Structural Analyses of Behavioral Errors: The Case of Risk and Time Preferences

LUKMAN HAKIM

DOCTORAL THESIS

Structural Analyses of Behavioral Errors: The Case of Risk and Time Preferences

---

Author:
Lukman Hakim

Supervisors:
Prof. Hong Il Yoo
Prof. Morten Igel Lau

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

in the

Department of Economics and Finance
Durham University Business School
Durham University

July 2022

Structural Analyses of Behavioral Errors: The Case of Risk and Time Preferences

Lukman Hakim

Abstract

This thesis revives the interest in behavioral errors by evaluating their application in the decision making under risk in the first essay and observing their association to the econometric modeling of decision heuristics in time preferences in the second essay. The first essay considers task complexity as the underlying source of randomness in the decision making under risk. We evaluate three heteroskedastic models that specify the disturbance's standard deviation as a function of *ad hoc* measures of task complexity. Our contribution is to adapt a stochastic model from the consumer behavior discipline to binary choice tasks over lotteries, which is the most common experimental method to elicit risk preferences. Our empirical results emphasize the importance of accommodating the impact of task complexity on behavioral errors by rejecting homoskedasticity in favor of at least two heteroskedastic models. Our analyses suggest that the models' statistical goodness-of-fit is contributed by their ability to capture broader risk behaviors at individual level, which is useful for distinguishing decision makers. The second essay critically evaluates the decision heuristic models, which have been claimed to have better accuracy in explaining the individual's time preferences than the structural discounting models. Our contention is that the seemingly superior performance of the heuristic discounting models is irrelevant to their ability to capture particular aspects of discounting behaviors but more an artefact of their oversimplified econometrics modeling. Their specification exhibits a linear approximation of a finite mixture of two behavioral error stories, which is not comparable to the structural discounting specifications with the representative agent model. The heuristic specifications can then be used as a simple diagnostic device for choosing between behavioral error specifications for the structural models. Finally, we contribute to the economic literature by showing that neglecting utility curvature completely reverses inferences about the relative performance of heuristic and structural discounting models.

Supervisors: Prof. Hong Il Yoo and Prof. Morten Igel Lau

Table of Contents

Declaration

This thesis is submitted in fulfillment of requirements for the degree of Doctor of Philosophy in Durham University Business School at the University of Durham, Durham, United Kingdom. I declare that this thesis is based on my own original work except for quotations and citations, which I have duly acknowledged. Where information has been derived from other sources, I confirm that this has been indicated in the work. The work in this thesis is based on research carried out in regular consultation with academic supervisors at the Department of Economics and Finance of Durham University Business School, Durham University, UK. No part of this thesis has been submitted elsewhere for any other degree or qualification.

List of Tables

Chapter 1

Chapter 2

List of Figures

Chapter 1

Chapter 2

Acknowledgements

I could not have completed my PhD journey without the unfailing support and guidance of my supervisors, Professor Hong Il Yoo and Professor Morten Igel Lau. My deepest gratitude goes to both for sharing their insights, expertise, and time with me. It has been a free gift that I cannot repay, but one that I will always remember.

I much acknowledge the Financial support from Durham University Business School and the Centre for Experimental Methods in Business Research (EMBR) to pursue this PhD Programme.

My final words of gratitude go to my parents and my family, who always give their prayers and support throughout my lifetime. Finally and foremost, I am thankful to my wife, Restu Prabandini, for her unlimited support, love, and care, and to my son Rauf and my daughter Rumi who always cheered me up and motivated me to fly halfway across the globe to pursue higher education. I would not have reached this far without all of you.

Introduction

It was evident in a sizeable experimental study that choice under risk or time preferences exhibits a strong stochastic component. Decision makers who mostly behave according to a deterministic theory may sometimes make contradict choices, be it due to random fluctuations in the states of mind, "slip of the finger" when reviewing computerized choice screens, or intended randomization (Agranov and Ortoleva [2017] and Dwenger, Kübler and Weizsäcker [2018]). In the context of data analysis from discrete choice experiments, several behavioral error stories, more formally known as stochastic models, have been proposed. Then structural estimation of economic theories proceeds by combining a non-linear index function that is based on a particular theory with a stochastic model. The subjects of extensive enquiry in applying a model of stochastic choice has been primarily to elicit the structural parameters of the deterministic model (see, for example, Andersen, Harrison, Lau, and Rutström [2008] and Lau and Yoo [2021]), to test one theory against another (Loomes and Sugden [1995], Ballinger and Wilcox [1997], and Buschena and Zilberman [2000]), and to measure welfare cost of the preference randomness (Alekseev, Harrison, Lau, and Ross [2019]).

In a similar vein, this thesis makes behavioral errors as its central theme. We analyze data from an artefactual field experiment with a representative sample randomly drawn from the adult Danish population to study the application of behavioral error stories in risk and time preferences. The first essay of this thesis is titled "Modeling the Effects of Decision Complexity on Choice Behavior under Risk." The essay considers three heteroskedastic error models, namely the Contextual Utility, the Decision Field Theory, and the Entropy model, which specify the disturbance's standard deviation as a function of a measure of task complexity. We empirically evaluate the three heteroskedastic error models relative to the homoskedastic Fechner specification in terms of their performance in explaining observed choices and predicting the choices in the

out-of-sample data. We use maximum simulated likelihood to estimate the full statistical model that accounts for unobserved preference heterogeneity and find that the estimation of risk parameters under EUT is robust across alternative stochastic models. However, we find some sensitivities in the risk parameters under RDU. Our in-sample analyses and out-of-sample predictions emphasize the need to allow behavioral error to vary with task complexity by rejecting homoskedastic Fechner in favor of at least two heteroskedastic models. Our analysis is distinctive as we go beyond the goodness-of-fit comparison by elucidating the background of the models' relative performance using the information on the individual-level parameter derived from the hierarchical Bayes procedure. We find that the superior performance of a stochastic model can be explained by its ability to capture a more heterogeneous behavior in terms of utility curvature or decision weight, as they contribute to higher utility differences or lower behavioral errors, or a combination of the two. Those factors lead to a contrast preference, thus a higher probability of choosing a particular option, unless the decision maker is indifferent between the available options. As the choices at points far from indifference are mostly consistent, the model generates higher log-likelihood values and better fitness to the data.

The title of the second essay is "An Econometric Analysis of Intertemporal Choice Heuristics." The essay scrutinizes the claim of the better accuracy of linear index models of intertemporal choice heuristics in predicting the individual's time preferences relative to the structural discounting models. Using incentivized field experiment data, we are able to emulate the main inference based on hypothetical responses in the other studies by Ericson, White, Laibson and Cohen [2015] and Wulff and van den Bos [2018]. However, our use of a more modern approach to structural analyses uncovers two issues in their relative efficacy. The first issue is related to this thesis' main theme. Our analytic observation finds that the econometric modeling of the heuristic models implies a stochastic structure akin to what one may induce by combining the structural

discounting models with a finite mixture of two behavioral error stories, namely Fechner and Luce errors. Our contribution to the literature is related to the second issue. While plentiful experimental literature emphasizes the necessity to allow for non-linear utility function to remedy an upward bias in the estimated discount rates, the studies of decision heuristics assume linear transformation of the monetary reward. In accordance with these analytic observations, we empirically find a relatively weaker performance of the heuristic discounting models once we adopt a mixture error specification or control for utility curvature, or a combination of the two, even relative to the most inflexible specification in the exponential discounting model.

This thesis is structured as follows. Chapter 1 presents the essay "Modeling the Effects of Decision Complexity on Choice Behavior under Risk." Chapter 2 discusses the essay "An Econometric Analysis of Intertemporal Choice Heuristics." The final section provides concluding remark.

Chapter 1
Modeling the Effects of Decision Complexity on Choice Behavior under Risk

Abstract. This paper is concerned with how the effect of task complexity on behavioral error is better specified. Four behavioral error models are considered, namely the Fechner errors, the Contextual Utility, the Decision Field Theory, and the Entropy model. The latter three formulations accommodate the complexity effect by specifying the disturbance's standard deviation as a function of ad hoc measures of task complexity. Our results show that the estimation of risk parameters under EUT is robust across alternative stochastic models, whereas the risk parameters under RDU show some sensitivities. Our in- and out-of-sample analyses emphasize the need to allow behavioral error to vary with task complexity by rejecting homoskedastic Fechner in favor of two of the three heteroskedastic models. The Entropy model delivers the best fit under the Expected Utility Theory, and Decision Field Theory provides the best fit to the data under the Rank-Dependent Utility. The superior performance of a stochastic model can be explained by its ability to capture a more heterogeneous behavior in terms of utility curvature or decision weight, as it leads to higher utility differences or lower behavioral errors, or a combination of the two.

1. Introduction

It is well known that choice behavior exhibits substantial inconsistency, which is commonly known as behavioral errors. Without properly accounting for behavioral errors, one may derive false inferences from tests of alternative theories. The earliest stochastic specifications commonly model the randomness in choice behavior as linear additive and homoskedastic disturbances to a discrete choice model's latent dependent variable. However, several empirical studies find that the tendency to err when choosing between risky alternatives varies across decision tasks, thus confronting the homoskedasticity assumption. Alternative models of heteroskedasticity have been proposed in response, although only a few explicitly consider the rationale for the randomness.

Given its fundamental effects on many contexts of decision making, task complexity becomes an ideal candidate to explain the variation in behavioral errors. In the context of decision making under risk, Agranov and Ortoleva [2017] and Dave et al. [2010], for example, found that decision-makers express a more noisy behavior when confronted with more complex tasks. Early research then seeks to accommodate task complexity's impact on a behavioral error by specifying the disturbance's standard deviation as a function of *ad hoc* measures of task complexity. Bruhin et al. [2010], for instance, multiplies the standard deviation of the payoff range. Sonsino et al. [2002], on the other hand, specify the behavioral error as a function of the number of payoffs. These early extensions have been superseded by models that introduce heteroskedasticity on more rigorous behavioral foundations, where two such examples are Contextual Utility (CU)(Wilcox, [2011]) and Decision Field Theory (DFT)(Hey, Lotito and Maffioletti [2010]).[1]

[1] The Contextual Utility is motivated by evidence from studies in psychology that people find it more difficult to distinguish one option from another when the range of experimental stimuli is greater, and assumes that the standard deviation of behavioral errors is proportional to the maximum utility outcome difference within a choice set. The Decision Field Theory is based on the namesake theory in cognitive science, due to Busemeyer and Townsend [1993], and assumes that the standard deviation of behavioral errors is proportional to that of utility outcome differences between risky alternatives.

We introduce the Entropy (EN) model by Swait and Adamowicz [2001] as an alternative heteroskedastic specification. The authors considered market complexity in discrete choice models of consumer demand for products, where all attributes are known to the decision-maker. We adapt their model to binary choice tasks over lotteries, which is the most common experimental method to elicit risk preferences.

The EN model associates task complexity with task similarity and measures task complexity by Shannon's entropy of predicted choice probabilities (Shannon [1948]).[2] Under EN, a more complex decision task involves less disparity in choice probabilities between the two options. Accordingly, the task is deemed as less complex by the decision-maker when she has a more apparent preference towards a particular option.

EN is more flexible than CU and DFT in two ways. First, EN allows one to incorporate the decision-maker's preferences into the evaluation of which task is more complex than another. Under CU and DFT, the ranking of alternative decision tasks in terms of complexity is invariant to the decision-maker's risk preferences. Second, EN captures a more flexible association between the standard deviation of behavioral errors and task complexity. Several hypotheses concerning the association between the standard deviation and task complexity can be evaluated. Apart from the monotonic association postulated by CU and DFT, EN is also able to capture various types of non-monotone associations.[3] Swait and Adamowicz [2001], for example, consider the hypothesis that the standard deviation of behavioral errors increases from low to moderate levels of complexity,

---

[2] Similarly, Agranov and Ortoleva [2017] define a similar choice as the one that is not obviously better than the other. Buschena and Zilberman [2000] measure the choice similarity by the difference between cumulative densities associated with risky alternatives. In psychology, the similarity of two stimuli is usually represented by a monotonically decreasing function of psychological distance between those stimuli (Luce [1961] and Nosofsky [1984, 1990]).

Several studies discovered the effect of choice similarity to choice behavior. Bhatia and Mullett [2018] and Dashiell [1937], for example, found that the decision makers use longer time to choose between similar options. Rubinstein [1998] and Leland [1994] developed a decision heuristic based on the similarity on payoff and probability attributes to explain the Allais paradox.

[3] Wilcox [1993] observed that such association is not always monotone.

but decreases at high levels of complexity as the decision-maker becomes more careful in the evaluation of possible choices. Finally, unlike CU and DFT, the EN model includes the Fechner specification as a special case.

Our aim is to find out how the effect of task complexity on behavioral error is better specified. We consider several evaluation methods. We first compare the sensitivity of the inferred risk attitudes under Expected Utility Theory (EUT) and Rank-Dependent Utility (RDU) to which the stochastic model is used. It is important because we ultimately want to know if our choice of behavioral error model produces more reliable estimates of the risk parameters. We use maximum simulated likelihood (MSL) to estimate the full statistical model that captures unobserved heterogeneity in risk preferences of the underlying population by modeling all structural parameters as random coefficients that follow a population distribution. The estimation procedure is preferable due to its ability to guard against potential bias in the representative agent estimation (see, for example, Wilcox [2006]).

The simulation-based method is also highly flexible in the sense that it enables us to locate subject-specific risk parameters in the estimated population distribution by examining the subject's past choices. Stated explicitly, we derive the posterior distribution of each individual's risk parameters conditional on the subject's observed choices and distribution of risk parameters of the entire population.[4] Then, as our second objective, we compare the posterior distribution to the estimated population distribution to examine how large is the proportion of the total variation in the

[4] Our approach is similar to Moffatt [2005] and von Gaudecker et al. [2011]. An obvious alternative to infer individual-specific parameters is to directly compute a set of maximum likelihood estimation for each individual in the sample (see, for example, Monroe [2020]). The procedure is, however, onerous as it is only feasible when we have large choice observation for every individual. Another potential issue is that the estimation may fail to converge for some individuals, which force the researcher to exclude the individuals from the analysis (see, for example, Harrison and Ng [2016]).

estimated risk attitude parameter across decision makers can be captured by the variation in the individual-level parameter.

Third, we compare the performance of the stochastic specifications in approximating the true data-generating process of choices under risk, and benchmarking it to the homoskedastic Fechner. We consider log-likelihood, the Akaike Information Criterion (AIC; Akaike [1973]), and the Bayesian Information Criterion (BIC; Schwarz [1978]) values to evaluate the models' statistical goodness-of-fit. To alleviate the potential overfitting issue, we also evaluate the models' out-of-sample forecasting performance error models on the second experiment data.

Fourth, we approximate the representativeness of each error model to decision-makers stochastic behaviors in the population, derived from the log-likelihood contribution by each subject to the "grand log-likelihood" of the MSL estimation. Precisely, we calculate the percentage of subjects for whom an error model's log-likelihood value is the highest at the individual level.[5]

We are also interested in gaining insight into the factors that contribute to a better performance of a stochastic model. To do it, we use the inferred individual-level parameters from our hierarchical estimation to calculate the utility difference on each choice made by each individual.[6] We then compare the distribution of the utility differences for each error model to see how heavy the distribution around the indifference point as well as at each tail. We also compare the shape and magnitude of the standard deviation of behavioral errors at the points close to and far from indifference, as its combination with the utility difference affects the magnitude of log-likelihood in each choice observation by each individual.

---

[5] Although generating a similar output, our approach is different to the Finite Mixture Model (see, for example, Bruhin et al. [2010], Conte et al. [2011], and Harrison and Rutström [2009]), which jointly estimate two or more specifications to obtain the discrete distribution for each specification in the population.

[6] In a similar vein, Moffat [2005] uses the individual-level parameter derived from the population distribution to examine the relationship between decision effort and task complexity, with one of the considered complexity measures is the utility difference between two options.

We draw three main conclusions from our statistical analyses. First, we find that inferred risk attitudes under EUT are not sensitive to which stochastic model is used. This is, however, not the case for RDU, under which FN, DFT and EN lead to similar levels of risk aversion in terms of the utility function, but CU leads to a lower level. Risk aversion under RDU also depends on the shape of the probability weighting function. We find that the estimates of all models display a regular S-shape that induces risk aversion when the best outcome occurs with a small probability and risk-seeking when it occurs with a large probability.

Second, the in- and out-of-sample evaluations favor at least two heteroskedastic models over the homoskedastic FN. EN emerges as the best-fitting model under EUT, followed by CU. Meanwhile, DFT is outperformed by FN. Despite its unsatisfactory result under EUT, DFT is the best-fitting model under RDU. CU and EN remain the better stochastic models than FN under the theory. The findings imply that simply allowing for heteroskedasticity with respect to task complexity does not necessarily improve the model's goodness of fit; the type of heteroskedasticity allowed matters.

Third, the best-fit performance of a stochastic model can be explained by its ability to capture a more diverse behavior in terms of utility curvature and decision weight. When the underlying theory is EUT, we observe that EN is better at capturing the substantial mass of highly risk averse and risk seeking decision makers. This leads to utility differences which are more widely spread out from zero. With a larger proportion of high EU differences in its absolute value, the probability of choosing particular option increases. Because choices are more likely to be consistent when the decision maker has clearer preferences, a higher choice probability potentially leads to a higher log-likelihood value.

Meanwhile, despite a comparable distribution of expected utility differences, the weaker performance DFT under EUT is contributed by its higher standard deviation of behavioral errors.

As we account for the non-linear probability weighting function, the distribution of RDU differences in DFT now has the highest variance of rank-dependent utility differences that contributes to its much-improved goodness-of-fit. In addition, the higher population variance of the parameter that controls the curvature of the probability weighting function lowers the magnitude of DFT's standard deviation of behavioral errors. Specifically, its standard deviation of behavioral errors is now very low and approaches zero as we move away from the indifference points. Unless the decision maker is indifferent between two options, the resulting index of utility difference is then higher, leading to a higher choice probability. As the choices at points far from indifference are less noisy, the higher choice probability generates a higher log-likelihood value and better fitness to the data.

