differ in the probability that exchange will be forced regardless of their statement. In Condition Low, subjects face a 0% chance that exchange will be forced. That is, this condition is equivalent to a standard exchange setting common to endowment effect experiments. In Condition High, subjects are forced to exchange their object with 50% chance regardless of choice. The chance of forced exchange was based on a draw from a lotto drum that was visible to all subjects. Within EBRD models, the Low and High conditions induce different expectations of the final object to be obtained and so induce different reference points. This, in turn, leads to different willingness to exchange across the two conditions. In the neoclassical model or models with backwards looking reference points, the probability of forced exchange should have no effect on optimal choice.

Procedures and Pre-Registration. The objects used for the exchange experiment were a USB stick, a set of three erasable pens, a picnic mat, and a thermos.[28] We selected these four objects on the basis of a pre-experimental survey evaluation of 12 candidate objects. We put particular emphasis on ruling out complementarities between items across rounds. The former two (USB stick and pens) and the latter two objects (picnic mat and

[27] We present our analysis with robust standard errors in the main text and Appendix Table A11 reproduces our results with standard errors clustered at the session level. Statistical significance is enhanced with clustering, and so we decided to provide the more conservative values in the main text.

[28] Pictures and information presented to subjects are reproduced in Appendix D.

thermos) each constituted a pair. Across the two stages, each subject encountered each pair of objects exactly once. The use of each pair as the Stage 1 pair was counterbalanced at the session level.

The total sample for the exchange experiment consists of 1024 subjects recruited from the BonnEconLab at University of Bonn in Germany. In total, 59 percent (603 of 1024 subjects) were randomly assigned to Probabilistic Forced Exchange. An initial sample of 607 subjects participated in June and July 2015, and a pre-registered replication sample of a further 417 subjects participated in July 2018 (Goette, Graeber, Kellogg and Sprenger, 2018, AEARCTR-0003124).[29] Subjects received a participation fee of 6 euros and also two of the four objects used in the experiment according to their endowments, choices, and chance. A full set of screenshots for our experiment, implemented in *ztree* (Fischbacher, 2007), can be found in Appendix D.

3.2 Identifying Gain-Loss Attitudes and Heterogeneous Theoretical Predictions

We again derive theoretical predictions using the Kőszegi and Rabin (2006, 2007) EBRD model, now applied to the exchange setting with two objects. A reader familiar with the Kőszegi and Rabin (2006, 2007) model may wish to skip this section and proceed directly to the predictions spelled out at the end of Section 3.2.2 or the results presented in Section 3.3. We consider individual i's two-dimensional utility function over object X and object Y,

$$u_i(\mathbf{c}|\mathbf{r}) = m_X + m_Y + \mu_i(m_X - r_X) + \mu_i(m_Y - r_Y),$$

where $\mathbf{c} = (m_X, m_Y)$ refers to consumption utility associated with the quantity of each object, and $\mathbf{r} = (r_X, r_Y)$ similarly refers to reference utility. Thus, an individual's utility

[29] While the experiment carried out in June and July 2015 was not pre-registered, the one carried out in July 2018 was pre-registered. In the main body of the paper we pool the results from both experiments, but Appendix Table A10 shows that the main results replicate in both samples. There were a few very minor differences between the original sessions and those in the replication, which are also described in Appendix B.6

function consists of two components: consumption utility, $m_X + m_Y$, and gain-loss utility, $\mu_i(m_X - r_X) + \mu_i(m_Y - r_Y)$. We let $m_X, r_X \in \{0, X\}$, and $m_Y, r_Y \in \{0, Y\}$ denote both the outcome and the corresponding utility of zero or one unit of object X, and zero or one unit of object Y, respectively. For our primary analysis we assume utilities, X and Y to be homogeneous in the population, as our design does not allow us to simultaneously estimate heterogeneity in gain-loss utility and consumption utility. As before, we assume piecewise linear gain-loss attitudes, with potential heterogeneity in loss-aversion or gain-seeking, λ_i, and $\eta = 1$ for all individuals. We consider the estimation of gain-loss attitudes in Stage 1 of our experimental design, and the CPE comparative statics in Stage 2 of our experimental design.

3.2.1 Stage 1 Estimates of Gain-Loss Attitudes

In Stage 1 of our design subjects are explicitly endowed with an object and then asked to provide preference statements about that object and an alternative. These statements are made without knowledge of any possibility of actual exchange. Hence, theoretically, the reference point is fixed at the endowed object.[30] An individual endowed with X will state a preference in the form of a higher liking value for X, higher wanting value for X, or hypothetical choice of X if $u_i(X, 0|X, 0) - u_i(0, Y|X, 0) > \delta$, where δ captures the possibility of equal rating levels.[31] Under our functional form assumptions such a preference statement occurs if

$$(1 + \lambda_i) - 2\frac{Y}{X} - \delta_X > 0,$$

[30]Though implausible given our design, potential alternative formulations might be to assume that subjects believe they can change their reference point from X to Y or to assume subjects consider retaining their endowed object, X, *and* gaining the alternative, Y (evaluating utility of Y as $X + (1 + \eta)Y$). Importantly, both of these formulations would imply that Stage 1 statements reveal no information on gain-loss attitudes, λ_i. Hence, both would yield null predictions for heterogeneous treatment effects in Stage 2. As such, the results we document invalidate these formulations.

[31]Note that $\delta = 0$ for our hypothetical choice data as there was no possibility of stating indifference.

where $\delta_X \equiv \frac{\delta}{X}$. Similarly, an individual endowed with X would state a preference for Y if

$$2\frac{Y}{X} - (1 + \lambda_i) - \delta_X > 0.$$

An individual would state equal preferences if neither inequality were satisfied. These two equations provide an intuitive formulation for identifying gain-loss attitudes. Controlling for the relative utility of the two objects, $\frac{Y}{X}$, an individual with a greater value of λ_i should be more likely to prefer their endowment and less likely to prefer the alternative.

This simple intuition on identification motivates a reduced-form measure of gain-loss attitudes based on residual preference for endowed objects. First, we conduct a principal components analysis on the three preference statements in Stage 1 and reduce the data to the first principal component. Within our data the first component captures around 70 percent of the variation in relative wanting, relative liking, and hypothetical choice statements. We then regress this component on Stage 1 object assignment. The residuals of this regression summarize a residual preference for the endowed or the alternative object accounting for the average preference. An individual who disproportionately likes their assigned object relative to average preferences is plausibly more loss averse than one who exhibits a residual in the opposite direction. Hence, we consider these residuals as a reduced form measure of gain-loss attitudes, $\hat{l}_i$.

For this analysis, we assume that residual preference for the assigned object is solely reflective of gain-loss attitudes. Alternatively, one could consider residual preference for the assigned object to be reflective of heterogeneity in the intrinsic relative utility $\frac{Y}{X}$. Importantly, if heterogeneity in Stage 1 behavior were driven by variation in $\frac{Y}{X}$, by design we would have zero predictive power for Stage 2. In Stage 2, subjects are randomly assigned to different objects than Stage 1, rendering their prior behavior irrelevant. Any classification of $\hat{l}_i$, or its structural counterpart discussed below $\hat{\lambda}_i$, would be spurious, a product of specification error and orthogonal to Stage 2 treatment effects. This interpretation of the data is rejected by the strong heterogeneous treatment effects observed in Stage 2. At-

tributing behavior partially to heterogeneity in relative good values similarly works against our obtained effects.

In order to provide a structural estimate of the parameter, λ_i, we make the following assumptions. First, rather than assuming deterministic choice, we posit that an individual endowed with X will state a relative preference for X with probability

$$\pi_{X|X} = Prob((1 + \lambda_i) - 2\frac{Y}{X} - \delta_X > \epsilon),$$

a relative preference for Y with probability

$$\pi_{Y|X} = Prob(2\frac{Y}{X} - (1 + \lambda_i) - \delta_X > \epsilon),$$

and, where appropriate, would provide equal ratings for the two objects with probability $\pi_{E|X} = 1 - \pi_{X|X} - \pi_{Y|X}$. Symmetric formulations are assumed for individuals endowed with object Y.

Second, we assume $Prob(\cdot)$ is the logistic function, and that λ_i is drawn from a log-normal distribution with $log(\lambda_i) \sim N(\mu_\lambda, \sigma_\lambda^2)$, leading to a mixed logit formulation.

Third, we assume that there exists a deterministic component of relative utility, $\frac{Y}{X}$, homogeneous in the population. Within this structure, the parameter ϵ can be interpreted as capturing idiosyncratic variation in relative utility or noisiness in response. Thus, we parametrically admit idiosyncratic variation around the mean relative utility, $\frac{Y}{X}$, that we estimate. Unfortunately, our three observations—hypothetical choice, relative liking, and relative wanting—generate data limitations that prevent us from estimating a distribution of relative good values alongside the distribution of gain-loss attitudes. Intuitively, the three observations in the data support the estimation of three parameters of interest: μ_λ, σ_λ^2, and $\frac{Y}{X}$. This limitation means that systematic individual variability in relative good values may lead to mis-estimation of the distribution of gain-loss attitudes. As noted above,

given random assignment of new objects in Stage 2, such misspecification works against our obtained heterogeneity in treatment effects.[32]

Fourth, and last, we fix $\delta_X = 0.55$. This value was selected based on prior aggregate estimates without heterogeneity in gain-loss attitudes from our initial analysis of the exchange experiment (the original working paper is available at `https://papers.ssrn.com/sol3/papers.cfm?abstractid=3170670` and Appendix B.6 reproduces the prior estimates). Due to the same data limitations noted above, we cannot provide an estimate of δ_X alongside the distribution of gain-loss attitudes and relative utility. Importantly, we found substantial sensitivity for the value σ_λ^2 to different restrictions on δ_X.[33] The challenge is intuitive: a larger value of δ_X implies individuals should more frequently give the two objects equal ratings. All else equal, a higher variance of gain-loss attitudes is required to justify the relative infrequency of such observations.

With the above assumptions in hand, we estimate the parameters of the distribution of gain-loss attitudes $N(\hat{\mu}_\lambda, \hat{\sigma}_\lambda^2)$ based on Stage 1 data.[34] Moving from the estimated distribution of gain-loss attitudes to an expected value of $\hat{\lambda}_i$ for each individual is a straightforward step. As proposed in Train (2009), from the estimated distribution, $N(\hat{\mu}_\lambda, \hat{\sigma}_\lambda^2)$, we simulate the distribution of gain-loss attitudes and the corresponding distributions of preference statements. We then calculate the expected simulated value of loss aversion for each possible combination of Stage 1 preference statements and use this as our measure of $\hat{\lambda}_i$ for

[32] In Appendix B.2 we consider the possibility of homogeneous gain-loss attitudes and estimate heterogeneous relative good values using an analogous estimation structure. These estimates provide substantially worse fit. Additionally, they predict a constant treatment effect in Stage 2, in sharp contrast with the data.

[33] Appendix Table A7 and Figure A9 provide analysis setting δ_X at several different values and demonstrating corresponding sensitivity for the variance of gain-loss attitudes and other parameters. The estimated variance of gain-loss attitudes is notably sensitive to varying the assumptions on δ_X. Moreover, the likelihood is effectively flat across different specifications, reinforcing the essential data limitations noted above.

[34] It is also straightforward to alter the assumptions of this formulation to estimate heterogeneity in intrinsic utilities, $\frac{Y}{X}$, rather than gain-loss attitudes. Such an exercise is presented in Appendix B.2, and yields estimates of aggregate loss aversion and substantial variation in object valuations. As noted above, interpreting Stage 1 measures as being driven by heterogeneous utilities rather than heterogeneous gain-loss attitudes leads to the prediction of no heterogeneous treatment effects in Stage 2, and thus is rejected by the data.

all subjects who make a given combination of Stage 1 statements.[35] Appendix B.2 provides additional details and Appendix Table A8 provides examples of the corresponding mappings from preference statements to $\hat{\lambda}_i$.