2. Data

We use existing binary choice data from a risk preference experiment that was part of a broader artefactual field experiment documented in Andersen, Harrison, Lau, and Rutström (AHLR) [2014] and Harrison, Lau, and Yoo [2020]. An initial sample of 50,000 adult Danes was randomly drawn from the population aged between 18 and 75. The sample was then stratified by geographic area to send out 1,996 invitations. Lastly, the experiment obtained a final sample of 413 subjects. The descriptive statistics show that the sample is representative of the Danish adult population. Their average age was 48.7, 48.2% were female, and 56.5% were married. As a comparison, the initial sample of 50,000 Danes had an average age of 49.8, 50.7% were female, and 50.1% married.

Each subject was confronted with the four blocks of risk aversion tasks. Each block comprised ten binary choices, in which the subjects had to make a choice from each pair of lottery presented. Each block differs in lottery prize combination, and each lottery has two possible prizes.

The prize sets employed are: [A1: 2000 and 1600; B1: 3850 and 100], [A2: 1125 and 750; B2: 2000 and 250], [A3: 1000 and 875; B3: 2000 and 75] and [A4: 2250 and 1000; B4: 4500 and 50], with randomized order for each subject. The ten tasks in each block were presented one at a time in an ordered manner. At the first row of each block, each lottery offered a 10% chance of receiving the high prize and a 90% chance of receiving the low prize. For example, the subject picks lottery A that gives the subject a 10:90 chance of receiving 2000 kroner or 1600 kroner or lottery B that has a 10:90 chance of receiving 3850 kroner or 100 kroner. As subjects proceed down the block, the probability of receiving the high prize increases by 10% until the last row had the subject choose between two degenerated lotteries. Since the last choice in each block was provided as a stochastic dominance test and does not directly contribute to the identification of risk attitudes, we exclude it from our analysis and focus on modeling the remaining 36 choices per subject.[7] Under this experimental design, only risk seeking subjects would choose lottery B in the first row, and only highly risk averse subjects would pick lottery A in the ninth row.

One of the subject's 40 choices was randomly selected and played out at the end of the experiment, and the subject had a 10% chance of realizing the lottery prize. The average payment for the risk preference experiment was 242 kroner, equivalent to 48 U.S. dollars using the exchange

[7] As we will see in the next section, unlike other behavior error models considered in this study, violation of first-order stochastic dominance (FOSD) is intuitively impossible under the Decision Field Theory (DFT). DFT predicts zero error on a choice of degenerated lotteries; thus, decision makers will never choose dominated lotteries. This advantage comes with a numerical problem in the estimation. Suppose some decision-makers choose dominated lottery in the FOSD pair. In that case, DFT could not calculate choice probabilities for that pair (because the expected utility difference between two lotteries should be divided by zero standard error). Indeed, we found 115 transparent FOSD violations by 48 decision makers in the data, which prevented us to estimates risk parameters with DFT even if we used a representative agent model with maximum likelihood estimation. It is possible to computationally fix the problem by adding the tremble error (Harless and Camerer [1994]) alongside the "main" error stories (see, for example, Blavatskyy and Pogrebna [2009], Conte et al. [2011], Hey et al. [2010] and Loomes et al. [2002]). However, as we aim to see the pure performance of those "main" error stories and provided that the pair is merely a test of monotonicity, we exclude the choice of degenerated lotteries from our analysis.

rate at the time of the experiment. Additionally, each subject received a fixed show-up fee of 300 kroner or 500 kroner.

Six months later, the experiment was repeated with 182 of the 413 subjects who participated in the first experiment. The experiment was conducted in private for each subject and carried out at the same hotel as the first experiment or other convenient location for them, such as their private residence. Such a costly implementation precluded the experimenters from using the same payment structure as in the first experiment. Constrained by the budget, all subjects in the second experiment were only paid a fixed participation fee of 300 kroner.[8]

3. Econometrics

We first write out a simple structural model of individual risk attitudes by combining Expected Utility Theory (EUT) with a Fechner error specification, which assumes that behavioral errors are homoskedastic with respect to the complexity of decision tasks. We then extend the model to account for unobserved heterogeneity in individual risk attitudes as well as three forms of heteroskedasticity that incorporate different notions of task complexity. In particular, we consider Contextual Utility and Decision Field Theory specifications, which assume that the standard deviation of behavioral errors increases in task complexity, and a flexible Entropy specification that allows one to evaluate the effect of task complexity on the standard deviation as an empirical question. Finally, we generalize EUT to the Rank-Dependent Utility (RDU) model that incorporates probability weighting in the evaluation of lotteries.

[8] Despite the hypothetical payoffs, using the data from the second experiment to evaluate the stochastic models' out-of-sample forecasting performance would not be a significant issue since Harrison, Lau and Yoo [2020] found evidence of stable risk preference in the longitudinal data used in this study.

A. Expected Utility Theory with Fechner Error Term

Consider first the estimation of individual risk preferences under EUT, which is one of the most popular models of decision making under risk. Let M_{ij} be the j^{th} prize of lottery $i \in \{A, B\}$ and $j \in \{1, 2\}$. Assume that utility is given by the constant relative risk aversion (CRRA) function

$$U(M_{ij}) = M_{ij}^{(1-r)}/(1-r) \quad (1)$$

where r is the CRRA coefficient, and $r \neq 1$. Under EUT, a positive r denotes risk aversion, a negative r denotes risk seeking behavior, and $r = 0$ represents risk neutral behavior.

Lotteries are evaluated by their expected utility, and the choice between lottery A and B depends on the difference in EU between the two lotteries. Let $p(M_{ij})$ be the probability of outcome M_{ij}, which is induced by the experimenter. The EU of lottery i is simply the weighted average of the utility of each outcome

$$EU_i = p(M_{i1}) \times U(M_{i1}) + p(M_{i2}) \times U(M_{i2}), \quad (2)$$

where $p(M_{i2}) = 1 - p(M_{i1})$ in the decision tasks. Let y denote a binary indicator of the subject's choice between lottery A ($y = 1$) or lottery B ($y = 0$), and let $\mathbf{I}[.]$ denote an indicator function. The observed choice under EUT can then be written as $y = \mathbf{I}[(EU_A - EU_B) > 0]$.

Next we combine the EUT model with a stochastic behavioral error term that allows observed choices to deviate from deterministic theoretical predictions. Consider the Fechner (FN) error specification, popularized by Hey and Orme [1994], which assumes that the error term has constant variance across all decision tasks regardless of their complexity.[9] The observed choice now depends on the difference in EU as well as the random error term $\boldsymbol{\varepsilon}$, such that $y = \mathbf{I}[(EU_A - EU_B) +$

[9] One may introduce heteroskedasticity into the Fechner specification by specifying μ as a function of observable characteristics of decision-makers or decision tasks. For example, Lau, Yoo and Zhao [2019][2021] estimated individual risk attitudes under Cumulative Prospect Theory (CPT) with a Fechner specification and allowed μ to vary with the lottery characteristics. Another example is Galarza [2009], who considered μ to be a linear function of the individual's demographic characteristics.

$\varepsilon > 0$]. Assume that ε is normally distributed with mean 0 and standard deviation μ, $\varepsilon \sim N(0, \mu^2)$. The likelihood of each choice can be specified as

$$P(r, \mu) = \Phi(\nabla EU^{FN})^y \times (1 - \Phi(\nabla EU^{FN}))^{(1-y)} \quad (3)$$

where $\Phi(\nabla EU^{FN})$ is the standard normal CDF evaluated at the index value ∇EU^{FN} given by

$$\nabla EU^{FN} = (EU_A - EU_B)/\mu. \quad (4)$$

As the noise parameter μ approaches 0, this specification converges to a deterministic EUT model with no behavioral errors. Conversely, when μ gets sufficiently large, the specification converges to a random choice model driven entirely by noise, with both lotteries having a 50:50 chance of being selected regardless of the underlying difference in EU.

We index the CRRA coefficient r_n by subject n to specify it as an individual-specific random coefficient drawn from a population distribution of risk preferences to account for unobserved preference heterogeneity across individuals. We denote the density function for the random CRRA coefficient as $f(r_n; \theta)$, where θ is a set of parameters that characterize the distribution.[10] The choice-level likelihood function in (3) is written as $P_{nt}(r_n, \mu)$ to represent subject n's choice in decision task t.

Direct estimation of the set of parameters θ is possible once the density function $f(r_n; \theta)$ is fully specified. Assume that r_n is normally distributed so that $\theta = (\bar{r}, \sigma_r)$, where $\bar{r}$ and σ_r are the population mean and standard deviation of the CRRA coefficient r_n, respectively. Conditional on the CRRA coefficient draws, the following equation specifies the conditional likelihood of observing a series of choices made by subject n

$$CL_n(r_n, \mu) = \prod_t P_{nt}(r_n, \mu). \quad (5)$$

[10] We assume the standard deviation of the behavioral error to be fixed for every subject. This modeling choice is preferred as we want to isolate the effects of risk preferences on the different measures of task complexity which may explain the error model's relative performance. Additionally, we intend to compare to evaluate the heteroskedastic error models relative to the homoskedastic Fechner specification.

Since we model r_n as random coefficients, the "unconditional" (Train [2009, p.146]) likelihood of subject n's choices is derived by taking the expected value of $CL_n(r_n, \mu)$ over the distribution $f(r_n; \theta)$

$$L_n(\bar{r}, \sigma_r, \mu) = L_n(\theta, \mu) = \int CL_n(r_n, \mu) f(r_n; \theta) dr_n. \quad (6)$$

Many textbook models for panel data, such as random effects probit (Wooldridge [2010, p.613]), similarly integrate out unobserved heterogeneity as in equation (6). A distinguishing feature in our application is that unobserved heterogeneity enters the index of expected utility difference non-linearly via the CRRA coefficient. It is not possible to solve the integral analytically, but we can compute it by simulation methods (Train [2009, p.144-145]). In particular, we draw a value of r_n from population distribution $f(r_n; \theta)$ and calculate the conditional likelihood $CL_n(r_n, \mu)$ with this draw. We repeat the process using 100 Halton draws and the results are then averaged over the draws. We use maximum simulated likelihood to estimate the distribution of preference parameters θ and the behavioral error parameter μ, where the maximand is a simulated analogue to the sample log-likelihood function $\sum_n \ln(L_n(\theta, \mu))$ across the 413 subjects who participated in the experiment.

B. Contextual Utility Specification

The simplifying assumption of homoskedasticity across decision tasks that potentially vary in complexity has been challenged by empirical evidence (Alós-Ferrer and Garagnani [2021], Hey [1995][2001], and von Gaudecker et al. [2011]). We now turn to alternative models of behavioral errors that address this drawback by augmenting the Fechner error specification with a heteroskedastic structure. Specifically, the models add a proxy measure of task complexity as a multiplicative factor to the standard deviation of the behavioral error term.

The Contextual Utility (CU) model of Wilcox [2011] extends the homoskedastic Fechner error specification by adding a scalar that normalizes the difference in expected utility between the two lotteries. Let $\upsilon \times \varepsilon$ denote the resulting error term, where $\varepsilon \sim N(0, \mu^2)$ as before and υ is a

positive scalar that captures the effect of heteroskedasticity from the variation in complexity across the decision tasks. Under CU, the measure of task complexity υ is given by $(U_{max} - U_{min})$, where U_{max} and U_{min} are the maximum and minimum of the four potential outcome utilities: $U(M_{A1})$, $U(M_{A2})$, $U(M_{B1})$ and $U(M_{B2})$. The likelihood of each choice observation can then be obtained by replacing the index ∇EU^{FN} in (4) with

$$\nabla EU^{CU} = (EU_A - EU_B)/ (\upsilon \times \mu) \qquad (7)$$
$$= (EU_A - EU_B)/ [(U_{max} - U_{min}) \times \mu].$$

Note that replacing the Fechner specification with CU does not require estimation of an extra parameter as $(U_{max} - U_{min})$ depends only on the potential outcomes and the existing r parameter. MSL estimation of $\theta = (\bar{r}, \sigma_r)$ and μ can proceed in the same way once the joint likelihood function in (6) has been modified to incorporate the new index function in (7).

With $\upsilon = U_{max} - U_{min}$, CU only requires the best and the worst outcomes in a lottery pair to measure task complexity, and the degree of task complexity does not change with alternative configurations of intermediate outcomes. With the design of a lottery pair where $M_{B1} > M_{A1} > M_{A2} > M_{B2}$, the postulated task complexity of CU does not vary with the characteristics of lottery A. The degree of task complexity under CU is also invariant to outcome probabilities. CU considers a task with $p(M_{A1}) = p(M_{B1}) = 0.5$ as complex as tasks with $p(M_{A1}) = p(M_{B1}) = 0.9$, as long as M_{B1} and M_{B2} are similar for both tasks. Thus, the definition of task complexity under CU is rather restrictive by only looking at minimum and maximum prizes in paired lotteries.

C. Decision Field Theory

Based on Decision Field Theory (DFT) by Busemeyer and Townsend [1993], Hey, Lotito and Maffioletti [2010] and Wilcox [2015] consider alternative specifications of multiplicative heteroskedasticity where task complexity is measured by the variance of the difference in utility of

outcomes between two lotteries. This variance changes across lottery pairs and depends on the dispersion of outcomes as well as the probabilities of the outcomes. Task complexity under DFT thus includes a more complete set of lottery characteristics than CU.

Recall from Section 2 that the probability of the best outcome is the same for both lotteries in each decision task, but the prizes in the two lotteries are different: that is, $p(M_{A1}) = p(M_{B1})$ and $M_{A1} \neq M_{B1}$. The difference in utility of outcomes between the two lotteries can therefore be modeled as a Bernoulli random variable, where $U(M_{A1}) - U(M_{B1})$ occurs with probability $p(M_{A1}) = p(M_{B1})$, and $U(M_{A2}) - U(M_{B2})$ with probability $1 - p(M_{A1}) = 1 - p(M_{B1})$. Let σ_{A-B}^2 denote the variance of this Bernoulli random variable, we then have[11]

$$\sigma_{A-B}^2 = [p(M_{A1}) \times (1 - p(M_{A1}))] \times [U(M_{A1}) - U(M_{B1}) + U(M_{B2}) - U(M_{A2})]^2. \quad (8)$$

Under DFT, the heteroskedastic factor υ is set to σ_{A-B}, and the index ∇EU^{FN} in (4) is now replaced with

$$\nabla EU^{DFT} = (EU_A - EU_B) / (\sigma_{A-B} \times \mu). \quad (9)$$

Replacing the Fechner specification with DFT does not require estimation of an extra parameter: σ_{A-B} is a function of the risk aversion parameter r, prizes and probabilities in the lotteries. For MSL estimation of $\theta = (\bar{r}, \sigma_r)$ and μ, we only need to modify the joint likelihood function in (6) to incorporate ∇EU^{DFT} in lieu of ∇EU^{FN}.

The measure of σ_{A-B} in equation (8) implies that the degree of task complexity in DFT is determined by two aspects. First, the probability of the best and worse outcomes. It is represented by the product of the two probabilities in the first bracket of equation (8). The standard deviation of

[11] In the original specification of DFT (Busemeyer and Townsend [1993, p.438]), the variance of utility differences between two lotteries is given by $\sigma_{A-B}^2 = \sigma_A^2 + \sigma_B^2 - 2\sigma_{AB}$, where σ_A^2 and σ_B^2 are the variances of the utility of each monetary outcome in lottery A and lottery B, respectively. The last term σ_{AB} represents the covariance between the utilities in each lottery. The variance of the utility of outcomes in lottery $i \in \{A, B\}$ is given by $\sigma_i^2 = \sum_j p(M_{ij}) \times [U(M_{ij}) - EU_i]^2$, and the covariance is given by $\sigma_{AB} = E[(U(M_{Aj}) - EU_A) \times (U(M_{Bj}) - EU_B)]$. In our case, $p(M_{Aj}) = p(M_{Bj})$, which implies that $\sigma_{AB} = \sum_j p(M_{Aj}) \times [U(M_{Aj}) - EU_A] \times [U(M_{Bj}) - EU_B]$, and σ_{A-B}^2 can then be simplified to equation (8).

a Bernoulli random variable reaches the maximum when the best and worst outcomes are equally likely, and declines towards a minimum of zero as either outcome probability moves to 1.[12] Hence, a task with $p(M_{A1}) = p(M_{B1}) = 0.5$ is considered more complex under DFT than a task with $p(M_{A1}) = p(M_{B1}) = 0.9$, given the same set of outcomes $[M_{A1}, M_{A2}, M_{B1}, M_{B2}]$ for both tasks. Second, the dispersion of outcomes in the lottery pair, represented by the second bracket in equation (8). It controls the magnitude of the standard deviation of behavioral errors and changes with the outcome combination $[M_{A1}, M_{A2}, M_{B1}, M_{B2}]$. Any changes in the set of outcomes that increase $U(M_{A1}) - U(M_{B1})$ and $U(M_{B2}) - U(M_{A2})$ will also translate into an increase in σ_{A-B}. Unlike CU, task complexity under DFT also depends on the intermediate outcomes M_{A1} and M_{A2} in addition to the best and worst outcomes M_{B1} and M_{B2}.

Despite including more lottery information than CU in the measure of task complexity, DFT relies on the assumption that the standard deviation of behavioral errors is proportional to the measure of task complexity. This assumption does not take into account any possible non-monotonic response to increased task complexity. For example, it is possible that behavioral errors increase with task complexity up to some degree and then decrease due to more careful examination by the decision maker. Such varying efforts exerted by the decision maker in response to changes in task complexity may lead to a non-monotonic association between the standard deviation of

[12] Given the range of possible values of σ_{B-A}, violation of first-order stochastic dominance (FOSD) is intuitively impossible under DFT. When σ_{B-A} is equal to zero, DFT predicts that decision makers never make an error, thus never choosing dominated lotteries in a FOSD pair. With RDU specification, we generalize the function σ_{B-A} by allowing decision weights to enter into the complexity measure, as in Wilcox [2015]. On the other hand, Hey et al. [2010] use the objective probability of outcomes to calculate σ_{B-A} even when the analysis is under the cumulative prospect theory. When accounting for the unobserved heterogeneity among decision makers, it is possible to find some of them with extreme estimated weighting parameters. In that case, the decision makers may extremely underweight probability of 0.1 as it is equal to zero and overweight probability of 0.9 as it is equal to unity, which then leads to σ_{B-A} equals zero, although when the lottery pair is not FOSD. To avoid any numerical problems, we assume that σ_{B-A} is in the interval of $(0, \infty)$ when the underlying decision theory is RDU.

behavioral errors and task complexity. We now turn to a more general error specification that can accommodate this potential scenario.