3.2.2 Heterogeneous Effects of Stage 2 Low vs. High Conditions

Consider Condition Low, in which subjects are asked whether endowed with object X they prefer X or Y. In this setting, the two potential CPE selections are $\{(X,0),(0,Y)\}$ (the first reflecting the choice to keep, and the second the choice to exchange). The individual can support keeping their endowed object in a CPE if $u_i(X,0|X,0) \geq u_i(0,Y|0,Y)$. Given our assumptions, this condition is satisifed for all values of X at or above a Low condition threshold $X \geq X_{Low,i} = Y$. That is, the individual can support keeping their endowed object if it has weakly greater consumption utility than the alternative.[36]

Next, consider the environment in Condition High. With probability 0.5, the agent, assumed endowed with X, will be forced to exchange X for Y regardless of their choice. If the individual wishes to retain their object, they are subject to a stochastic reference point, as with probability 0.5 their object will be exchanged regardless of their choice. Now, the potential CPE selections for someone endowed with X are $\{0.5(X,0)+0.5(0,Y),(0,Y)\}$, with the first element reflecting attempting to keep the endowed object and the second reflecting exchange, as before. They can support attempting to keep their object as a CPE

[35]The recovered distribution is dependent on the assumption that the preference statements are evaluated at the homogeneous consumption utility ratio for each choice set. Naturally, our recovered distribution of gain-loss attitudes rests upon the aforementioned assumptions; any given individual could have consumption preferences strong enough to flip our estimate of λ_i from gain seeking to loss averse. However, this is orthogonal to the predictive ability of our Stage 1 gain-loss parameter on the Stage 2 treatment effect.

[36]It has been noted before that the CPE formulation predicts that individuals exchange in standard endowment effect designs only on the basis of consumption utility, and so fails to predict an endowment effect. The Kőszegi and Rabin (2006, 2007) EBRD model is also equipped with several alternative equilibrium concepts and refinements, Personal Equilibrium (PE) and Preferred Personal Equilibrium (PPE), the former of which can rationalize and endowment effect. Importantly, PE, PPE, and CPE all share common comparative statics for the change from Low to High conditions: loss-averse individuals should grow more willing to exchange in High relative to Low, while gain-seeking individuals should grow less willing to exchange in High relative to Low. Appendix B.1 presents all three forms of the Kőszegi and Rabin (2006, 2007) model's application to this design for completeness.

if

$$u_i(0.5(X,0)+0.5(0,Y)|0.5(X,0)+0.5(0,Y)) \geq u_i(0,Y|0,Y),$$

which, under our functional form assumptions, requires that X be at or above a revised threshold

$$X \geq X_{High,i} = \frac{1+0.5(\lambda_i-1)}{1-0.5(\lambda_i-1)}Y.$$

The manipulation of probabilistic forced exchange changes the CPE threshold for not exchanging from $X_{Low,i} = Y$ in Condition Low to $X_{High,i} = \frac{1+0.5(\lambda_i-1)}{1-0.5(\lambda_i-1)}Y$ in Condition High.

Note that the value of λ_i determines the difference between the thresholds in Low and High conditions. If individuals are loss-averse, $\lambda_i > 1$, then $X_{Low,i} < X_{High,i}$. If higher values for object X are required to support not exchanging in Condition High, this implies that loss-averse individuals should be more willing to exchange in High than in Low. In contrast, if individuals are gain-seeking $\lambda_i < 1$, then $X_{Low,i} > X_{High,i}$, and gain-seeking individuals are less willing to exchange in High than in Low. The empirical analogs for these theoretical relationships are the focus of our experiment.

We define the *Treatment Effect (TE)* as the percentage of individuals who exchange in Condition High minus those who exchange in Condition Low. The development above leads to the following empirical predictions for heterogeneous treatment effects.

Prediction 4. The empirical treatment effect, TE, in the exchange experiment increases in loss aversion ($\hat{\lambda}_i$).

Prediction 5. The empirical treatment effect, TE, in the exchange experiment is positive for loss-averse individuals ($\hat{\lambda}_i > 1$).

Prediction 6. The empirical treatment effect, TE, in the exchange experiment is negative for gain-seeking individuals ($\hat{\lambda}_i < 1$).

3.3 Results From The Exchange Experiment

3.3.1 Stage 1: The Distribution of Gain-Loss Attitudes in Exchange.

Fifty-seven percent of subjects state that they would hypothetically choose their endowed object, 45 percent provide a higher liking rating for their endowed object compared to 33 percent for the alternative, and 45 percent provide a higher wanting rating for their endowed object compared to 32 percent for the alternative. The different preference statements are remarkably correlated within individual. The pairwise Pearson correlations between hypothetical choice, relative liking, and relative wanting statements all exceed 0.7.

Given random assignment of endowed objects and the counterbalanced design, the distributions of preference statements should, in principle, be identical between endowed and alternative objects. Instead, all three distributions show a clear preference for the subject's endowed object relative to the alternative. For each measure we reject the null hypothesis that stated preferences are equal over the endowed and alternative objects.[37] These collected preference statements show a clear endowment effect, and so are indicative of loss aversion on average. However, we also document substantial heterogeneity. Thirty-eight percent of subjects (385 of 1024) state that they would hypothetically choose, strictly like, and strictly want their endowed object. And, twenty-six percent of subjects (262 of 1024) exhibit the opposite pattern of hypothetically choosing, strictly liking, and strictly wanting the alternative object. While this heterogeneity in statements likely partly reflects variation in valuations for the different objects, the predictive power of these Stage 1 statements for Stage 2 behavior with different randomly assigned objects demonstrates that an important component is driven by heterogeneous gain-loss attitudes.[38]

[37]Two sided t-tests comparing "Endowed>Alternative" to "Alternative>Endowed" are significant for all statements (Liking: $t = 5.48$, Wanting: $t = 5.86$, Hypothetical Choice: $t = 6.06$, $p < 0.01$ for all comparisons).

[38]Subjects preferring their endowed object aligns with the endowment effect, while a preference for the alternative object suggests a negative endowment effect. Assuming no differences in object valuation, these results align with the 30% negative endowment effects documented by Chapman et al. (2017).

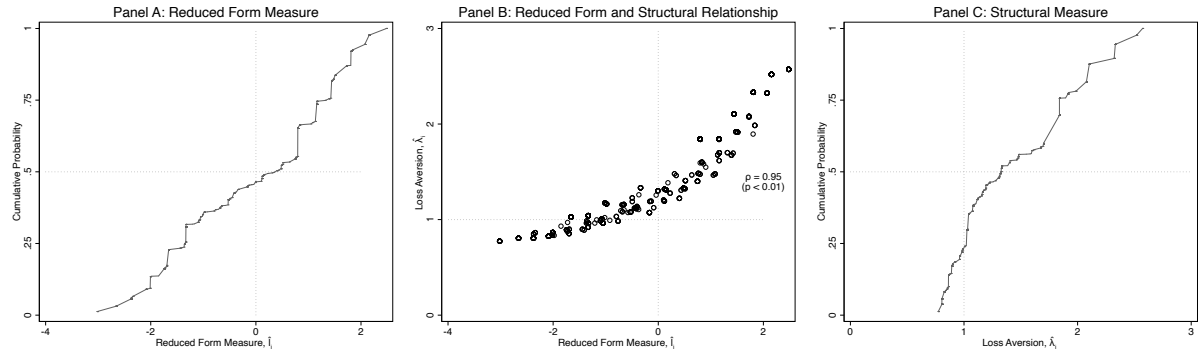

Figure 3: Stage 1: Gain-loss attitudes in the exchange experiment

Notes: Panel (a) and (c) show CDFs of the reduced form and structural measures of gain-loss attitudes, respectively. Panel (b) displays the relationship between the two measures ($r = 0.95,\ p < 0.01$)

Figure 3 shows the distributions of our reduced form and structural measures of gain-loss attitudes associated with the heterogeneity in Stage 1 preference statements, along with the relationship between the two. As in the labor supply study, we document substantial variation in gain-loss attitudes, irrespective of which measure we rely on. Appendix Table A8 provides the structural estimates for the distribution of gain-loss attitudes, $N(\mu_\lambda, \sigma^2_\lambda)$, alongside the auxiliary parameters for relative utilities, $\frac{Y}{X}$, for each pair of objects; and Appendix Table A8 provides the mapping from preference statements to individual estimates of $\hat{\lambda}_i$ under these estimates. Within our sample, $\hat{\lambda}_i$ has mean 1.49 and median 1.34. In line with our labor supply findings, we calculate that 76% of subjects are loss-averse, $\hat{\lambda}_i > 1$, while 24% are gain-seeking $\hat{\lambda}_i < 1$. Also as in our labor supply experiment, we observe a strong correlation between the reduced form and structural measures of gain-loss attitudes (Pearson's $r = 0.95,\ p < 0.01$).

3.3.2 Stage 2: Heterogeneous treatment effects of Low vs. High

Stage 1 behavior, measured with one pair of objects for each subject, delivers estimates of gain-loss attitudes that can be used to analyze Stage 2 choices in the Low and High conditions measured with a different pair of objects. Figure 4 provides a visual illustration of the connections between Stage 1 gain-loss attitudes and Stage 2 behavior. In both

panels, we construct 15 equally spaced bins of Stage 1 $\hat{\lambda}_i$ and connect this measure of gain-loss attitudes to a relevant choice or treatment effect in Stage 2 to test Predictions 4-6.

Analyses of Prediction 4. Figure 4 Panel A documents the relationship between $\hat{\lambda}_i$ and estimated treatment effects: we estimate larger treatment effects among subjects with greater values of $\hat{\lambda}_i$. This connection between gain-loss attitudes and treatment effects is closely in line with the theoretical implications of EBRD and Prediction 4.

Table 2 provides corresponding regression results for Prediction 4. In column (1), we regress the likelihood of exchanging in Stage 2 on an indicator for Condition High without accounting for heterogeneous gain-loss attitudes. In Condition Low, 38 percent of subjects choose to exchange. Comparing this value to the neoclassical benchmark of 50 percent indicates a significant endowment effect in Condition Low, $F_{1,1022} = 25.66$, $(p < 0.01)$. The estimated coefficient on the indicator for Condition High is 0.00 (clustered s.e. $= 0.03$), showing that the substantial endowment effect observed in Condition Low is unaffected by probabilistic forced exchange on average. In contrast to the prediction of EBRD models with universal loss aversion (which would predict a positive treatment effect), we fail to reject that this treatment effect is different from zero.

The precisely estimated aggregate null effect in Table 2 column (1) masks substantial heterogeneity in treatment effects over gain-loss attitudes. Without accounting for heterogeneous gain-loss attitudes, the average treatment effect reported in column (1) potentially aggregates different-signed effects of loss-averse and gain-seeking subjects. This aggregation creates a number of theoretical issues challenging reliable identification.[39] Columns (2) and (3) accommodate the heterogeneous treatment effects suggested by Figure 4: we

[39]In Appendix B.4, we show that in the exchange setting the relationship between λ_i and treatment differences for exchange probability can be concave, with the negative effects for gain-seeking individuals being of greater absolute magnitude than the positive effects for loss-averse individuals. This leads to substantial aggregation issues in our setting as the average treatment effect may be substantially understated relative to the treatment effect of the average preference. This may help to explain why the average treatment effect is indeed null.