D. Entropy Model

The Entropy (EN) specification of Swait and Adamowicz (SW) [2001] relies on the intuitive notion that a decision task is less complex when the decision maker has a clearer preference for a particular option. It follows that the most complex task is one where the decision maker is indifferent between all available options. SW measure the decision maker's preference for an option by the probability of choosing that option. Information on the probability of choosing each option is then used to measure task complexity by employing Shannon's entropy (Shannon [1948]) of choice probabilities. In our setup, the entropy can be specified as

$$H = -1 \times [\Phi(\nabla EU^{FN}) \times \ln(\Phi(\nabla EU^{FN})) + (1 - \Phi(\nabla EU^{FN})) \times \ln(1 - \Phi(\nabla EU^{FN}))], \qquad (10)$$

where $\Phi(\nabla EU^{FN})$ and $1 - \Phi(\nabla EU^{FN})$ are the probabilities of choosing lottery A and lottery B, respectively, under the Fechner error specification. The relationship between Shannon's measure of entropy and the choice probabilities takes an inverse-U shape where H reaches a maximum value of 0.693 when the decision maker is indifferent between the two lotteries, *i.e.* when $\Phi(\nabla EU^{FN}) = 0.5$. Entropy H is a symmetric function of the choice probability $\Phi(\nabla EU^{FN})$ and it converges to 0 when the choice probability approaches zero or unity. Recall that the complexity measure in DFT also takes a similar shape, which is symmetric around its maximum. However, the measure differs from EN, as the entropy function H is based on the decision maker's probability of choosing each lottery. Whereas, DFT depends on the objective probabilities of obtaining the lottery's prizes.

The incorporation of the decision maker's preferences into complexity measurement H is one feature that distinguishes EN from CU and DFT. Under the latter two, the ranking of task complexity is invariant to the decision maker's risk preference parameter r, although the cardinal

measure of task complexity may change with the CRRA coefficient. The following example illustrates the differences between the three error specifications.

Consider three decision makers with different risk attitudes facing three decision tasks. The first decision maker is risk seeking, whose CRRA coefficient is −1.8. The second decision maker is slightly risk averse with a CRRA coefficient of 0.20, and the third decision maker is assumed to have a CRRA coefficient of 1.50. The three decision tasks, denoted by T1, T2, and T3, have the same set of lottery prizes $[M_{A1}, M_{A2}; M_{B1}, M_{B2}] = [1000, 875; 2000, 75]$, but differ in the probability of obtaining the highest prize in each lottery. In task T1, the probability of obtaining the highest prize in each lottery is 0.1, while they are 0.5 and 0.9, respectively, in T2 and T3.[13] As before, $p(M_{A1}) = p(M_{B1})$ and, accordingly, $p(M_{A2}) = p(M_{B2}) = 1 - p(M_{A1}) = 1 - p(M_{B1})$.

According to CU, the three tasks are equally complex for each decision maker because the difference between the highest and lowest utility is the same for each task, given by U(3850) − U(100). Under DFT, as T2 offers a 50:50 probability of receiving either prize in each lottery, the three decision-makers unanimously agree that T2 is the most complex task. The other two tasks are equally complex. The different degrees of risk aversion do not change the ordinal complexity ranking but the cardinal level of task complexity.[14] This applies to both CU and DFT. For example, despite agreeing that T2 is the most complex task under DFT, the more risk averse decision maker would give a higher complexity score on the task than what would be given by the decision maker with a slightly risk averse behavior.

[13] Recall that in the experiment design, T1 and T3 are located at the first and second-to-last row of the choice list, while T2 is located at the middle of the list.

[14] As illustrated in Figure A1 in Appendix A, CU or DFT maintains the same complexity ranking of the decision tasks for any given level of risk aversion. However, the magnitude of the task complexity changes proportionally with the CRRA coefficient. Under EN, the ranking of task complexity varies with parameter r.

The ordinal complexity ranking in EN, however, depends on the degree of risk aversion. The risk seeking decision maker will be indifferent between the two lotteries in T1 and will have a lower probability of choosing lottery A in T2 and T3. As measured by H, the task complexity is higher for T1 than T2 and T3. The slightly risk averse decision maker will have a 50:50 probability to choose either lottery in T2, but less probable to choose lottery B in T1 and most likely to choose lottery B in T3. For this type of decision maker, T2 is the most complex task. Finally, the more risk averse decision maker considers T3 as the most complex task, as she will be indifferent between the two lotteries in T3 and have a higher probability of choosing lottery A in T2 and is almost certain to pick lottery A in T1.

Another distinct feature of the entropy model is how the measure of task complexity enters the heteroskedastic specification. Behavioral errors are still specified as $\upsilon \times \varepsilon$, where $\varepsilon \sim N(0, \mu^2)$ and υ is a multiplicative factor of heteroskedasticity. However, in contrast to CU and DFT, the EN specification allows for the possibility that υ, and hence the standard deviation of behavioral errors, varies non-monotonically with task complexity. This flexibility is motivated by the idea that the net effect of increased complexity on the standard deviation should be evaluated empirically since the decision maker may respond to increased complexity by more carefully evaluating the available options. Following SW, we specify $\ln(\upsilon)$ as a quadratic function such that $\upsilon = \exp(\beta_1 \times H + \beta_2 \times H^2)$, where β_1 and β_2 are estimated parameters. The index ∇EU^{FN} in (4) is then replaced by

$$\nabla EU^{EN} = (EU_A - EU_B)/[\exp(\beta_1 \times H + \beta_2 \times H^2) \times \mu]. \quad (11)$$

Note that the evaluation of complexity H in (9) is still based on the Fechner index ∇EU^{FN} as shown in (4). One can use the joint likelihood function in (6) a basis for the MSL estimation, bearing in mind that replacing ∇EU^{FN} with ∇EU^{EN} in (4) implies that the joint likelihood function depends on β_1 and β_2, as well as the risk preference parameters $\theta = (\bar{r}, \sigma_r)$ and the baseline noise parameter μ.

The association between the standard deviation of behavioral errors and task complexity can take any shape. SW hypothesize that the association takes an inverse-U shape, that is, $\beta_1 > 0$ and $\beta_2 < 0$. It implies that at moderate levels of complexity, an increase in complexity increases the chance of making large behavioral errors (greater standard deviation). Yet, as the complexity levels continue to increase, the decision maker pays more attention to the decision tasks and deliberate her choices more carefully; thus, decreasing the behavioral errors.

The EN specification also allows us to evaluate other hypotheses concerning the association between the standard deviation of behavioral errors and task complexity. The Fechner specification is nested within the EN specification as a special case of $\beta_1 = \beta_2 = 0$, a joint hypothesis that can be easily tested. DFT and CU do not have this flexibility since they do not nest the Fechner specification. When $\beta_1 \geq 0$ and $\beta_2 \geq 0$ and at least one of the inequalities holds strictly, the standard deviation increases in task complexity as postulated under CU and DFT, albeit the measure of task complexity is different.

E. Rank-Dependent Utility

The Rank-Dependent Utility (RDU) is a popular generalization of EUT proposed by Quiggin [1982] that allows the non-linear transformation of the objective probabilities of obtaining lottery prizes. Recall that each lottery in our data has only two monetary outcomes, with $M_{i1} > M_{i2}$. Denote w(p) as the probability weighting function, the rank dependent expected utility of lottery i can then be specified as

$$RDU_i = [\, w(p(M_{i1})) \times U(M_{i1}) \,] + [\, (1 - w(p(M_{i1}))) \times U(M_{i2}) \,]. \quad (12)$$

Provided its flexibility, we consider a probability weighting function by Prelec [1998] which is given by

$$w(p) = \exp\{-\eta(-\ln p)^{\varphi}\}, \quad (13)$$

and is defined for $0 < p < 1$, $\eta > 0$ and $\varphi > 0$. We model φ and η as log-normally distributed random coefficients φ_n and η_n, respectively, that vary across individuals. This two-parameter function exhibits inverse-S probability weighting (overweighting of small probabilities and underweighting of large probabilities) for $\varphi < 1$, and S-shaped probability weighting (underweighting of small probabilities and overweighting of large probabilities) for $\varphi > 1$. When $\varphi = 1$, the function exhibits global concave probability weighting (overweighting of all probabilities) for $\eta < 1$, and globally convex probability weighting (underweighting of all probabilities) for $\eta > 1$. The function exhibits linear probability weighting that reduces RDU to EUT when $\varphi = \eta = 1$.

The logic behind our econometric specifications allows natural extension from EUT to RDU once we replace equation (2) with equation (12). The introduction of the probability weighting function in RDU implies that the risk premium consists of two components: the non-linear utility function governed by the parameter r and the probability weighting function governed by the parameters φ and η.

Accounting for the non-linear probability weighting function has no effect on the measurement of task complexity in CU. It still depends on the difference between the utility of the best and worst outcome, which in our experiment data is given by $U(M_{B1}) - U(M_{B2})$.

With RDU, the most complex task under DFT does not necessarily offer a 50:50 chance of obtaining the best or the worst outcome. As the non-linear probability function also enters the complexity measure in equation (8), it is now given by a task where $w[p(M_{A1})] = w[p(M_{B1})] = 0.5$. To see the pure effect of the probability weighting function on the task complexity measure in DFT, first, consider the Dual Theory of choice under risk by Yaari [1987] with linear utility and a power probability weighting function. It is equivalent to $r = 0$ and $\varphi = 1$, so η is the only parameter characterizing risk behavior. The upper panel of Figure A2 in Appendix A shows that, unlike CU, the location of maximum point of the task complexity changes with parameter η. Let us reconsider

the tasks T1, T2, and T3 discussed in the previous subsection. A risk seeking decision maker with $\eta = 0.3$ considers T1 the most complex task as she overweights the presented probability so that $w[0.1] = 0.5$. The most complex task for a very risk averse decision maker with $\eta = 6.5$ is T3, as she underweights the objective probability of 90% to 50%. For both decision makers, T2 is considered the second complex task.

To disentangle the effect of variation in φ to task complexity under DFT, we again assume risk neutrality and set η equal to 1. As illustrated in the lower panel of Figure A2, the complexity measure in DFT maintains the same highest value, but its symmetric shape becomes more dense around the maximum value as φ increases. The standard deviation of behavioral errors is thus lower in the tasks that offer the best prizes with a low or high probability, which potentially leads to higher ∇RDU indices and higher log-likelihood values. Conversely, when φ is very low, the change in the shape of the standard deviation of behavioral errors is less sensitive to the variation of the probability of receiving high prizes. The resulting ∇RDU indices and log-likelihood values are then potentially lower.

Under EN, the function H now depends on the choice probability $\Phi(\nabla\text{RDU})$. Hence, in addition to the utility curvature parameter, the probability weighting parameters also determine the order of task complexity. With the risk neutrality assumption, the inclusion of the non-linear probability weighting has a similar effect to the changes of task complexity as in DFT (see the top panel of Figure A2). By setting φ equals 1, the parameter η determines the location of the maximum point of complexity measure H in the choice list. For example, the risk seeking decision maker, whose parameter η is equal to 0.3, considers T1 the most complex task because she will have a clearer preference toward a particular option (*i.e.*, lottery B) in T2 and T3 than in T1. According to a risk averse decision maker whose $\eta = 6.5$, the tasks T1 and T2 are less complex than T3 as she will have a stronger preference toward an option (*i.e.*, lottery A) in T1 and T2 than in T3.

The bottom panel of Figure A2 shows that the symmetric shape of the complexity measure H also becomes tighter around its peak as φ increases while setting $r = 0$ and $\eta = 1$. However, the standard deviation of behavioral errors potentially differs from DFT because EN allows for a possible non-monotonic relationship between task complexity and behavioral errors by incorporating the task complexity measure into its heteroskedastic form through a quadratic function.

4. Results

Our first objective is to evaluate the behavioral error specifications presented in Section 3 when they are combined with Expected Utility Theory (EUT) and Rank-Dependent Utility (RDU). We first evaluate the sensitivity of inferred risk attitudes to alternative stochastic specifications, both based on the estimated population distribution and posterior distribution of individual-level parameters. We then compare how well the models describe the observed choices when they are estimated in the in-sample data provided by the first experiment. The longitudinal design of the experiment allows us to also compare their out-of-sample forecasting performance using the data from the second experiment. Specifically, we evaluate each error model's log-likelihood in the second experiment data at the parameter estimates obtained from the first experiment. We measure the models' goodness-of-fit in both evaluations by means of AIC, BIC, and Vuong test. Finally, we calculate the representativeness of the error models to decision makers' stochastic behaviors by comparing their log-likelihood values at the individual level.

Our second objective is to decipher the factors that contribute to the goodness-of-fit the heteroskedastic error models. Using the individual-specific risk parameters derived from the population distribution, we calculate the error rate of the choice prediction made by each stochastic model in the in-sample data. We use the mean absolute error (MAE) as the measure of the accuracy

of choice prediction, defined as the absolute difference between the predicted choice probability and the observed choice. We then calculate the EU and RDU differences in each choice observation for every individual evaluated at their individual-level risk parameters. We then compare the distributions of EU and RDU differences in each model. We also compare the shape of the standard deviation of behavioral errors in each model to discern how its magnitude changes near or far from the indifference point. If the choice prediction is correct, the combination of a high EU/RDU difference and low standard deviation of behavioral errors leads to a higher log-likelihood value. Accordingly, inaccurate choice prediction penalizes a model with a lower log-likelihood value. The empirical study generally found higher inconsistency when a choice in the decision task is perceived by the decision maker as not obviously better than the other (see, for example, Agranov and Ortoleva [2017]). Given the average risk preference and, accordingly, the indifference point, the pattern of MAE in the choice list may give us insight into which part of the choice list the prediction by a stochastic model is generally more accurate. As the choices far from the indifference point are generally more consistent, then the standard deviation of behavioral errors should be lower. The combination of large EU/RDU differences and lower standard deviation of behavioral errors at points far from indifference potentially generates a higher log-likelihood value, which, in turn, supports the statistical goodness-of-fit of a stochastic model. Conversely, the choices around the indifference point are usually noisier; thus, it is more likely for a stochastic model to make incorrect choice predictions. The higher noise is also better represented by a higher standard deviation of the error term. As the standard deviation gets arbitrarily large, the probability of choosing either option is close to 50:50. As a result, the stochastic model is better at minimizing the penalty on the log-likelihood due to many incorrect choice predictions.[15]

[15] For that reason, in a case of an almost deterministic model, represented by the homoskedastic Fechner specification with almost zero standard deviation of behavior errors, we may observe higher log-likelihood values for choices at points far from different. However, with noisier choices around the indifference point, any mistakes in predicting the choices will heavily penalize the model with lower

A. Inferred Risk Attitudes under Expected Utility Theory (EUT)

We start our analysis by comparing the sensitivity of the inferred risk behaviors to which the stochastic model is used. We first consider the population distribution of the risk parameter, and Table 1 reports the parameter estimates under each stochastic specification. In all stochastic models, we can reject the hypothesis that the population means of CRRA coefficient are equal to zero with p-values < 0.001, implying that, on average, the decision makers in the population exhibit a non-linear utility. As illustrated by Figure 1, the MSL estimation of EUT with every behavioral error specification yield an almost identical population mean of the r parameter. With FN, we find that the population distribution of the r parameter has an estimated mean of 0.535 and an estimated standard deviation of 0.608. The estimate of the population mean of the CRRA coefficient in CU is 0.531, while its standard deviation is 0.635. We also observe that the estimated mean of the population distribution of the r parameter in DFT is 0.569, while the estimated standard deviation is 0.689. Finally, the estimates of the population mean and standard deviation of the utility curvature parameter in EN are 0.547 and 0.801, respectively. As reported in Table B1 in Appendix B, we cannot reject the null hypothesis of the same population means of the r parameter in each pair of stochastic specifications with p-values of at least 0.410.

To compare the combined effects of the mean and standard deviation across different models, consider the implied percentage of decision makers who are risk averse. Under FN, the estimates suggest that risk averse decision makers make up 81.1% of the population. The three other models produce comparable figures: 80% under CU, 79.7% under DFT and 75.2% under EN.

log-likelihood values.

B. Individual-level Parameter under EUT

Using the estimated population distribution of risk parameters from the pooled data for all subjects, we now infer the location of each subject's risk parameters in the population distribution by examining the series of choices that the subject made. Recall from Section 3 that $f(r_n; \theta)$ is the population distribution of CRRA coefficients of the entire population. Denote the observed sequence of choices made by subject n as y_n. Now, let $g(r_n | y_n, \theta, \mu)$ be the distribution of individual-specific r_n values conditional on the subject's observed choices, the population distribution of risk parameters, and the standard deviation of the behavioral error term. By Bayes' rule, $g(r_n | y_n, \theta, \mu) = CL_n(r_n, \mu)f(r_n; \theta) / L_n(\theta, \mu)$. Recall from equation (5) and (6), the denominator of this ratio is the integral of the numerator. The utility curvature parameter specific for decision maker n is the expected value of that decision maker's conditional distribution of the r parameter given by $\tilde{r}_n$ $=E(r_n | y_n, \theta, \mu) = \int r_n \, g(r_n | y_n, \theta, \mu)dr_n = \int r_n \, CL_n(r_n, \mu)f(r_n; \theta)dr_n \, / \, \int CL_n(r_n, \mu)f(r_n; \theta)dr_n$.[16]

Table 2 reports the descriptive statistics of the individual-level parameter $\tilde{r}_n$ for the 413 subjects. In each error model, we observe that the sample average of the conditional means of the r parameter is similar to its estimated population mean $\bar{r}$ given in Table 1. FN, CU, and DFT generate means of individual-level parameter $\tilde{r}_n$ of 0.547, 0.539, and 0.574, respectively, while the mean of $\tilde{r}_n$ in EN is 0.557. Recall from Table 1, the estimated means of the utility curvature parameter for FN, CU, and DFT are 0.535, 0.531, and 0.569, respectively, and that for EN is 0.547.

The standard deviation of the individual-level parameter $\tilde{r}_n$, denoted as $\sigma_{\tilde{r}n}$, in FN is 0.547. Measured by the $\sigma_{\tilde{r}n}/\sigma_r$ ratio, the closeness of $\sigma_{\tilde{r}n}$ to the estimate of σ_r also indicates how large is the proportion of the total variation in the estimated risk aversion parameter across decision makers can be captured by the variation in the individual-level parameter $\tilde{r}_n$.[17] By comparing it to the estimated

[16] None of these integrals exist in closed form, but can be approximated using simulation methods (Train [2009, p. 263] and Greene [2020, p. 723]).