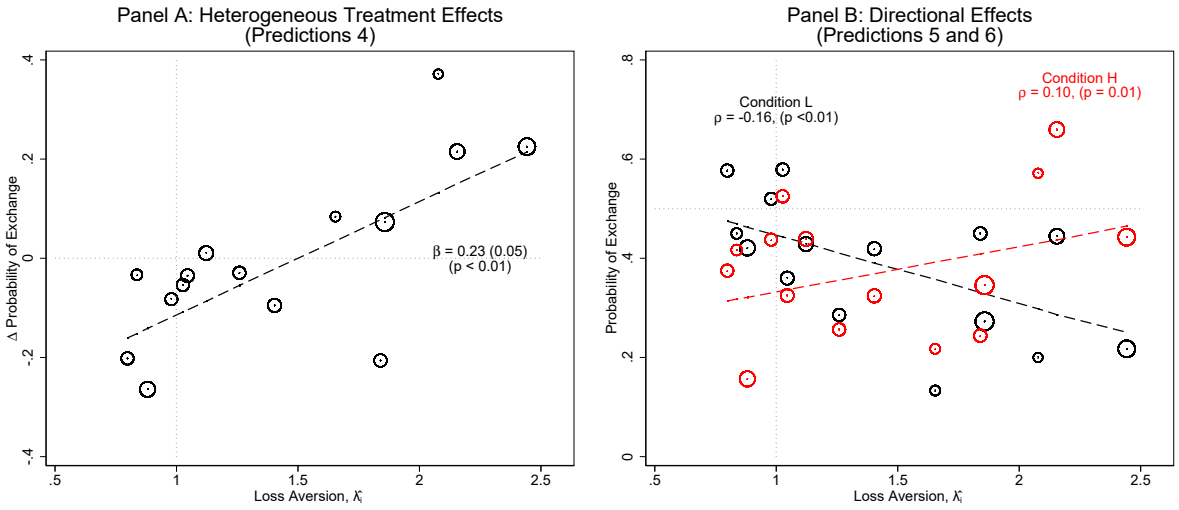

Figure 4: Stage 2: Heterogeneous treatment effects in the exchange experiment

Notes: Panel A shows the relationship between $\hat{\lambda}_i$ and treatment effects in fifteen equally sized bins of $\hat{\lambda}_i$. Panel B plots the relationship between $\hat{\lambda}_i$ and willingness to exchange in Condition Low (black markers) and Condition High (red markers).

interact treatment with reduced form and structural measures of gain-loss attitudes. Both measures are highly positively correlated with the effect of treatment, consistent with Prediction 4. More loss averse subjects have greater increases in their willingness to exchange as they move from the Low to the High condition. Consistent with EBRD, individuals respond to the change in expectations across Low and High conditions, and differentially so depending on their gain-loss attitudes. Alternative formulations of the reference point predict zero treatment effect and zero heterogeneity therein, and, thus, are rejected by our exchange study results. Importantly, as in the labor supply experiment, when accounting for heterogeneous treatment effects over gain-loss attitudes, a substantially greater proportion of behavior is explained; the R^2 values increase by more than a factor of 10.

In addition to the standard regressions presented in columns (2) and (3) of Table 2, we also present bootstrap analyses to account for the potential issue of using the values $\hat{l}_i$ and $\hat{\lambda}_i$, generated from prior estimation procedures, as regressors. As in the labor supply study, we bootstrap the entirety of Stage 1 estimation and the evaluation of heterogeneity in Stage 2 treatment effects. The resulting average bootstrap coefficient and its standard

deviation are presented in brackets in Table 2, columns (2) and (3).[40] The conclusions from the bootstrap analyses are qualitatively similar to the original analysis.

Analyses of Prediction 5 and 6. Panel A of Figure 4 also provides analyses associated with Predictions 5 and 6: that the level of the treatment effect is positive for loss-averse individuals but negative for gain-seeking individuals. We find that individuals with $\hat{\lambda}_i > 1$ are systematically more likely than those with $\hat{\lambda}_i < 1$ to exhibit a positive treatment effect. Panel A shows negative estimated treatment effects for all bins with $\hat{\lambda}_i < 1$, and positive treatment effects for 55% (6 out of 11) of the bins with $\hat{\lambda}_i > 1$.

To shed light on the drivers of the heterogeneous treatment effect in Panel A, Panel B of Figure 4 plots the empirical frequency of exchanging separately for the High and Low conditions, for the 15 different bins of $\hat{\lambda}_i$. First, we observe a negative relationship in Condition Low: more loss-averse subjects are less likely to exchange their endowment for the alternative, $\rho = -0.16$ ($p < 0.01$). Second, this relationship reverses in Condition High: willingness to exchange increases in $\hat{\lambda}_i$, $\rho = 0.10$ ($p < 0.01$). Within the Kőszegi and Rabin (2006, 2007) model's CPE construct, the positive correlation between $\hat{\lambda}_i$ and willingness to exchange in Condition High is predicted. However, the negative correlation between $\hat{\lambda}_i$ and willingness to exchange in Condition Low lies outside the CPE formulation; exchange in Condition Low should be independent of gain-loss attitudes under CPE. Interestingly, however, in their model's alternative Personal Equilibrium (PE) construct, this correlation is admitted (see Appendix B.1.3 for details). Given that some portion of our observed heterogeneous treatment effects falls outside of the CPE framing, our results may speak to the relevance of this alternative equilibrium construct.[41]

[40] Given the computational intensity of the task, we limit the analysis to 500 bootstrap iterations. Not every bootstrap for the mixed-logit estimation converged, and some bootstraps delivered extreme outlier regression coefficients for the Condition High treatment effect. Column (3) thus presents bootstrapped coefficients and standard errors winsorized at 5th and 95th percentile for the Condition High treatment effect (conditional on converging), yielding 425 total bootstraps.