[17] The variance of individual-level parameters is generally smaller than the unconditional distribution of the parameter (Sarrias [2020]) because it does not account for the variance of the parameter around its

population standard deviation of the r parameter, the variation in the individual-level CRRA coefficient in FN captures 90% of the total estimated population variation in r_n. CU generates a standard deviation of $\tilde{r}_n$ of 0.586, capturing 92% of the estimated standard deviation of the r parameter in the population. The standard deviations of $\tilde{r}_n$ in DFT and EN are 0.641 and 0.724, respectively. As a result, they capture 93% and 90% of the estimated population standard deviation of the r parameter, respectively, for DFT and EN. Those broad coverages then allow us to use the conditional mean of the r parameter for each individual in elucidating the background of the stochastic models' performance which will be momentarily discussed.[18]

C. Goodness of Fit under EUT

We now evaluate the stochastic models based on their performance in approximating the true data-generating process, using log-likelihood, AIC, and BIC values. In contrast to the former, a lower score of the latter two measures represent better statistical goodness-of-fit. AIC is given by $AIC = -2 \times LL + 2 \times \varkappa$, where LL is the maximized log-likelihood value and $\varkappa$ is the number of estimated parameters. BIC is given by $BIC = -2 \times LL + \varkappa \times \ln(N)$, where N is the number of individuals in the sample. An improvement in LL, therefore, has the same effects on both AIC and

expectation for each decision-maker. As previously described, each individual's conditional distribution is derived from and proportional to the unconditional distribution. Altogether, the conditionals distributions for all decision makers aggregate to the unconditional distribution of the population. As a result, the unconditional variance is a combination of the expectation of the variance of the conditional distribution and the variance of the conditional means (Revelt and Train [2000] and Daly et al. [2012]). To be specific, the unconditional variance of the r parameter can be written as $Var(r_n) = E[Var(r_n|y_n, \theta, \mu)] + Var[E(r_n|y_n, \theta, \mu)]$. Therefore, $Var(r_n) \geq Var(\tilde{r}_n)$. As the number of choice observations for each individual gets arbitrarily large, the expectation of the variance of the conditional distribution $E[Var(r_n|y_n, \theta, \mu)]$ converges in probability to zero, and the variance of the conditional means $Var(\tilde{r}_n)$ converges in probability to the variance of unconditional distribution $Var(r_n)$.

[18] Distinguishing decision makers using the individual-level parameter is also useful for further analysis. For example, Gao, Harrison, and Tchernis [2020] use the individual-level risk parameter to identify the welfare effect from insurance products at the individual level. The individual-level parameter is also used to determine the optimal temporary cost reduction in a targeted marketing campaign (Allenby and Rossi [1998]). Sandorf, Campbell, and Hanley [2017], on the other hand, use the individual-specific parameter to derive the individual-specific willingness-to-pay.

BIC, but the penalty for including extra parameters varies across the two criteria, with the former being more lenient than the latter.

With EUT as the underlying decision theory, EN is the best error model with the ranking of EN $>_{in}$ CU $>_{in}$ FN $>_{in}$ DFT. The relation sign $>_{in}$ implies that the former model is preferred to the latter in the in-sample comparison, and the ranking is the same for AIC and BIC. Panel A of Table 3 reports the "grand log-likelihood" value aggregated from the 413 individuals, AIC, and BIC of each stochastic model. The maximized log-likelihood of EN is −5,838, which is much higher than the log-likelihood values of the other stochastic models, which are −6,096, −6,344, and −6,245, respectively, for CU, DFT, and FN. The only factor that potentially differentiates the information criteria ranking from the log-likelihood ranking is the number of parameters. With 413 individuals and two extra parameters, the AIC and BIC penalties on the twice log-likelihood value for EN are only 4 and 12.05, respectively. Therefore, even after considering the penalty for extra parameters, AIC and BIC agree with the log-likelihood ranking by selecting EN as the best fit model under EUT. With those log-likelihood values, the AIC and BIC scores of EN are higher than the remaining models. The AIC (BIC) generated by EN is 11,686 (11,706), while CU, FN, and DFT generate AIC (BIC) scores of 12,199 (12,211), 12,496 (12,508), and 12,695 (12,707), respectively.

The comparison of information criteria between error models dichotomies the better and worse models arbitrarily as it makes little difference on the ranking of the models whether the AIC or BIC difference is 1 or 100. To allow for a more meaningful interpretation, we now turn to the Vuong test that considers the significance of disparity in goodness-of-fit between models. Figure 2 displays the distribution of the individual log-likelihoods that add up to the maximized log-likelihood value.[19] The Vuong test assumes that the log-likelihood differences between two error

[19] We derive each decision maker's log-likelihood contribution by inserting the estimated distribution parameters of CRRA coefficient, $f(r_n; \hat{\theta})$, and the estimated noise parameter $\hat{\mu}$ into the likelihood functions in (6).

specifications are asymptotically normally distributed and conducts a z-test on them. The null hypothesis is that the two error models are equally close to the true data generating process, against the alternative that one model is closer. The Vuong tests reported in Table 4 supports the results under the information criteria.[20] All tests consistently favor EN and disfavor DFT. The test between CU and FN yield a z-statistic that points to the direction of CU. The tests show the evidences of significant difference in performance between error models with p-values of at least 0.007.

We now compare the out-of-sample forecasting performance of the stochastic models using the data from the second experiment. Based on statistical goodness-of-fits reported in Panel B of Table 2, we find that the out-of-sample ranking agrees with the in-sample result given by EN $>_{out}$ CU $>_{out}$ FN $>_{out}$ DFT. The out-of-sample log-likelihood of EN from the 182 decision makers is -2,730. The out-of-sample predictions by the remaining error models yield log-likelihood values of −2,946, −3,017, and −3,104, respectively, for CU, FN, and DFT. To compare the significance of the difference in out-of-sample log-likelihood between error models, we use the Vuong test without any penalty for extra parameters. The results presented in Table 5 show that the differences in the goodness-of-fit between error models are significant, with p-values of at least 0.001.

From the individual log-likelihood contributions, we calculate the percentage of decision makers for whom an error model's log-likelihood value is the highest at the individual level. We then use the results to represent the share of each error model in the population. As illustrated in Figure 3, we find that the ranking of the four specifications in the population shares is almost consistent with the earlier goodness-of-fit rankings based on log-likelihood. Among the 413 decision makers in the in-sample data, 57.9% exhibit behavioral error that is consistent with EN. CU makes up the

[20] Following Desmarais and Harden [2013], we use the modified Vuong test that imposes a penalty on the log-likelihood difference due to less parsimony. The corrected log-likelihood difference of two models, say model 1 and model 2, in each observation i is given by $\nabla LL_i = \ln L_{1i} - \ln L_{2i} + k \times \ln(N) / 2N$, where $k = \#_2 - \#_1$.

majority of decision makers in the population, with a share of 18.4%, while FN represents 16.2% of subjects in the population. Finally, DFT characterizes 7.5% of decision makers in the population.

Comparing the individual-level log-likelihood of 182 decision makers in the out-of-sample data, we find that EN composes 62.1% shares of the population. The population share of FN is larger than CU, which is 20.3% against 15.4%. Finally, there are only 2.2% of decision makers in the population whose behavioral error is better characterized by DFT. The population share rankings in the out-of-sample are then different from the ranking based on log-likelihood value, where FN now outperforms CU. This difference is possible because, in this individual log-likelihood comparison, we only determine which model best fits the decision maker's stochastic behavior, regardless of the magnitude of individual log-likelihood differences between models.

To sum up, both in- and out-of-sample analyses show that two of the three heteroskedastic models perform better than FN, with EN being the best model in explaining the stochastic choice behavior of decision makers. The inferior performance of DFT relative to FN implies that merely allowing for heteroskedasticity with respect to task complexity does not necessarily improve the goodness-of-fit of a model. What matters is which form of heteroskedasticity is introduced.

D. The Background of Error Models' Performances under EUT

We now seek to explain why a stochastic model is superior to the others. We first consider the error rate of choice prediction in the in-sample data, represented by the mean absolute error (MAE). It is defined as the absolute difference between the predicted choice probability $P_{nt}(\tilde{r}_n, \hat{\mu})$, calculated using the individual-level parameter $\tilde{r}_n$ and the estimated error parameter $\hat{\mu}$, and the observed choice y_{nt}. For example, when the observed choice by subject n at decision task t is lottery A, and the predicted probability of choosing A by that subject at that task is 60%, the MAE is 40%.

Accordingly, when the predicted probability of choosing A is 60%, while the observed choice is B, the MAE is 60%.

Figure 4 displays the MAE of each stochastic model when combined with EUT. The best performance of EN is supported by its lower error of 0.206. The MAEs of CU and FN are 0.234 and 0.242, respectively. As the worst model under EUT, DFT has the highest MAE, which is 0.247. As MAE measures the deviation of predicted choice probability from choosing a lottery with certainty, the lower MAE suggests a more apparent dichotomy in the probability of choosing a particular option. In other words, with the lowest MAE, EN predicts a higher probability of choosing lottery A or lottery B than its counterparts. When the choice prediction is correct, the higher predicted choice probability leads to a higher log-likelihood. The low MAE thus provides an overview of how the EN model can provide the best performance.

We now scrutinize the factors that contribute to the predicted choice probability. Recall that the four stochastic specifications yield a similar population mean of CRRA coefficient, which is between 0.531 and 0.569. The estimates of the CRRA coefficient indicate that, on average, the decision makers switch between lottery A to lottery B when the highest prizes are offered with a probability of 50% or 60%. In each choice block, the indifference point is thus located around the middle of the list. The average location of indifference point is also confirmed by the trend of prediction error illustrated in Figure 4. In line with the finding by Alós-Ferrer and Garagnani [2021] and Dickhaut et al. [2013], the choice inconsistency is higher around the indifference point, that is at the lottery choices that offer the highest prizes with 50% or 60% probability.[21] As we move away

[21] Similarly, the trend in the success rate displayed in Figure C1 in Appendix C also shows lower choice consistency around the point where the decision makers are indifferent between the two options. The success rate is calculated based on the percentage of correct choice prediction. Specifically, the choice prediction is successful when the observed choice is A, and the predicted probability of choosing A is at least 0.5, and *vice versa*.

from the indifference point to the beginning and the end of the choice list, the decision makers have a clearer preference and, therefore, a higher choice consistency.

At around the indifference point, that is when the decision makers are uncertain about their preference over the two options, the probability of choosing a particular option is around 50:50, and the ∇EU index should be close to zero. As we move away from the indifference point, the index will be higher in its absolute value. Recall from Section 3, a higher ∇EU index is contributed by a larger EU difference ($EU_A - EU_B$) or a lower standard deviation of behavioral errors, or a combination of the two. As choices at points far from indifference are primarily consistent, it is innocuous to compare the EU difference and the standard deviation of the error term of the four stochastic models at around the beginning and the end of the choice list to explain the background of the models' relative fitness.

Figure 5 depicts distributions of expected utility differences, which are calculated using the individual-level parameter. The means of the EU difference in EN and DFT are 0.934 and 0.616, respectively. The mean of EU differences in CU is similar to that in FN, which is 0.497 against 0.492. The test statistics show that those mean values are statistically significant from zero (*p*-values ≤ 0.001). With its highest population variance of the r parameter, EN has the highest standard deviation of EU differences, which is 4.771. DFT and FN have lower standard deviations of EU differences at 2.318 and 1.933, respectively. Finally, the standard deviation of the EU difference in CU is the lowest, which is 1.877. We can also compare the 25th and 75th percentiles of the distributions to see the contribution of EU difference to the model's performance. The 25th percentiles of the four models are similar, which are -0.337, -0.350, -0.332, and -0.359, respectively, for FN, CU, DFT, and EN. The 75th percentiles, however, vary between models. Those in FN and CU are similar which are, 0.820 and 0.829, respectively. Meanwhile, the upper quartiles in DFT and EN are higher at 0.901 and 0.882, respectively. With more deviate EU differences relative

to zero, EN has a larger proportion of large EU differences in the absolute term, potentially leading to a higher ∇EU index and, eventually, a higher log-likelihood value and better fit. However, such an indication is not apparent for the ranking of the remaining stochastic models, which requires us to examine the standard deviation of the error term to explain their relative performance.

Figure 6 illustrates the shape and magnitude of the models' standard deviation of the error term. We first compare EN with FN based on the results of the test statistics on the estimates of β_1 and β_2. Recall that if each entropy coefficient equals zero, the heteroskedastic factor υ in EN becomes unity, which means EN simplifies to FN. Table 1 shows that estimates of β_1 and β_2 are 5.314 (p-value is 0.003) and −11.895 (p-value <0.001), respectively. A Wald test also rejects the null hypothesis that β_1 and β_2 are jointly equal to zero with a p-value <0.001. Both single and joint hypothesis tests thus reject FN in favor of EN. The significant estimated values of β_1 and β_2 also imply that the standard deviation of the error term is a function of complexity. With a positive estimate of β_1 and a negative estimate of β_2, we find that the relationship between the standard deviation of the error term and task complexity is represented by an inverse-U function, thereby rejecting the monotonic relationship postulated by CU and DFT. This may also explain the best performance of EN.

Figure 6 shows that the standard deviation of the error term in EN is mostly low except when the highly risk averse decision makers with CRRA coefficient between 1-1.5 are confronted with two lotteries that offer the high prizes with a low probability. However, the combined effect of the mean and standard deviation of the r parameter in EN suggests that those decision makers only comprise 28.6% of the population. Additionally, with this type of decision makers, EN can still benefit from a combination of higher positive EU differences and lower standard deviation of behavioral errors at lottery choices with a moderate to a high probability of receiving the high prizes. For the less risk averse or less risk seeking decision makers, the standard deviation of the error term

is primarily low in any decision task. The low standard deviation of the error term at the majority of choice tasks and for majority decision makers contribute to the better relative performance.

As discussed in Section 3, the standard deviation of behavioral errors in CU and DFT vary in stakes and change monotonically with the CRRA coefficient. The difference between the two is that the behavioral errors in the latter model vary with outcome probability, while the former does not. The inverse U-shape of the standard deviation of behavioral errors is determined by the product of the probability of the best and worse outcomes, so it reaches its maximum value when each lottery offers the two prizes with a 50:50 probability. The standard deviation of the error term in CU increases abruptly from its lowest points at the beginning and the end of the choice list, that is when the high prizes is offered with a probability of 10% or 90% (see Panel C of Figure 6). When the probability of receiving the high prizes increases (decreases) to 20% (80%), the standard deviation of the error term increases by a factor of 1.78. It then multiplies 1.31 times when the probability of the high prize increases (decreases) further to 30% (70%). The increase rate continues to drop as we approach the maximum value. Figure 6 also illustrates how DFT has primarily higher magnitudes of the standard deviation of behavioral errors than the other stochastic models. The relatively higher standard deviations of the error term then severely scaled-down DFT's expected utility differences, leading to lower absolute values of ∇EU indices and lower log-likelihood values. It thus explains the weak performance of DFT despite its relatively higher mean and standard deviation of the EU difference.

That in FN is, of course, flat and fixed for any decision makers in any task characteristics. Its inflexibility implies drawbacks. A high standard deviation of behavioral errors will cost FN with a lower log-likelihood value relative to a more flexible heteroskedastic model as the high standard deviation scales down the high expected utility differences at points far from indifferent. When the standard deviation is too small, it indeed leads to higher ∇EU indices at points far from indifferent.

On the flip side, when the subjects make noisier choices around the indifference point, FN suffers from lower individual log-likelihood values due to many missed choice predictions. Despite being insensitive to the change of prize probability, CU is more flexible than FN as its standard deviation of the error term varies with the lotteries' highest and lowest prizes and the r parameter. CU can benefit from higher individual log-likelihood values contributed by less risk averse subjects whose choices are more consistent due to lower standard deviations of behavioral errors. The highly risk averse or highly risk seeking individuals contribute to the CU's higher log-likelihood by minimizing the penalty due to inaccurate predictions on noisier choice at around the indifference point.

To summarize, the better performance of a stochastic model is supported by its higher accuracy in prediction choice. The higher population standard deviation of parameter r also contributes to the relative fitness of a stochastic model by providing an advantage in terms of more dispersed EU differences, leading to higher ∇EU indices and potentially higher log-likelihood values. In addition, the heteroskedastic specifications also benefit from their non-constant standard deviation of behavioral errors. Their specification implies flexibility, which may better represent the pattern of choice inconsistency than the homoskedastic specification. When the choices are more consistent at particular task characteristics, the combination of lower standard deviations of behavioral errors and higher EU differences generate a higher log-likelihood value. An exception is found in DFT as its relatively higher standard deviation of behavioral errors diminishes its advantage of having higher mass in the tails of EU difference distribution.

E. Inferred Risk Attitudes under Rank-Dependent Utility Theory (RDU)

We now expand our analysis by embedding the stochastic models to RDU. We now find some sensitivity of the inferred risk attitudes, with Table 6 presenting the estimation results. First, consider the utility curvature parameter as illustrated in Panel A of Figure 7. We observe that the

estimates of the population mean of the r parameter in FN, DFT and EN are almost identical, taking values of 0.718, 0.790, and 0.748, respectively (p-values < 0.001). The estimated population standard deviation of the r parameter in FN is 0.450 (p-value < 0.001), while those in DFT and EN are 0.939 and 0.490, respectively (p-values < 0.001). The test statistics cannot reject the null hypothesis that the population means of the r parameter in the three stochastic models are equal with p-values≥0.160 (see Table D1 in Appendix D). The estimate of the population mean of CRRA coefficient in CU is smaller at 0.560 (p-value < 0.001), with an estimated standard deviation of 0.417 (p-value < 0.001), suggesting a lower degree of risk aversion in terms of the utility function. The test statistics for the difference in means now reject the null hypothesis that the population mean of the r parameter in CU is equal to those in other stochastic models with p-values < 0.001.

For all four stochastic models, we reject the hypothesis that, on average, the decision makers are risk neutral with p-values < 0.001. The combination of the estimated mean and standard deviation of the r parameter in each model suggests that FN, CU, EN shares an almost similar proportion of risk averse decision makers in the population, which are 94.5%, 91.0%, and 93.7%, respectively, with p-values < 0.001. Under DFT, the risk-averse decision-makers only make up 80.0% of the population (p-value < 0.001).