[41] Note that Table 2, column (3), documents endowment effects among subjects with high $\hat{\lambda}_i$, as the exchange probability in Condition Low is below 0.5. Conversely, it also documents negative endowment effects among subjects with very low values of $\hat{\lambda}_i$, for whom their exchange probability would be larger than 0.5.

Table 2: Heterogeneous treatment effects in the exchange experiment

Dependent Variable:	*Exchange* (= 1)		
	(1)	(2)	(3)
Condition High	-0.00	-0.00 [-0.00]	-0.34 [-0.38]
	(0.03)	(0.03) [(0.03)]	(0.09) [(0.15)]
Gain-loss attitude: Reduced form ($\hat{l}_i$)		-0.05 [-0.05]	
		(0.02) [(0.02)]	
Condition High × Reduced form ($\hat{l}_i$)		0.08 [0.08]	
		(0.02) [(0.02)]	
Gain-loss attitude: Structural ($\hat{\lambda}_i$)			-0.14 [-0.16]
			(0.04) [(0.08)]
Condition High × Structural ($\hat{\lambda}_i$)			0.22 [0.26]
			(0.05) [(0.12)]
Constant (Condition Low)	0.38	0.38 [0.38]	0.58 [0.61]
	(0.02)	(0.02) [(0.02)]	(0.07) [(0.10)]
R-Squared	0.00	0.01	0.02
# Individuals	1024	1024	1024
H_0 : Zero TE (High-Low)	$F_{1,1022} = 0.01$	$F_{1,1020} = 0.02$	$F_{1,1023} = 15.07$
	$(p = 0.91)$	$(p = 0.90)$	$(p < 0.01)$
H_0 : Gain-Loss $\perp$ Exchange in Low		$F_{1,1020} = 10.69$	$F_{1,1023} = 11.35$
		$(p < 0.01)$	$(p < 0.01)$
H_0 : Gain-Loss $\perp$ TE		$F_{1,1020} = 14.65$	$F_{1,1023} = 17.23$
		$(p < 0.01)$	$(p < 0.01)$

Notes: Ordinary least squares regression explaining each subject's decision to exchange their object. Values in brackets correspond to bootstrapped values from 500 bootstraps re-estimating gain-loss attitudes and reconducting regression in each bootstrap. Not every bootstrap for the mixed-logit estimation converged, and some bootstraps delivered extreme outlier regression coefficients for the Condition High treatment effect. Column (3) thus presents bootstrapped coefficients and standard errors winsorized at 5th and 95th percentile for the Condition High treatment effect (conditional on converging), yielding 425 total bootstraps. Null hypotheses tested for 1) zero treatment effect (Condition High coefficient= 0); 2) no relationship between gain-loss attitudes and behavior in Condition Low behavior ($\hat{\lambda}_i$ or $\hat{l}_i = 0$); 3) constant treatment effect over gain-loss attitudes (Condition High × $\hat{\lambda}_i$ or Condition High × $\hat{l}_i = 0$). F-statistics and two-sided p-values reported.

Limitations. One important observation to note in Panel A of Figure 4 is that we document negative treatment effects even for some bins of $\hat{\lambda}_i > 1$. This is inconsistent with the EBRD formulation and Prediction 5, which predicts positive treatment effects for all loss-averse individuals. By contrast, we document slightly negative levels of treatment effects for individuals with low levels of loss aversion, i.e. those that are estimated to be close to gain-loss neutrality. The linear fit shown in Panel A suggests a crossing point from negative to positive treatment effect at a level of loss aversion of $\hat{\lambda}_i \approx 1.5$, rather than at 1 as predicted by the theory. Similar to our findings on labor supply, one possibility is this reflects the inherent noisiness of our estimates of loss aversion and empirical estimates of treatment effects. Another possibility that we openly embrace is that the EBRD formulation of the reference point is incomplete, and that there are additional drivers of behavior in our study—such as status quo-based reference points, attention, anchoring, and cognitive limitations—for which we cannot account.

In sum, the results on the heterogeneity of gain-loss attitudes and its predictive power for the behavioral effect of a shift in the expectations-based reference point are compatible with the labor supply experiment. That said, due to the sensitivity of some structural findings to estimation choices—such as δ_X—we caution against drawing strong conclusions about the degree of similarity in the distribution of gain-loss attitudes across the two domains.

4 Conclusion

Prior work testing reference-dependent preferences assumes universal loss aversion. This paper studies the role of heterogeneity in gain-loss attitudes, and explores its implications for identifying models of the reference point. Failing to acknowledge heterogeneity in gain-loss attitudes is critical both because comparative statics used to test different formulations of the reference point can change sign depending on the level of gain-loss attitudes and because such heterogeneity is an empirical reality. In two laboratory experiments, we show

that once one accounts for heterogeneity in gain-loss attitudes, experimental tests are generally supportive of Expectations-Based Reference Dependence (EBRD) formulations of reference points.

Our large-sample pre-registered experiments show that the existing body of evidence on heterogeneity in gain-loss attitudes is not a mere artifact of measurement error or behavioral noise. Instead, by showcasing its out-of-sample predictive power, we document that gain-seeking behavior has a substantive interpretation that can be productively used in theory testing. The consistency of our findings across our two experimental settings attests to the robustness and importance of recognizing heterogeneity.

Conceptually, the importance of recognizing parameter heterogeneity in identifying behavioral predictions hinges on two issues: non-linearity in aggregation and statistical power. First, treatment effects need not aggregate linearly over the dimension of heterogeneity, so ignoring heterogeneity can confound inference. The severity of this concern differs by model and context, and we, ourselves, show a potentially more pronounced aggregation problem in our study of exchange behavior than in our study of labor supply. Similar concerns have been highlighted in other decision domains such as intertemporal choice (Weitzman (2001); Jackson and Yariv (2014)). Second, even under linear aggregation, heterogeneity influences power considerations. An empirical study that is theoretically well-powered under the assumption of preference homogeneity may be under-powered if there is actual heterogeneity, which may lead to false conclusions from null findings. Both issues are of first-order importance for interpreting empirical tests of theories that likely feature parameters with real-world heterogeneity.