Turning to the probability weighting function, we observe similar estimates of the population mean of the φ parameter in all stochastic models. Panel B of Figure 7 illustrates its distributions. The estimates of the population mean of φ in FN and CU are 2.560 and 2.574, respectively (p-values < 0.001). Meanwhile, those in DFT and EN are 3.012 and 2.772, respectively (p-values < 0.001). The estimates of the standard deviation of the φ parameter are 5.280 (p-value < 0.001) in FN , 5.199 (p-value < 0.001) in CU, and 5.669 (p-values = 0.010) in DFT. The estimation of RDU with EN yields a standard deviation of the φ parameter of 5.789 (p-value < 0.001). We use the method proposed by Zhou, Gao, and Hui [1997] to test the difference in means of two log-

normal distributions and cannot reject the null hypothesis that the population means of the φ parameter between stochastic models are equal with *p*-values of at least 0.167 (see Table D2).

Panel C of Figure 7 presents the estimates of the η parameter, which vary between models. DFT has the largest estimate of the population mean of the η parameter, that is 1.775 (*p*-value < 0.001). The estimated population standard deviation of the parameter is 2.061 (*p*-value = 0.002). The estimates of the population mean of the η parameter in FN, CU, and EN are 1.281, 1.687, and 1.350, respectively (*p*-values < 0.001). Their respective estimates of the population standard deviation of the η parameter are 1.153, 1.584, and 1.361, with *p*-values < 0.001. The test statistics reported in Table D3 now reject the null hypothesis of the same population means of the η parameter between two stochastic models with *p*-values = 0.001, except for those between CU and DFT (*p*-value = 0.257) and between FN and EN (*p*-value = 0.257). In all cases, we reject that the population means of φ and η equal unity, both using single and joint hypothesis tests at 5% significance level, implying the non-linear probability weighting function and rejection of EUT in favor of RDU.

Given the estimates, all error specifications agree that the probability weighting function, on average, takes an S-shape, implying pessimism for small probability and optimism for large probability (see Figure 8). Examining the population distribution of the φ parameter, we find that the decision makers with the S-shape function make up 61.6% of the population under FN (*p*-value < 0.001). Under CU and DFT, they make up 61.9% (*p*-value < 0.001) and 63.9% (*p*-value < 0.001) of the population, respectively. Under EN, their share in the population is 62.0% (*p*-value < 0.001).

F. Individual-level Parameters under RDU

Table 7 contains the descriptive statistics of the individual-level parameters under RDU given by each stochastic model. We still observe an almost negligible difference between the mean

of individual-level parameter $\tilde{r}_n$ and the estimated mean of the CRRA coefficient in each model. The mean of $\tilde{r}_n$ in FN is 0.701, while that in DFT is 0.702. EN and CU generate means of posterior distribution the r parameter of 0.723 and 0.558, respectively.

Similarly, we also observe an almost similar sample average of posterior mean of each probability weighting parameter to the estimated population mean. In RDU with FN, the means of $\tilde{\varphi}_n$ and $\tilde{\eta}_n$ are 2.115 and 1.282, respectively. With CU, the mean of individual-level $\tilde{\varphi}_n$ is 2.277, while the sample average of $\tilde{\eta}_n$ is 1.720. Under DFT, the sample mean of $\tilde{\varphi}_n$ of the 413 subjects is 2.459, while that under EN is 2.242. Finally, the mean of $\tilde{\eta}_n$ in those two error models are 1.783 and 1.362, respectively.

We find that the conditional means of the risk parameters in each model capture a fairly large portion of the variation in the risk aversion across decision makers, implying an almost similar figure of the standard deviation of the individual-level parameters and the estimated standard deviation of the parameters in the population. The standard deviations of the conditional mean of the r parameter in FN and DFT are 0.300 and 0.625, respectively, capturing 67% of their estimated population standard deviation of the parameter. With standard deviation of 0.270 and 0.321, respectively, for CU and EN, the conditional means capture 65% of the total variation in the population's CRRA coefficient estimated by CU and EN.

CU seems better in capturing variation in the estimated parameters of decision weight than the other models. With a standard deviation of $\tilde{\varphi}_n$ of 2.621, the $\sigma_{\tilde{\varphi}n}/\sigma_r$ ratio in CU is 52%. The ratio is higher than the other three models, which is 45%, generated by standard deviations of 2.597, 2.551, and 2.382, respectively, for DFT, EN, and FN. The standard deviations of the conditional mean $\tilde{\eta}_n$ in CU, DFT, and EN are 2.277, 2.459, and 2.242, respectively, while FN generates a standard deviation of 2.115. As the results, the posterior mean $\tilde{\eta}_n$ in CU captures a higher estimated

variation in the η parameter than the other models, which is 72% against 66% in FN and DFT. A slightly lower ratio is given by EN, which is 70%.

G. Goodness of Fits under RDU

When the error models are combined with RDU, we observe that DFT much improves their ranking in the in-sample evaluation. CU remains the second-best model, while EN is downgraded to the third position. The in-sample ranking is thus given by DFT $>_{in}$ CU$>_{in}$ EN $>_{in}$ FN. As presented in Panel A of Table 8, with the aggregate log-likelihood of −5,142, the estimation of RDU with DFT yields an AIC (BIC) score of 10,305 (10,345). The AIC (BIC) score of RDU with CU is 10,361 (10,401), generated by a maximized log-likelihood value of −5,170. Aggregated from the 413 subjects, the log-likelihood of EN are −5,189, leading to an AIC (BIC) score of 10,402 (10,451). Finally, the log-likelihood value of −5,219 in FN generates an AIC (BIC) score of 10,459 (10,449).

Different from EUT, the gap in performance between models is now narrowed. Figure 9 shows that the distributions of individual log-likelihoods of the stochastic models are almost similar. Comparing the individual-level log-likelihood values in each pair of error models, the Vuong tests presented in Table 9 cannot reject the null hypothesis of non-discriminated goodness-of-fit between CU, DFT, and EN at a 5% significance level with *p*-values between 0.054 and 0.247. However, the Vuong tests between FN against each heteroskedastic specification reject the null hypothesis of equivalent performance with *p*-values between 0.002 and 0.020.

We observe a consistent ranking to the in-sample comparison for the out-of-sample prediction, that is, DFT $>_{out}$ CU $>_{out}$ EN $>_{out}$ FN. Panel B of Table 8 shows that DFT, CU, and EN have log-likelihood values of −2,334, −2,344, and −2,352, respectively. Finally, FN's out-of-sample log-likelihood is −2,374. As presented in Table 10, we also find insignificant differences in

out-of-sample log-likelihoods between CU, DFT, and EN, and significant log-likelihood differences between FN and each heteroskedastic model.

Using the individual log-likelihood contribution, we now calculate the population share of the stochastic models under RDU. As shown in Figure 10, DFT has the largest shares in the in-sample data by best fitting the observed choices of 33.7% decision makers in the population. Among the 413 decision makers, 31.7% display a stochastic behavior consistent with EN. Finally, the shares of the population that suit the characteristic of CU and FN are 21.5% and 13.1%, respectively. In the out-of-sample data, we find that the population of 182 subjects consists of 40.7% DFT, 28.6% EN, 21.4% CU, and 9.3% FN. The population share rankings thus disagree with the fitness ranking based on log-likelihood, AIC, BIC values, in which EN now outperforms CU. Again, such a difference is possible as this population share evaluation employs an ordinal comparison, ignoring the magnitude of log-likelihood difference between models.

To summarize, the results of in- and out-of-sample analyses emphasize the need to allow behavioral error to vary with task complexity by rejecting homoskedasticity in favor of the three heteroskedastic models. From the worst fit model under EUT, DFT is now be the winner under RDU. CU remains the second-best model, outperforming EN and FN. Next, we discuss the background of the changing ranking from EUT to RDU.

H. The Background of Error Models' Performances under RDU

The better performance of the three heteroskedastic specifications is supported by their higher accuracy in predicting the decision makers' choices. As illustrated in Figure 11, the MAEs of CU, DFT, and EN are similar at 0.162, while that of FN is 0.165. The error rate of choice prediction

by RDU with any stochastic model also exhibits the same trend as in EUT.[22] Therefore, we can similarly compare the RDU differences ($RDU_A - RDU_B$) between models and their combination with the standard deviations of behavioral errors at points far from indifference to explain the models' underlying performance.

Figure 12 shows the distribution of the RDU differences ($RDU_A - RDU_B$) generated by individual-level parameters in each stochastic model. Overall, introducing the non-linear probability weighting function improves the error models' goodness-of-fit by increasing the proportion of large RDU differences in the absolute term. As the best stochastic model for RDU, DFT has the highest mean (standard deviation) of RDU differences, which is 0.651 (1.626). CU, the second-best model under RDU, provides the lowest mean (standard deviation) of RDU difference, which is 0.335 (0.971). The means (standard deviations) of the RDU difference in EN and FN are 0.434 (1.075) and 0.405 (1.024), respectively. The test statistics reject the null hypothesis of the insignificant mean of RDU difference in each model (p-values$\leq$0.001). Comparing the 25th percentiles of the distributions, CU has the smallest value of −0.443, while the remaining models have lower quartiles between −0.393 and −0.401. The 75th percentiles of the RDU difference in DFT, CU, and EN are 1.339, 1.044, and 1.099, respectively. That in FN is smaller at 1.037. The results imply that DFT has a larger proportion of large RDU differences in absolute value than its counterpart, contributing to its better performance. Albeit with less spreading RDU difference, the better fit of CU relative to EN and FN may be contributed by its platykurtic shape of the distribution of RDU difference that has more masses in its shoulders with values that are distant from zero. Additionally, the better performance of CU may also be explained by the feature of its heteroskedastic form.

[22] Similarly, the success rate of choice prediction also shows higher error or choice consistency at around indifference point (see Figure E1 in Appendix E).

The significant fitness improvement of DFT is also contributed by the non-linear probability weighting function that enters its heteroskedastic specification. Note that the estimated population mean of the η parameter in DFT is 1.775. With this estimate, the maximum value of the complexity measure σ_{A-B} is located at the tasks that offer a 70% chance of getting the highest prizes. This maximum value is thus still at around indifference points and in line with the choice pattern that exhibits high inconsistency around that point. DFT also generates the largest estimate of the population mean of the φ parameter. Recall from Section 3, as φ increases, the complexity measure σ_{A-B} becomes tighter around its maximum value and shallower around its edges. The resulting standard deviation of the error term is thus lower than when we assume linear probability, especially around the beginning and the end of the choice list (see Panel A of Figure 13). As a result, DFT gains an advantage from the combination of high RDU differences and very low standard deviations of behavioral errors at points far from indifference.

The better performance of CU relative to EN is contributed by its relatively lower estimate of the population mean of the r parameter than its counterparts. Under RDU, the other three stochastic models generate a higher population mean of the CRRA coefficient. That in CU is stable across the two underlying decision theories. The resulting complexity measure in CU under RDU is thus similar to that under EUT. Furthermore, the estimated error parameter μ of CU is lower in RDU than in EUT, that is 0.148 (p-value < 0.001) against 0.190 (p-value < 0.001), leading to a lower standard deviation of behavioral errors. CU also ignores the complexity arising from the change of lotteries' probabilities, making it insensitive to the magnitude of decision weight parameters. As the standard deviation of the error term in CU changes with the range between the lotteries' highest and lowest prizes and with the r parameter, the model exhibits a relative advantage over FN when the lottery stakes and the r parameter are low. Meanwhile, the standard deviation of the error term in FN is fixed for any decision task and for any type of decision makers, which impedes the model

from benefitting from the combination of higher RDU difference and lower standard deviation of behavioral errors at points far from indifference.

We can compare EN with FN based on test statistics on the estimates of β_1 and β_2. If both parameters equal zero, the EN specification is simplified to FN. The result from Table 4 shows the significant estimates of β_1 and β_2, which are −3.237 (p-value < 0.001) and 4.317 (p-value < 0.001), respectively. The Wald test also rejects the null hypothesis that β_1 and β_2 are jointly equal to zero (p-value < 0.001), thus rejecting FN in favor of EN.

With the estimated sign and magnitude of parameters β_1 and β_2, EN represents the relationship between the standard deviation of the error term and task complexity with a U-shape graph. The relationship implies that the decision makers respond to a more complex task with more careful evaluation, thus resulting in a lower standard deviation of the behavioral error. However, when the task becomes too complex, the decision makers simplify the evaluation process, leading to greater error variance.

The U-shape quadratic function in EN seems to impair its relative performance despite its relatively higher mean and standard deviation of RDU difference. As illustrated in Panel D of Figure 13, the standard deviation of behavioral errors is also now higher at the beginning and end of the choice list, thus attenuating the large RDU differences at points far from indifference. The standard deviation of behavioral errors is also higher at around the indifference point.

To sum up, the better fit of heteroskedastic error models relative to their homoskedastic counterpart is supported by their higher accuracy in predicting the decision makers' choices. The better fit of an error model under RDU can also be explained by the higher proportion of the large RDU differences. A significant improvement in the relative performance of DFT under RDU is also contributed by its largest estimates of the population mean of the parameter that controls the curvature of the probability weighting function. With such a large estimate, the standard deviation of

behavior errors is now tighter around its peak and shallower around its edges. As the indifference points are around the middle of the choice list, the combination of the high RDU differences and the lower standard deviation of behavioral errors at the beginning and end of the choice generates higher ∇RDU indices and log-likelihood values. The estimated shape relationship between task complexity and behavioral errors now costs EN with weaker performance under RDU, as the standard deviation of the error term around the indifference point is now higher than under EUT. It is also higher around its edge as compared to DFT and CU. The relatively steady performance of CU across the two decision theories is due to the stable estimate of the CRRA coefficient and its ignorance of the potential changing complexity due to variation of lotteries' probabilities. Finally, the results then suggest that the monotonically increasing functions in CU and DFT better explain the behavioral error story in RDU since the significant β_1 and β_2 in the EN model have a limited impact on the performance of EN relative to CU and DFT.

5. Conclusion

Apart from being a nuisance parameter, the behavioral error term is also substantively compelling. The fluctuation in its magnitude implies heterogeneity of behavioral responses to a different choice complexity. The study reported here examines how the relationship between decision complexity and behavioral error is better specified. It is done by evaluating alternative models of stochastic choice that incorporate the effects of task complexity on behavioral errors. We consider two specifications that have been applied in many experimental studies of decision making under risk, namely the Contextual Utility (CU) and the Decision Field Theory (DFT). We contribute to the literature by introducing the third heteroskedastic specification, namely the Entropy (EN) model, which is borrowed from a study of consumer behavior and applying it to decision making under risk. Each model operationalizes a distinct notion of which decision task is more complex

than another, and EN allows for a more flexible type of association between the standard deviation of behavioral errors and task complexity.

We embed each stochastic model with two alternative theories of decision making under risk, namely Expected Utility Theory (EUT) and Rank-Dependent Utility (RDU). Regarding inferred risk attitudes, we find that the EUT results are robust across alternative stochastic models, whereas the RDU results show some sensitivity. Our empirical findings emphasize the importance of accommodating the effects of task complexity on behavioral errors, by rejecting homoskedasticity in favor of two of the three heteroskedastic models. The in- and out-of-sample analyses show that EN provides the best fit to the data under EUT, and DFT provides the best fit under RDU.

Our analytical results show that the superior performance of EN under EUT and DFT under RDU in the in-sample analysis may be explained by their ability to capture a more diverse behavior at the individual level. The resulting distributions of expected utility and rank utility differences are more dispersed and have larger values in absolute terms. The changing of DFT's ranking from the worst in EUT to the best in RDU is contributed by its higher estimates of the population mean of the parameter that controls the curvature of the probability weighting function, leading to a lower standard deviation of behavioral errors, especially at points far from indifference.

There are two implications in our findings. First, the better performance of EN under EUT and DFT under RDU indicates that every aspect of the lotteries contributes to the complexity of the decision task and, eventually, affects behavioral errors. Second, the ability to better capture a more diverse risk attitude may be useful for designing a policy that targets individuals with a relatively extreme risk behavior in terms of utility curvature or probability weighting function.

Table 1: MSL Estimates of CRRA Coefficient and Error Parameter(s) under EUT

Variable	Estimate	St. Error	*p*-value	95% Confidence Interval	
			A. With Fechner Error (FN)		
$\bar{r}$	0.535	0.031	<0.001	0.474	0.596
σ_r	0.608	0.053	<0.001	0.504	0.712
μ	0.565	0.029	<0.001	0.508	0.622
			B. With Contextual Utility (CU)		
$\bar{r}$	0.531	0.033	<0.001	0.466	0.596
σ_r	0.635	0.052	<0.001	0.534	0.736
μ	0.190	0.008	<0.001	0.174	0.207
			C. With Decision Field Theory (DFT)		
$\bar{r}$	0.569	0.037	<0.001	0.497	0.641
σ_r	0.689	0.055	<0.001	0.580	0.797
μ	0.610	0.030	<0.001	0.550	0.669
			D. With Entropy (EN)		
$\bar{r}$	0.547	0.069	<0.001	0.411	0.683
σ_r	0.801	0.121	<0.001	0.565	1.038
μ	2.291	0.271	<0.001	1.760	2.823
β_1	5.314	1.766	0.003	1.852	8.776
β_2	−11.895	2.782	<0.001	−17.348	−6.442

Table 2: Descriptive Statistics of Posterior Mean of CRRA Coefficient under EUT

Model	Obs.	Mean	Std. Dev.	Min	Max
		Individual-Level CRRA Coefficient			
Fechner Errors (FN)	413	0.547	0.547	−1.355	2.248
Contextual Utility (CU)	413	0.539	0.586	−1.431	2.163
Decision Field Theory (DFT)	413	0.574	0.641	−1.559	2.278
Entropy (EN)	413	0.557	0.724	−2.063	2.827

Table 3: Summaries of The Statistical Goodness-of-Fit Measures under EUT

Stochastic Model	No. of Parameters	Log-likelihood	AIC	BIC	Rank
		A. In-Sample			
Fechner Error (FN)	3	−6,245	12,496	12,508	3
Contectual Utility (CU)	3	−6,096	12,199	12,211	2
Decision Field Theory (DFT)	3	−6,344	12,695	12,707	4
Entropy (EN)	5	−5,838	11,686	11,706	1
		B. Out-of-Sample			
Fechner Error (FN)	3	−3,017	6,040	6,049	3
Contectual Utility (CU)	3	−2,946	5,897	5,907	2
Decision Field Theory (DFT)	3	−3,104	6,214	6,223	4
Entropy (EN)	5	−2,730	5,469	5,485	1

Table 4: Vuong Tests Between the "Column" Specifications Against the "Row" Specifications under EUT in In-Sample Data
(The *p*-value is provided in the parenthesis)

	FN	**CU**	**DFT**
CU	−5.176 (<0.001)		
DFT	2.690 (<0.001)	13.380 (<0.001)	
EN	−7.283 (<0.001)	−5.521 (<0.001)	−9.389 (<0.001)

Table 5: Vuong Tests Between the "Column" Specifications Against the "Row" Specifications under EUT in Out-of-Sample Data
(The *p*-value is provided in the parenthesis)

	FN	**CU**	**DFT**
CU	−3.287 (0.001)		
DFT	3.372 (0.001)	10.941 (<0.001)	
EN	−5.808 (<0.001)	−5.144 (<0.001)	−7.483 (<0.001)

Table 6: MSL Estimates of CRRA Coefficient, Parameters of Probability Weighting Function, and Error Parameter(s) under RDU

Variable	Estimate	St. Error	*p*-value	95% Confidence Interval	
			A. With Fechner Error (FN)		
$\bar{r}$	0.718	0.035	<0.001	0.648	0.787
σ_r	0.450	0.059	<0.001	0.335	0.565
$\bar{\varphi}$	2.560	0.230	<0.001	2.109	3.010
σ_φ	5.280	0.988	<0.001	3.344	7.215
$\bar{\eta}$	1.281	0.116	<0.001	1.054	1.509
σ_η	1.153	0.198	<0.001	0.766	1.540
$\varrho_{r\varphi}$	0.068	0.026	0.008	0.018	0.119
$\varrho_{r\eta}$	−0.322	0.091	<0.001	−0.500	−0.144
$\varrho_{\varphi\eta}$	0.198	0.040	<0.001	0.119	0.277
μ	0.368	0.018	<0.001	0.332	0.403
			B. With Contextual Utility (CU)		
$\bar{r}$	0.560	0.038	<0.001	0.486	0.635
σ_r	0.417	0.084	<0.001	0.252	0.581
$\bar{\varphi}$	2.574	0.283	<0.001	2.020	3.128
σ_φ	5.199	1.159	<0.001	2.928	7.470
$\bar{\eta}$	1.687	0.178	<0.001	1.339	2.035
σ_η	1.584	0.278	<0.001	1.039	2.128
$\varrho_{r\varphi}$	0.031	0.038	0.416	−0.043	0.105
$\varrho_{r\eta}$	−0.258	0.124	0.038	−0.501	−0.015
$\varrho_{\varphi\eta}$	0.296	0.101	0.003	0.098	0.494
μ	0.148	0.010	<0.001	0.128	0.167
			C. With Decision Field Theory (DFT)		
$\bar{r}$	0.790	0.132	<0.001	0.531	1.049
σ_r	0.939	0.260	<0.001	0.430	1.449
$\bar{\varphi}$	3.012	0.775	<0.001	1.493	4.530
σ_φ	5.669	2.191	0.010	1.374	9.963

$\bar{\eta}$	1.775	0.359	<0.001	1.072	2.479
σ_η	2.061	0.671	0.002	0.746	3.376
$\varrho_{r\varphi}$	0.175	0.055	0.001	0.068	0.282
$\varrho_{r\eta}$	−0.412	0.080	<0.001	−0.568	−0.256
$\varrho_{\varphi\eta}$	0.120	0.051	0.019	0.019	0.221
μ	0.448	0.044	<0.001	0.362	0.535

D. With Entropy (EN)

$\bar{r}$	0.748	0.041	<0.001	0.667	0.829
σ_r	0.490	0.055	<0.001	0.383	0.597
$\bar{\varphi}$	2.772	0.270	<0.001	2.243	3.302
σ_φ	5.789	1.073	<0.001	3.686	7.892
$\bar{\eta}$	1.350	0.121	<0.001	1.113	1.586
σ_η	1.361	0.243	<0.001	0.885	1.838
$\varrho_{r\varphi}$	0.041	0.023	0.084	−0.005	0.087
$\varrho_{r\eta}$	−0.357	0.065	<0.001	−0.486	−0.229
$\varrho_{\varphi\eta}$	0.197	0.031	<0.001	0.136	0.259
μ	0.594	0.067	<0.001	0.462	0.725
β_1	−3.237	0.337	<0.001	−3.898	−2.577
β_2	4.317	0.879	<0.001	2.595	6.040

Table 7: Descriptive Statistics of Posterior Mean of Risk Aversion Parameters under RDU

Variable	Obs.	Mean	Std. Dev.	Min	Max
		A. Individual-Level CRRA Coefficient			
Fechner Errors (FN)	413	0.700	0.300	−0.166	1.409
Contextual Utility (CU)	413	0.558	0.270	−0.344	1.254
Decision Field Theory (DFT)	413	0.702	0.625	−1.469	2.249
Entropy (EN)	413	0.723	0.321	−0.372	1.505
		B. Individual-Level Parameter φ			
Fechner Errors (FN)	413	2.115	2.382	0.042	17.920
Contextual Utility (CU)	413	2.277	2.621	0.044	23.211
Decision Field Theory (DFT)	413	2.459	2.597	0.072	14.379
Entropy (EN)	413	2.242	2.551	0.043	20.791
		C. Individual-Level Parameter η			
Fechner Errors (FN)	413	1.282	0.762	0.129	5.306
Contextual Utility (CU)	413	1.720	1.146	0.169	10.243
Decision Field Theory (DFT)	413	1.783	1.360	0.186	6.796
Entropy (EN)	413	1.362	0.949	0.113	7.628

Table 8: Summaries of The Statistical Goodness-of-Fit Measures under RDU

Stochastic Model	No. of Parameters	Log-likelihood	AIC	BIC	Rank
		A. In-Sample			
Fechner Error (FN)	10	−5,219	10,459	10,499	4
Contectual Utility (CU)	10	−5,170	10,361	10,402	2
Decision Field Theory (DFT)	10	−5,142	10,305	10,345	1
Entropy (EN)	12	−5,189	10,402	10,451	3
		B. Out-of-Sample			
Fechner Error (FN)	10	−2,374	4,768	4,800	4
Contectual Utility (CU)	10	−2,344	4,707	4,739	2
Decision Field Theory (DFT)	10	−2,334	4,689	4,721	1
Entropy (EN)	12	−2,352	4,727	4,766	3

Table 9: Vuong Tests Between the "Column" Specifications Against the "Row" Specifications under RDU in In-Sample Data
(The *p*-value is provided in the parenthesis)

	FN	**CU**	**DFT**
CU	−2.329 (0.020)		
DFT	−2.782 (0.005)	−1.444 (0.149)	
EN	−3.100 (0.002)	1.158 (0.247)	1.924 (0.054)

Table 10: Vuong Tests Between the "Column" Specifications Against the "Row" Specifications under RDU in Out-of-Sample Data
(The *p*-value is provided in the parenthesis)

	FN	**CU**	**DFT**
CU	−2.319 (0.020)		
DFT	−2.180 (<0.001)	−0.382 (0.703)	
EN	−2.774 (0.006)	0.992 (0.321)	1.041 (0.298)

Figure 1: Population Distributions of Relative Risk Aversion under EUT

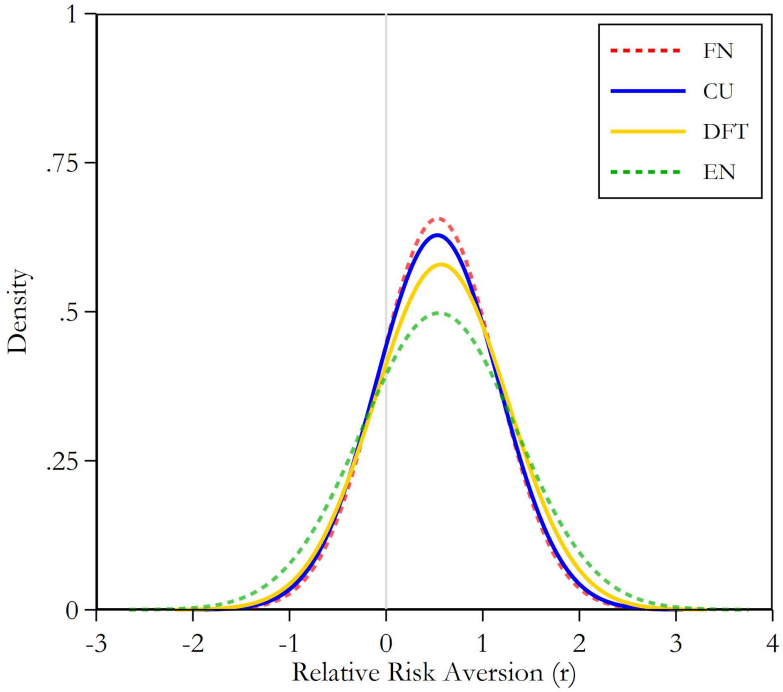


Figure 2: Distributions of Stochastic Models' Log-likelihood Values at Individual Level under EUT

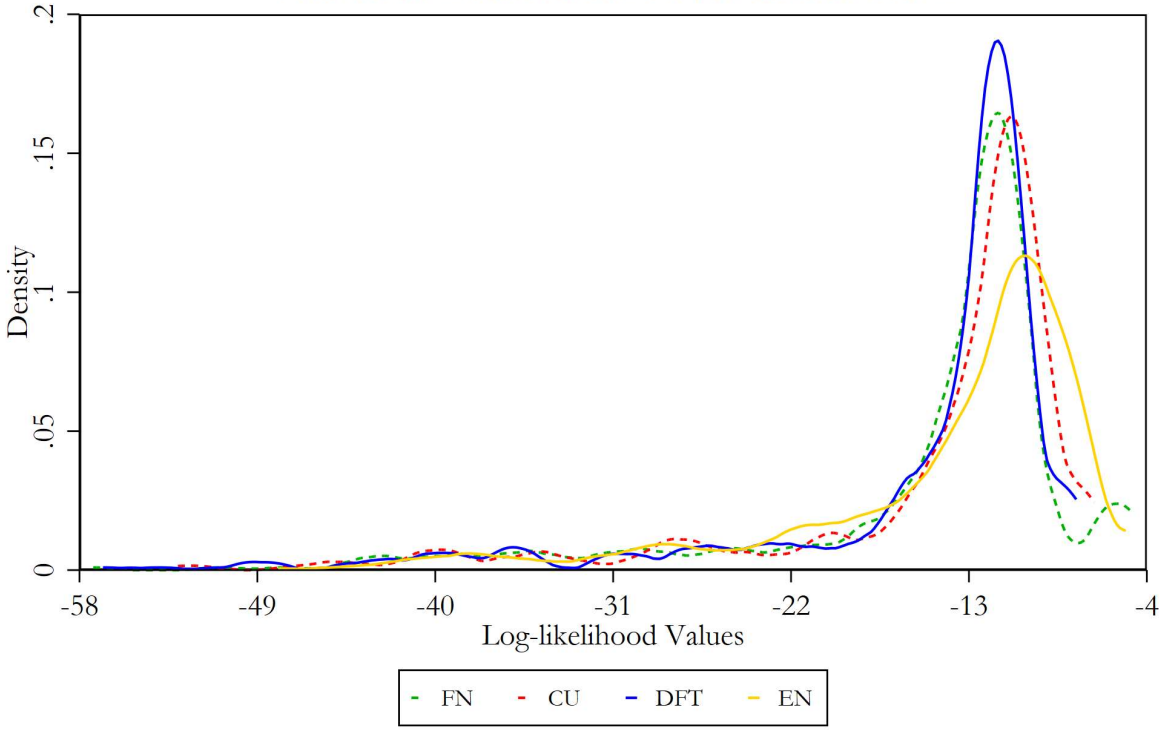


Figure 3: Classifying Subjects as CU, DFT, EN or FN Based on Individual-level Log-likelihood under EUT

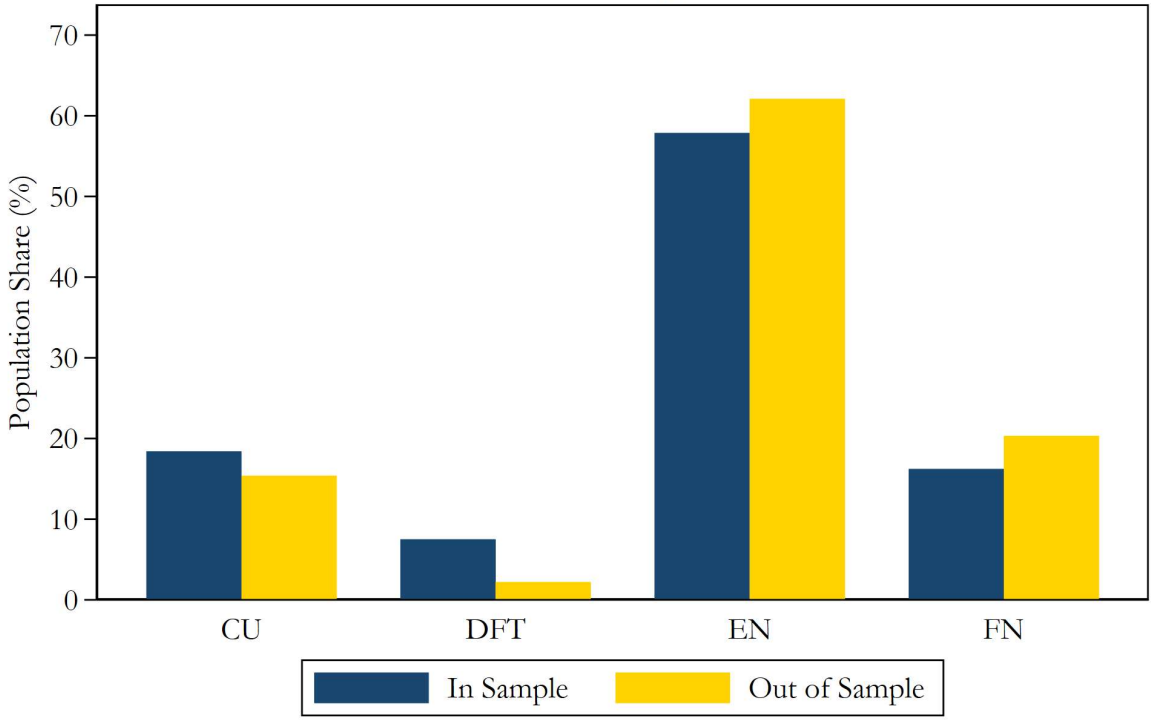


Figure 4: Mean Absoulte Errors of Choice Predictions under EUT

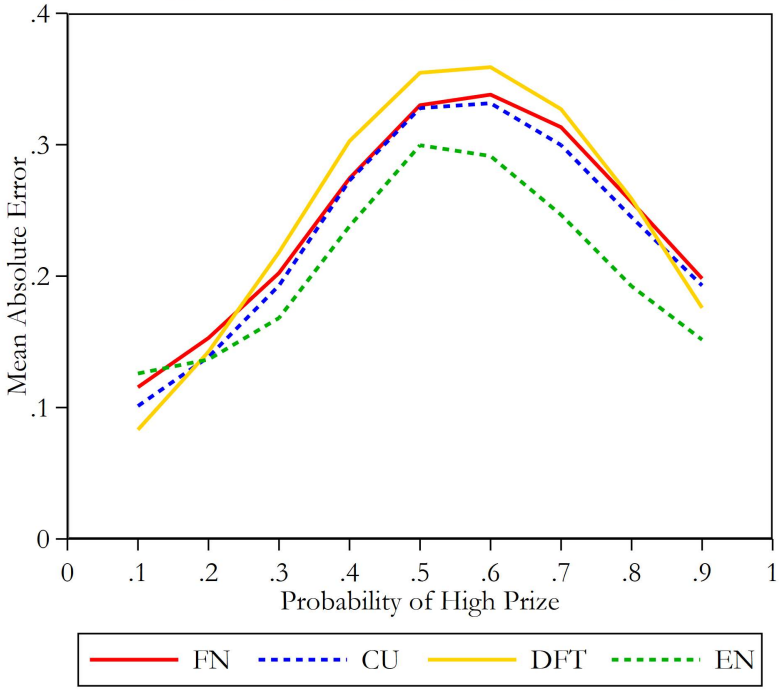
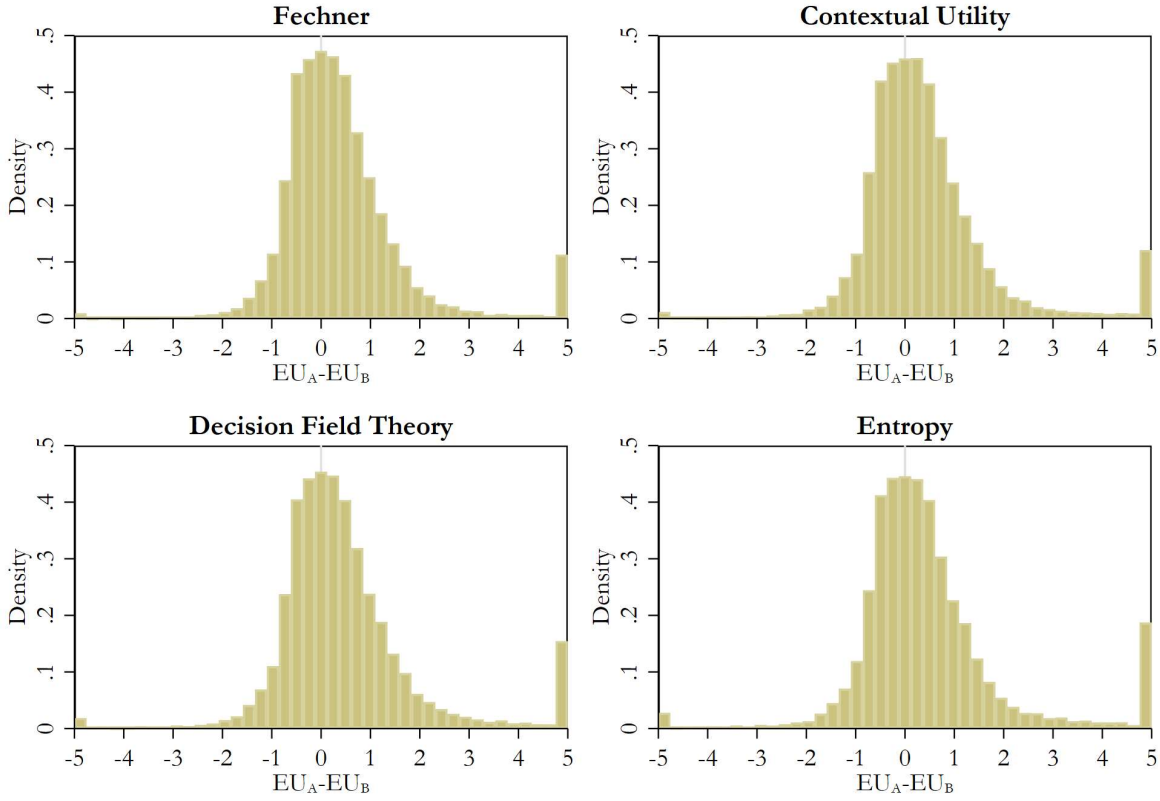


Figure 5: Distributions of Expected Utility (EU) Differences



Note: For a convenience presentation, we truncate the distributions by putting the values of 5 and higher in the same bin at the right tail and the values of -5 and lower in the same bin at the left tail. In each model, there are at least 2.6% of choice observations with EU difference which is equal to or greater than 5.