There is no universally accepted measurement of gain-loss attitudes, and each candidate has unique advantages and potential drawbacks. In the two designs presented in this manuscript, we elicit gain-loss attitudes in markedly different ways. In our labor supply study, we estimate gain-loss attitudes both from a large number of incentivized labor supply decisions and lottery choices. We treat each decision as isolated for the purposes of estimating gain-loss attitudes. Such approaches facilitate estimation, but fail to account

for the possibility that the reference point (EBRD or otherwise) depends upon the entire body of choice problems. In our exchange behavior study, by contrast, we estimate gain-loss attitudes from hypothetical non-choice data, circumventing this challenge but creating the concern that the measures are not incentivized. Importantly, key features of the estimated distribution of gain-loss attitudes are compatible across the different domains and measurement techniques. Whether measured using incentivized labor supply, lottery choices, or hypothetical exchange choices, around three quarters of subjects are measured to be loss averse and one quarter gain seeking.

Though we provide results on the role of EBRD in the two main paradigms used to test models of reference-dependent preferences, the considerations that motivate this paper equally apply to the role of gain-loss attitudes in other classes of theories and applications. Heterogeneity matters not only for tests of non-expectations-based forms of reference dependence, such as current or backward-looking elements (e.g., Bowman, Minehart and Rabin (1999)), but also for other field settings in which loss aversion has been shown to play a role, such as job search (DellaVigna, Lindner, Reizer and Schmieder (2017)), insurance choice (Barseghyan, Molinari, O'Donoghue and Teitelbaum (2013)) or tax compliance (Engström, Nordblom, Ohlsson and Persson (2015)). It may also provide new insights into how environmental factors shape gain-loss preferences (Fehr, Fink and Jack, 2022).

Beyond the context of gain-loss attitudes, our work contributes to a growing literature in behavioral economics that acknowledges the importance of (structurally) recognizing heterogeneity in behavioral parameters (see DellaVigna (2018) for a recent review). Our paper shows that taking the theoretical implications of heterogeneity seriously—instead of treating it as a nuisance—can deliver more comprehensive tests of behavioral theories and potentially reconcile conflicting evidence.

Data Availability Statement

The data and code underlying this research is available on Zenodo at:

https://dx.doi.org/10.5281/zenodo.16755528

References

Abeler, Johannes, Armin Falk, Lorenz Goette, and David Huffman, "Reference points and effort provision," *The American Economic Review*, 2011, pp. 470–492.

Augenblick, Ned and Matthew Rabin, "An experiment on time preference and misprediction in unpleasant tasks," *Review of Economic Studies*, 2019, *86* (3), 941–975.

Barseghyan, Levon, Francesca Molinari, Ted O'Donoghue, and Joshua C Teitelbaum, "The nature of risk preferences: Evidence from insurance choices," *American economic review*, 2013, *103* (6), 2499–2529.

Bell, David E., "Disappointment in Decision Making under Uncertainty," *Operations Research*, 1985, *33* (1), 1–27.

Bowman, David, Deborah Minehart, and Matthew Rabin, "Loss aversion in a consumption–savings model," *Journal of Economic Behavior & Organization*, 1999, *38* (2), 155–178.

Brown, Alexander L, Taisuke Imai, Ferdinand Vieider, and Colin Camerer, "Meta-analysis of empirical estimates of loss-aversion," *Available at SSRN 3772089*, 2021.

Buffat, Justin and Julien Senn, "Testing the speed of adjustment of the reference point in models of expectation-based reference-dependent preferences," *Available at SSRN 2526089*, 2015.

Camerer, Colin, Linda Babcock, George Loewenstein, and Richard Thaler, "Labor supply of New York City cabdrivers: One day at a time," *The Quarterly Journal of Economics*, 1997, pp. 407–441.

Campos-Mercade, Pol, Lorenz Goette, Alexandre Kellogg, and Charles Sprenger, "Reference-Dependent Effort Provision under Heterogeneous Loss Aversion: Pre-Analysis Plan," *AEA RCT Registry. March 02*, 2021.

Cerulli-Harms, Annette, Lorenz Goette, and Charles Sprenger, "Randomizing Endowments: An Experimental Study of Rational Expectations and Reference-Dependent Preferences," *American Economic Journal - Microeconomics*, February 2019, *11* (1), 185–207.

Chapman, Jonathan, Erik Snowberg, Stephanie W Wang, and Colin Camerer, "Looming large or seeming small? Attitudes towards losses in a representative sample," Technical Report, National Bureau of Economic Research 2024.

_ , Mark Dean, Pietro Ortoleva, Erik Snowberg, and Colin Camerer, "Willingness to Pay and Willingness to Accept are Probably Less Correlated Than You Think," Technical Report, National Bureau of Economic Research 2017.

Chen, Daniel L., Martin Schonger, and Chris Wickens, "oTree - An open-source platform for laboratory, online, and field experiments," *Journal of Behavioral and Experimental Finance*, March 2016, *9*, 88–97.

Dean, Mark and Pietro Ortoleva, "Is it all connected? A testing ground for unified theories of behavioral economics phenomena," *A Testing Ground for Unified Theories of Behavioral Economics Phenomena (August 13, 2015)*, 2015.

DellaVigna, Stefano, "Structural behavioral economics," in "Handbook of Behavioral Economics: Applications and Foundations 1," Vol. 1, Elsevier, 2018, pp. 613–723.

_ , Attila Lindner, Balázs Reizer, and Johannes F Schmieder, "Reference-dependent job search: evidence from Hungary," *The Quarterly Journal of Economics*, 2017.

Engström, Per, Katarina Nordblom, Henry Ohlsson, and Annika Persson, "Tax compliance and loss aversion," *American Economic Journal: Economic Policy*, 2015, *7* (4), 132–64.