Figure 6: Standard Deviations of The Error Term under EUT

Lottery Pair: [A: DKK1000 and DKK875; B: DKK2000 and DKK75]

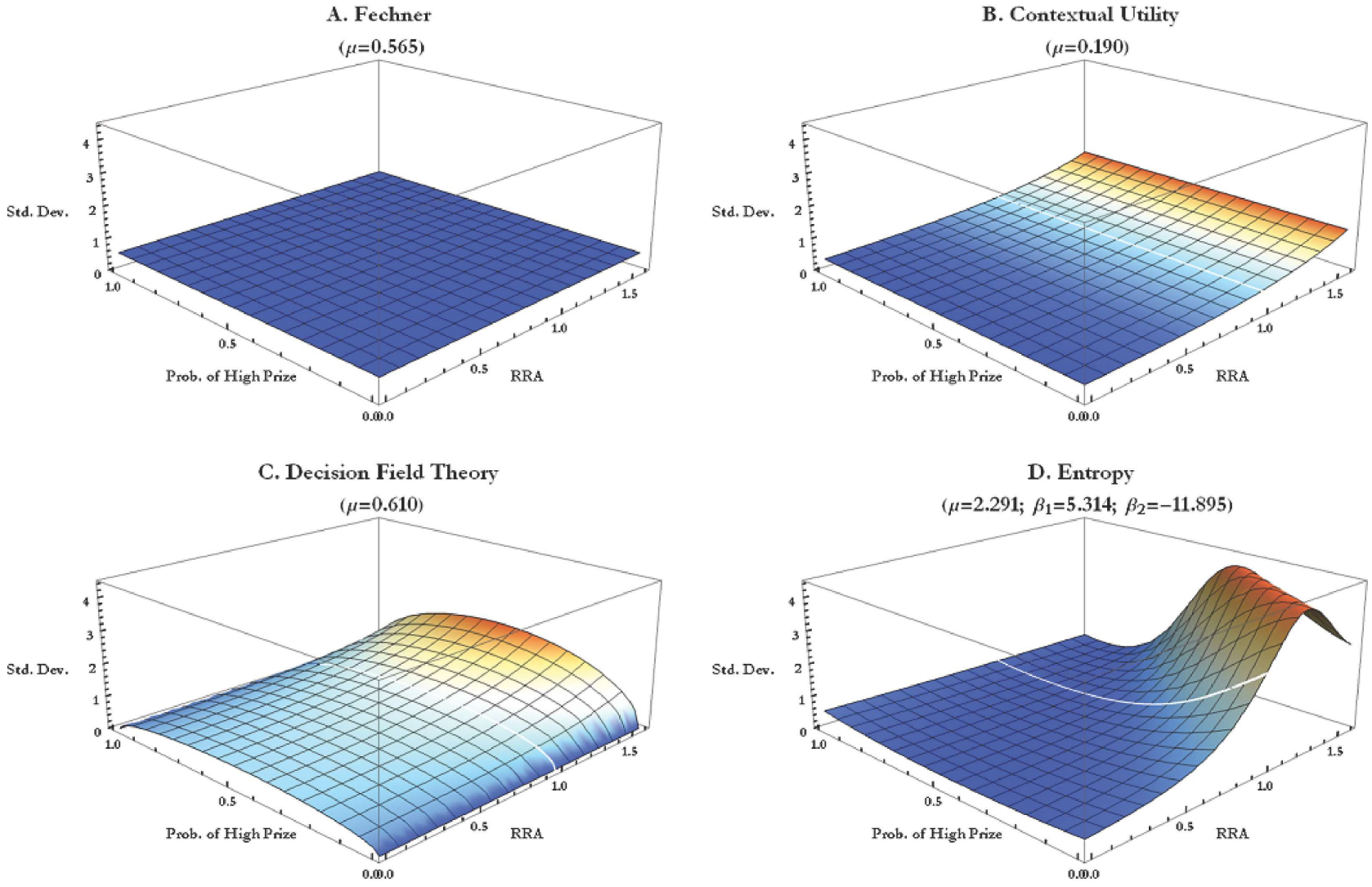


Figure 7: Population Distributions of Risk Aversion Parameters under RDU

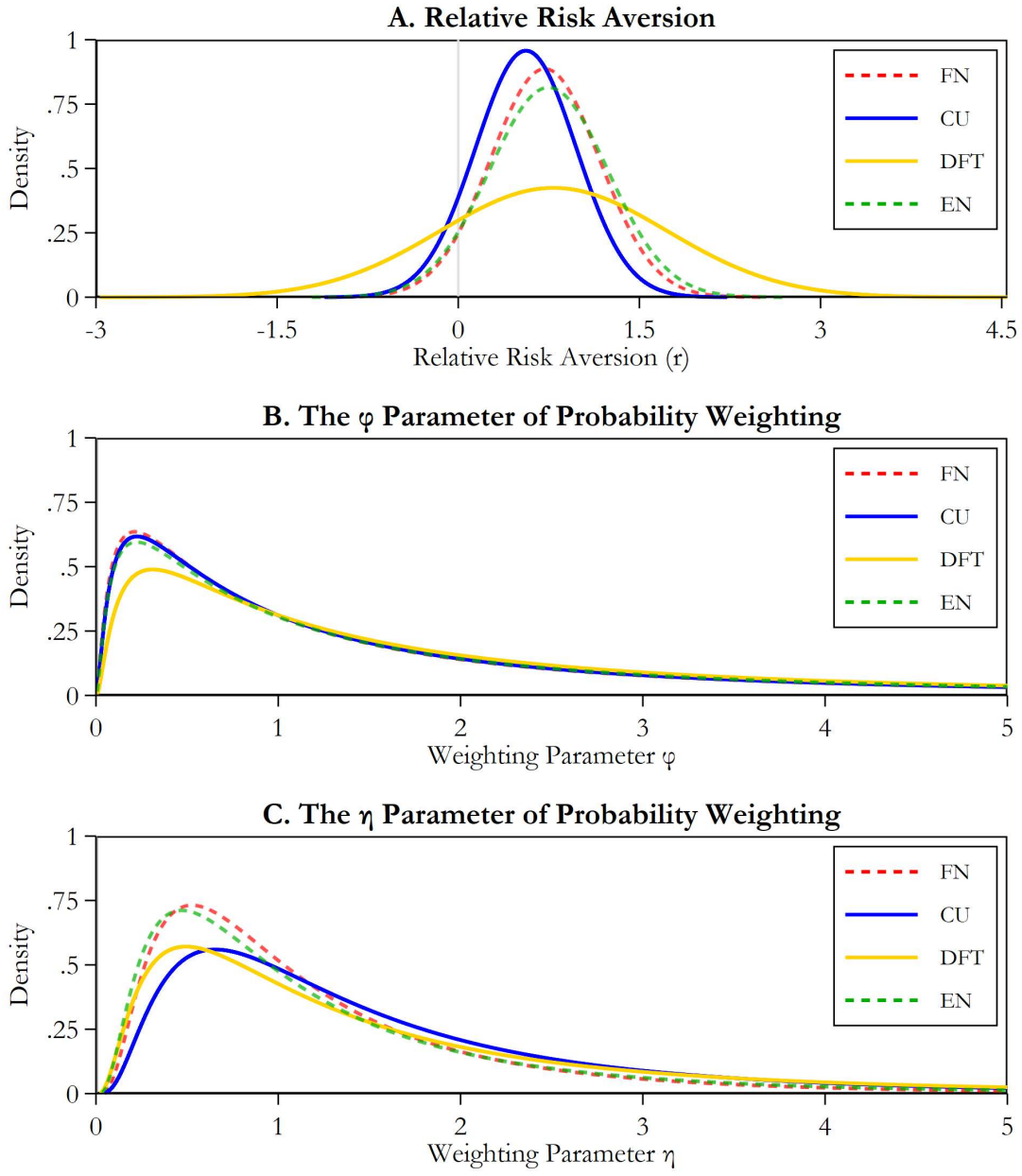


Figure 8: Shapes of Probability Weighting Function

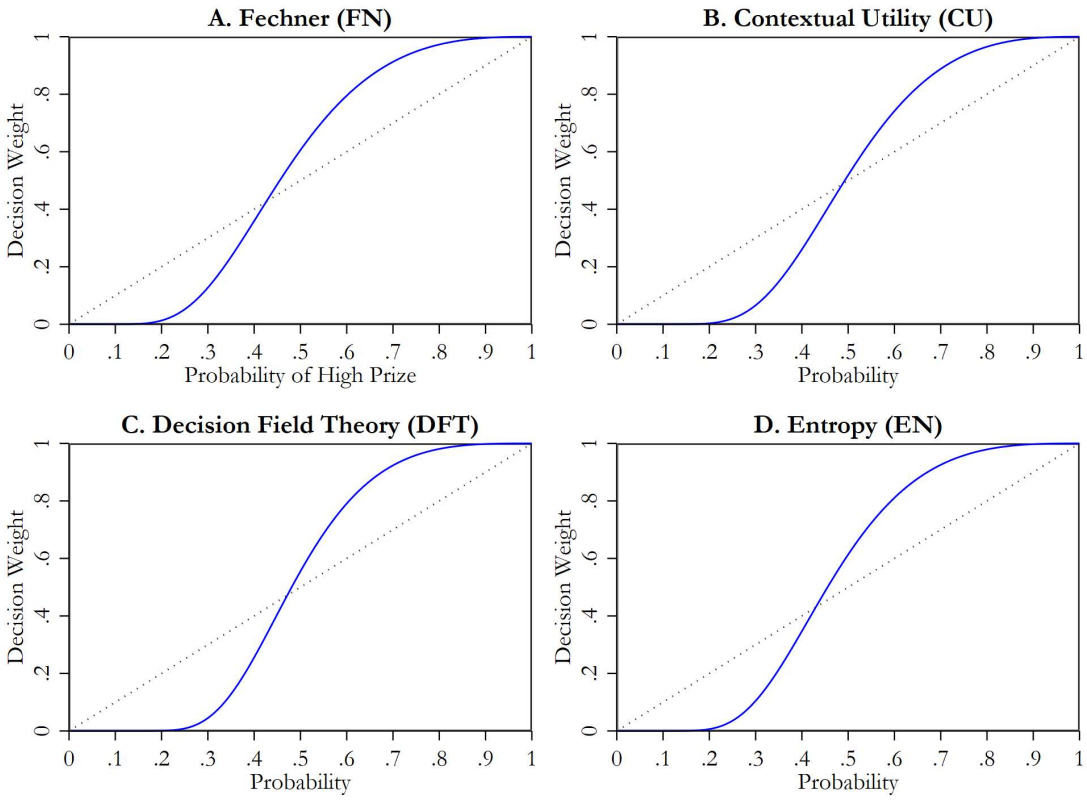


Figure 9: Distributions of Stochastic Models' Log-likelihood Values at Individual Level under RDU

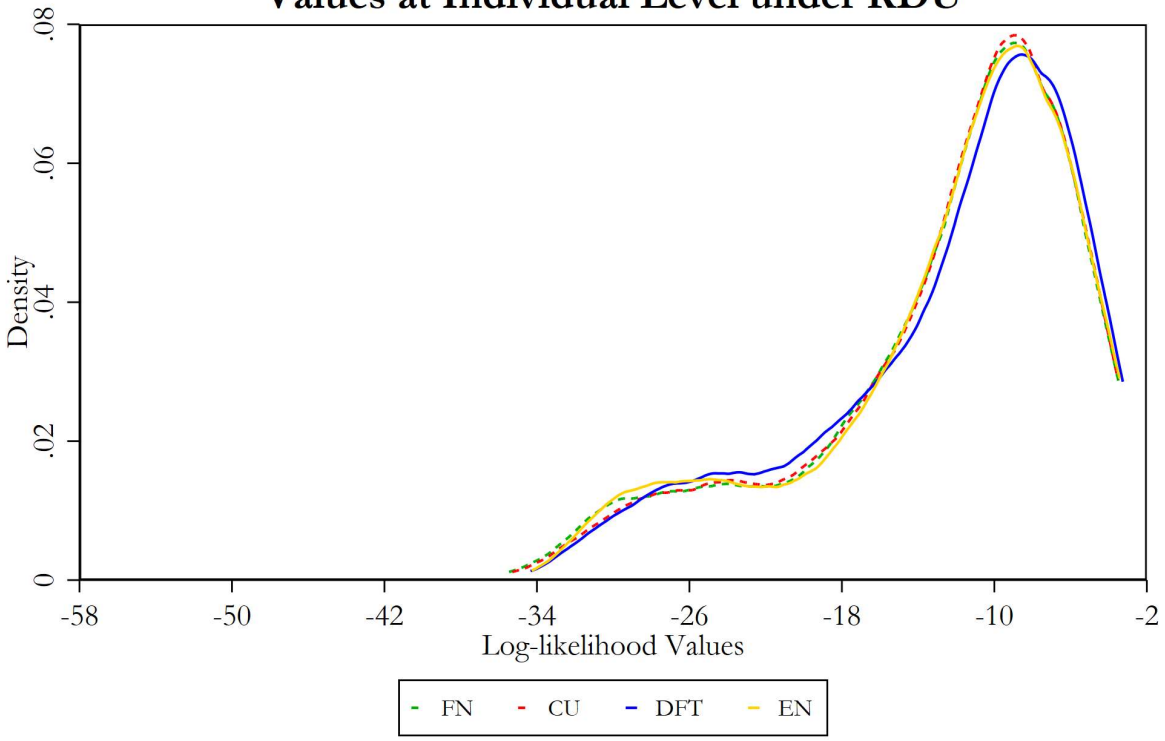


Figure 10: Classifying Subjects as CU, DFT, EN or FN Based on Individual-level Log-likelihood under RDU