Erev, Ido, Eyal Ert, and Eldad Yechiam, "Loss aversion, diminishing sensitivity, and the effect of experience on repeated decisions," *Journal of Behavioral Decision Making*, 2008, *21* (5), 575–597.

Ericson, Keith M. Marzilli and Andreas Fuster, "Expectations as Endowments: Evidence on Reference-Dependent Preferences from Exchange and Valuation Experiments," *The Quarterly Journal of Economics*, 2011, *126* (4), 1879–1907.

Fehr, Dietmar and Dorothea Kübler, "The endowment effect in the general population," 2022.

_ , Günther Fink, and B Kelsey Jack, "Poor and rational: Decision-making under scarcity," *Journal of Political Economy*, 2022, *130* (11), 2862–2897.

Fischbacher, Urs, "z-Tree: Zurich toolbox for ready-made economic experiments," *Experimental economics*, 2007, *10* (2), 171–178.

Gächter, Simon, Eric J Johnson, and Andreas Herrmann, "Individual-level loss aversion in riskless and risky choices," *Theory and Decision*, 2022, *92* (3), 599–624.

Gill, David and Victoria Prowse, "A Structural Analysis of Disappointment Aversion in a Real Effort Competition," *The American Economic Review*, 2012, *102* (1), 469–503.

Gneezy, Uri, Lorenz Goette, Charles Sprenger, and Florian Zimmermann, "The Limits of Expectations-Based Reference Dependence," *Journal of the European Economic Association*, 2017.

Goette, Lorenz, Thomas Graeber, Alexandre Kellogg, and Charles Sprenger, "Heterogeneity of Loss Aversion and Expectations-Based Reference Points: Replication," *AEA RCT Registry. July 05*, 2018.

Harinck, Fieke, Eric Van Dijk, Ilja Van Beest, and Paul Mersmann, "When gains loom larger than losses reversed loss aversion for small amounts of money," *Psychological Science*, 2007, *18* (12), 1099–1105.

Heffetz, Ori and John A. List, "Is the Endowment Effect an Expectations Effect?," *Journal of the European Economic Association*, 2014, *12* (5), 1396–1422.

Jackson, Matthew O and Leeat Yariv, "Present bias and collective dynamic choice in the lab," *American Economic Review*, 2014, *104* (12), 4184–4204.

Kahneman, Daniel and Amos Tversky, "Prospect Theory: An Analysis of Decision under Risk," *Econometrica*, 1979, *47* (2), 263–292.

_ , Jack L. Knetsch, and Richard H. Thaler, "Experimental Tests of the Endowment Effect and the Coase Theorem," *Journal of Political Economy*, 1990, *98* (6), 1325–1348.

Knetsch, Jack and Wei-Kang Wong, "The endowment effect and the reference state: Evidence and manipulations," *Journal of Economic Behavior & Organization*, 2009, *71* (2), 407–413.

Kőszegi, Botond and Matthew Rabin, "A Model of Reference-Dependent Preferences," *The Quarterly Journal of Economics*, 2006, *121* (4), 1133–1165.

_ and _ , "Reference-Dependent Risk Attitudes," *The American Economic Review*, 2007, *97* (4), 1047–1073.

_ and _ , "Reference-Dependent Consumption Plans," *The American Economic Review*, 2009, *99* (3), 909–936.

Loomes, Graham and Robert Sugden, "Disappointment and Dynamic Consistency in Choice under Uncertainty," *Review of Economic Studies*, 1986, *53* (2), 271–82.

Mrkva, Kellen, Eric J Johnson, Simon Gächter, and Andreas Herrmann, "Moderating loss aversion: Loss aversion has moderators, but reports of its death are greatly exaggerated," *Journal of Consumer Psychology*, 2020, *30* (3), 407–428.

Murphy, Kevin M and Robert H Topel, "Estimation and inference in two-step econometric models," *Journal of Business & Economic Statistics*, 2002, *20* (1), 88–97.

Nicolau, Juan L, "Asymmetric tourist response to price: loss aversion segmentation," *Journal of Travel Research*, 2012, *51* (5), 568–676.

Odean, Terrance, "Are Investors Reluctant to Realize Their Losses?," *The Journal of Finance*, 1998, *53* (5), 177–1798.

Rabin, Matthew, "Risk Aversion and Expected Utility Theory: A Calibration Theorem," *Econometrica*, 2000, *68* (5), 1281–1292.

Raven, John and Jean Raven, *Raven Progressive Matrices*, Boston, MA: Springer US,

Slovic, Paul, Baruch Fischhoff, Sarah Lichtenstein, Bernard Corrigan, and Barbara Combs, "Preference for insuring against probable small losses: Insurance implications," *Journal of Risk and insurance*, 1977, pp. 237–258.

Smith, Alex, "Lagged Beliefs and Reference-Dependent Utility," *Journal of Economic Behavior & Organization*, 2019, *167*, 331–340.

Sokol-Hessner, Peter, Ming Hsu, Nina G Curley, Mauricio R Delgado, Colin F Camerer, and Elizabeth A Phelps, "Thinking like a trader selectively reduces individuals' loss aversion," *Proceedings of the National Academy of Sciences*, 2009, *106* (13), 5035–5040.

Song, Changcheng, "An experiment on reference points and expectations," *Available at SSRN 2580852*, 2016.

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

Stigler, George J. and Gary S. Becker, "De Gustibus Non Est Disputandum," *The American Economic Review*, 1977, *67* (2), 76–90.

Swamy, Paravastu AVB, "Efficient inference in a random coefficient regression model," *Econometrica: Journal of the Econometric Society*, 1970, pp. 311–323.

Train, Kenneth E., *Discrete Choice Methods with Simulation*, Cambridge University Press, 2009.

Weitzman, Martin L, "Gamma discounting," *American Economic Review*, 2001, *91* (1), 260–271.