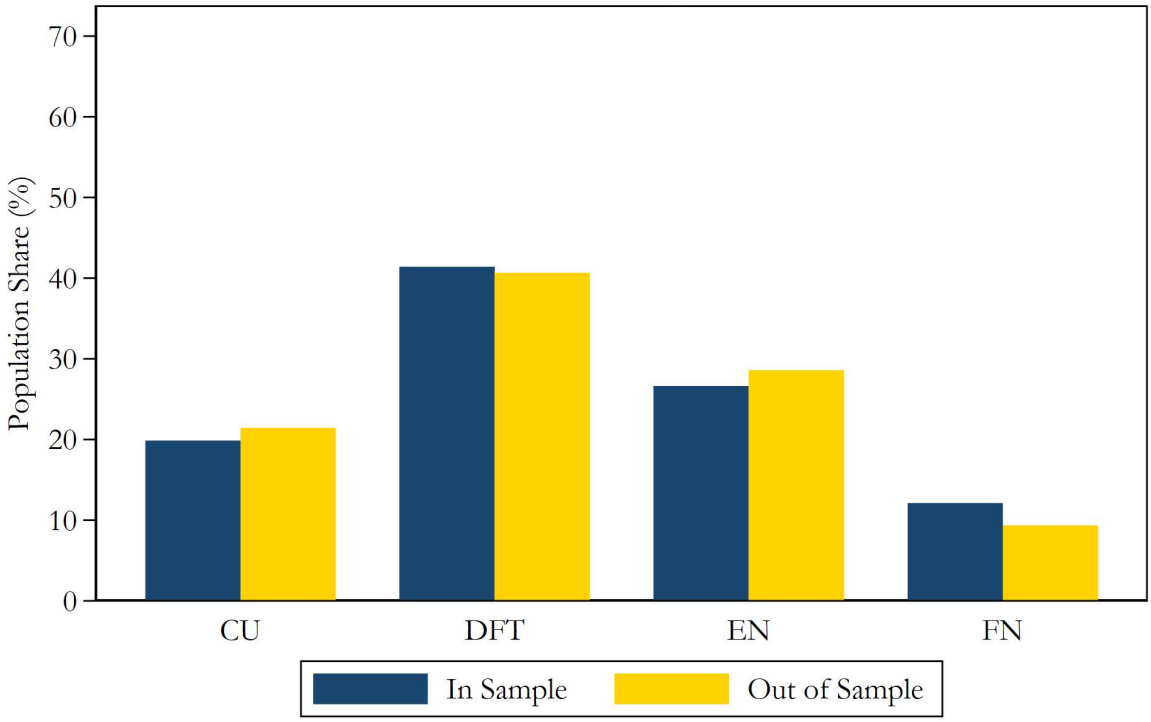


Figure 11: Mean Absoulte Errors of Choice Predictions under RDU

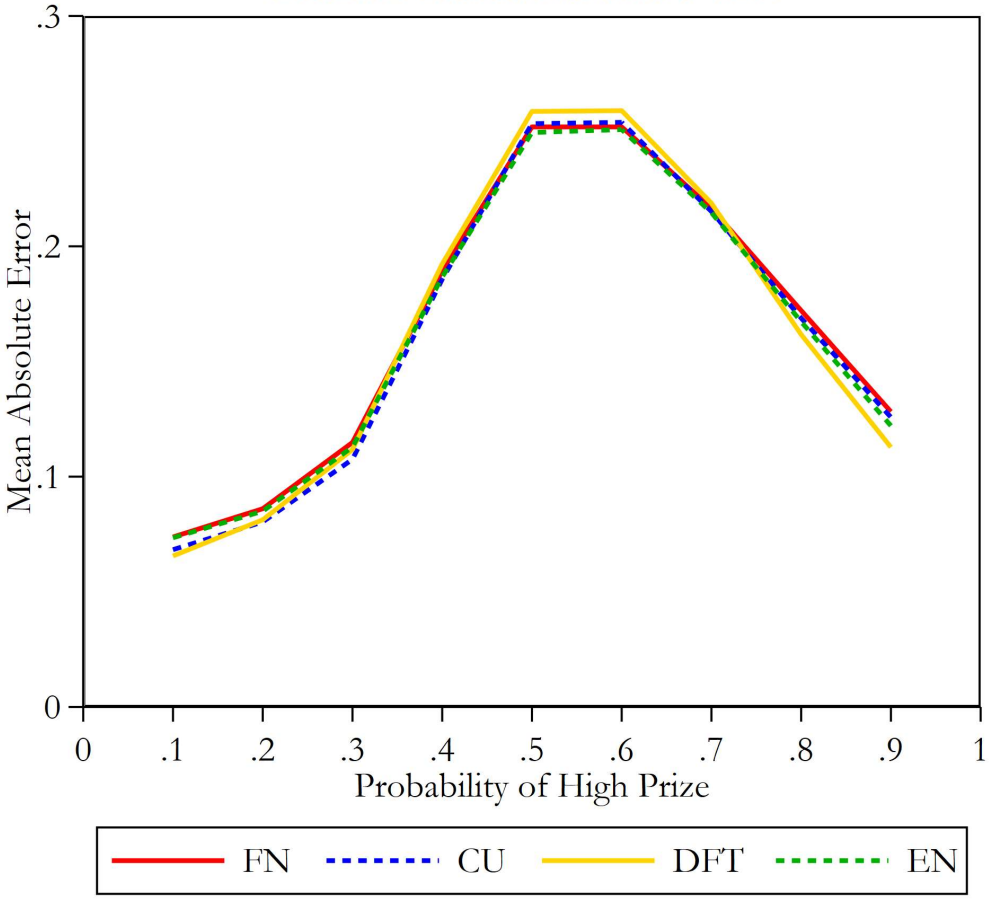


Figure 12: Distributions of Rank Dependent Utility (RDU) Differences

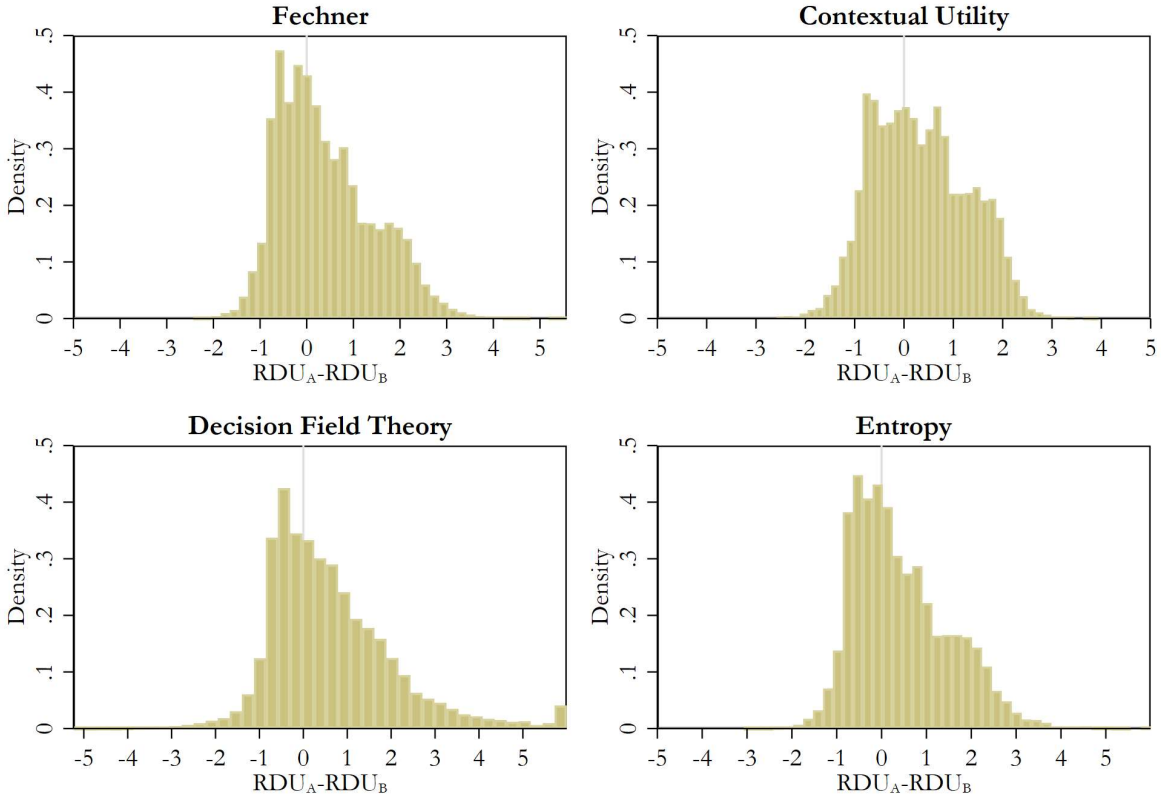
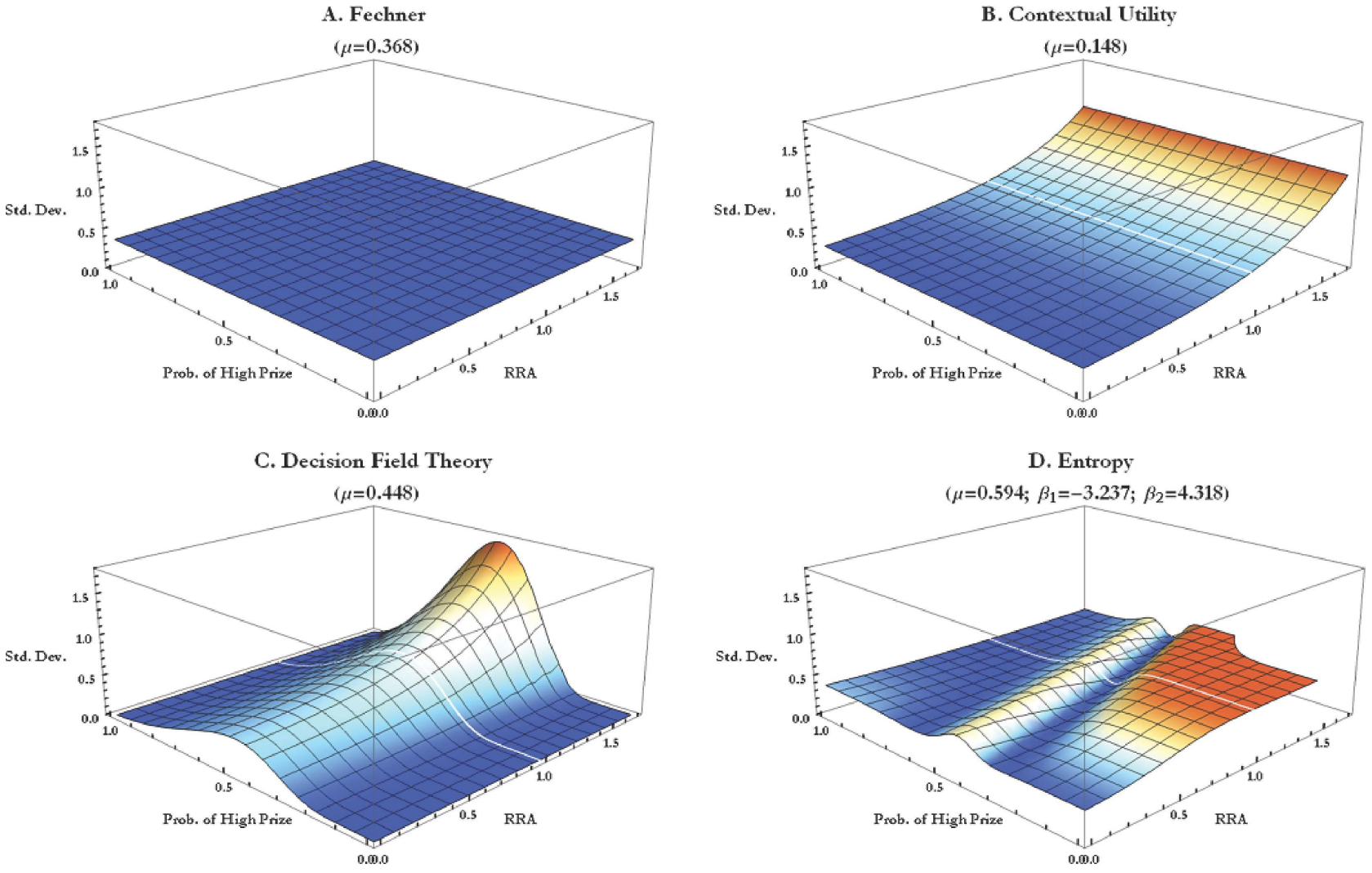


Figure 13: Standard Deviations of The Error Term under RDU

Lottery Pair: [A: DKK1000 and DKK875; B: DKK2000 and DKK75]



Appendix A: The Effect of Variation in Risk Parameter on The Complexity Measure of The Heteroskedastic Models

Figure A1: The Effect of Relative Risk Aversion
On The Complexity Measure of The Heteroskedastic Models
Lottery Pair: [A: DKK1000 and DKK875; B: DKK2000 and DKK75]

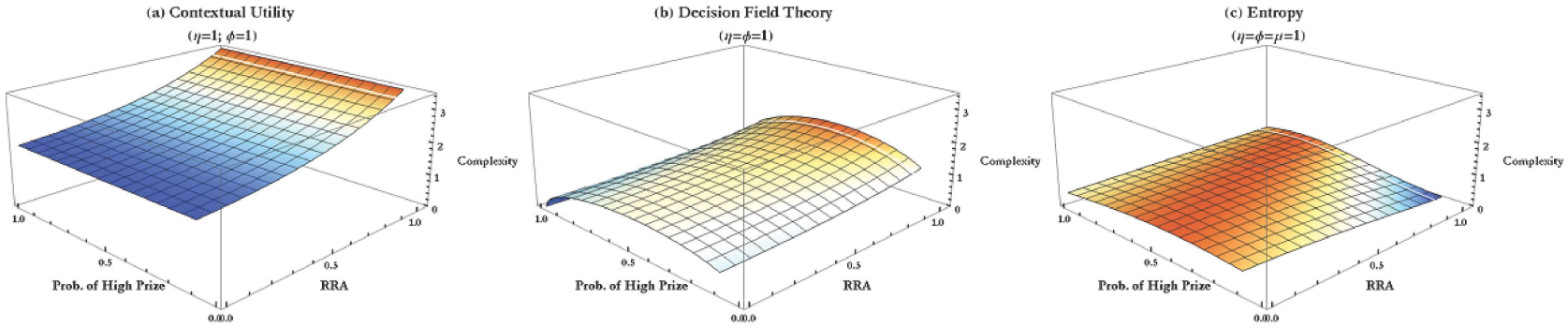
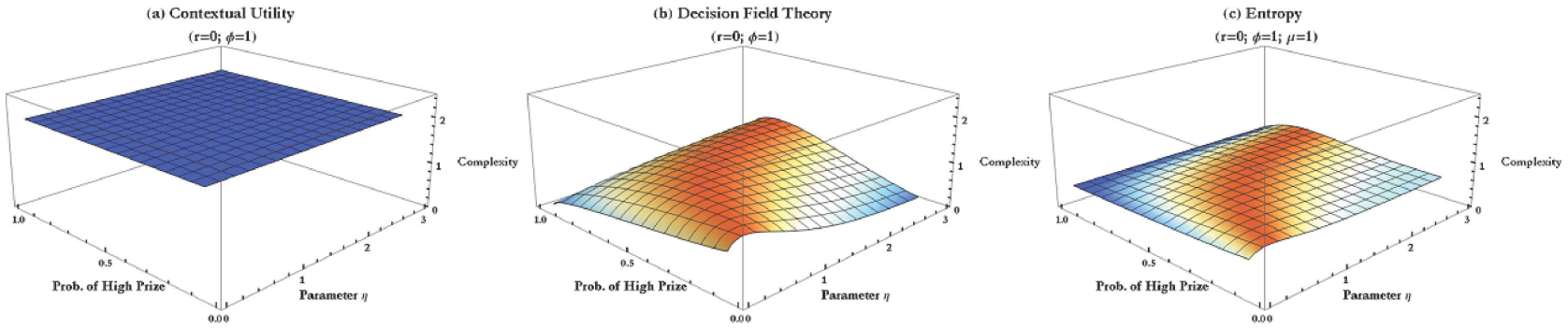


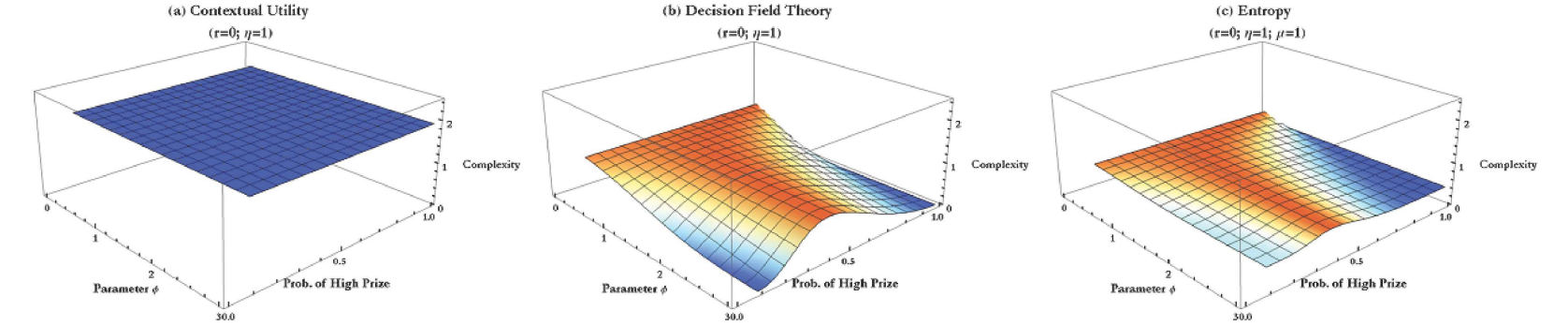
Figure A2: The Effect of Probability Weighting Function
On The Complexity Measure of The Heteroskedastic Models

Lottery Pair: [A: DKK1000 and DKK875; B: DKK2000 and DKK75]

A. Variation in Parameter η



B. Variation in Parameter φ



Appendix B: The Comparisons of The Estimated Population Means of Relative Risk Aversions under EUT

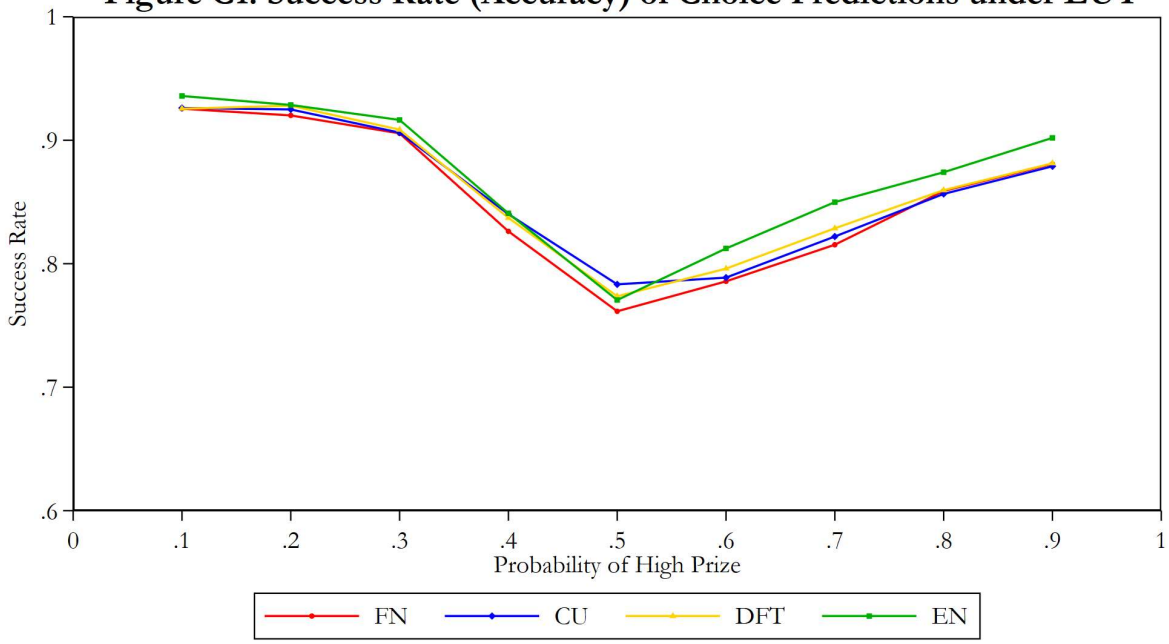
Table B1: Test Statistics of Difference in The Population Means of The CRRA Coefficient between Two Stochastic Models under EUT

(The *p*-value is provided in the parenthesis)

	FN	**CU**	**DFT**
CU	0.092 (0.927)		
DFT	-0.752 (0.452)	-0.824 (0.410)	
EN	-0.243 (0.808)	-0.318 (0.751)	0.423 (0.672)

Appendix C: Success Rates of The In-Sample Choice Prediction Under EUT

Figure C1: Success Rate (Accuracy) of Choice Predictions under EUT



Appendix D: Comparisons of The Estimated Population Means of Risk Aversion Parameters under RDU

Table D1: Test Statistics of Difference in The Population Means of The CRRA Coefficient between Two Stochastic Models under RDU

(The *p*-value is provided in the parenthesis)

	FN	**CU**	**DFT**
CU	5.234 (<0.001)		
DFT	−1.405 (0.160)	−4.549 (<0.001)	
EN	−0.916 (0.359)	−5.938 (<0.001)	0.806 (0.420)

Table D2: Test Statistics of Difference in The Population Means of The φ Parameter between Two Stochastic Models under RDU

(The *p*-value is provided in the parenthesis)

	FN	**CU**	**DFT**
CU	0.047 (0.963)		
DFT	1.383 (0.167)	1.346 (0.178)	
EN	0.654 (0.513)	0.612 (0.541)	−0.702 (0.483)

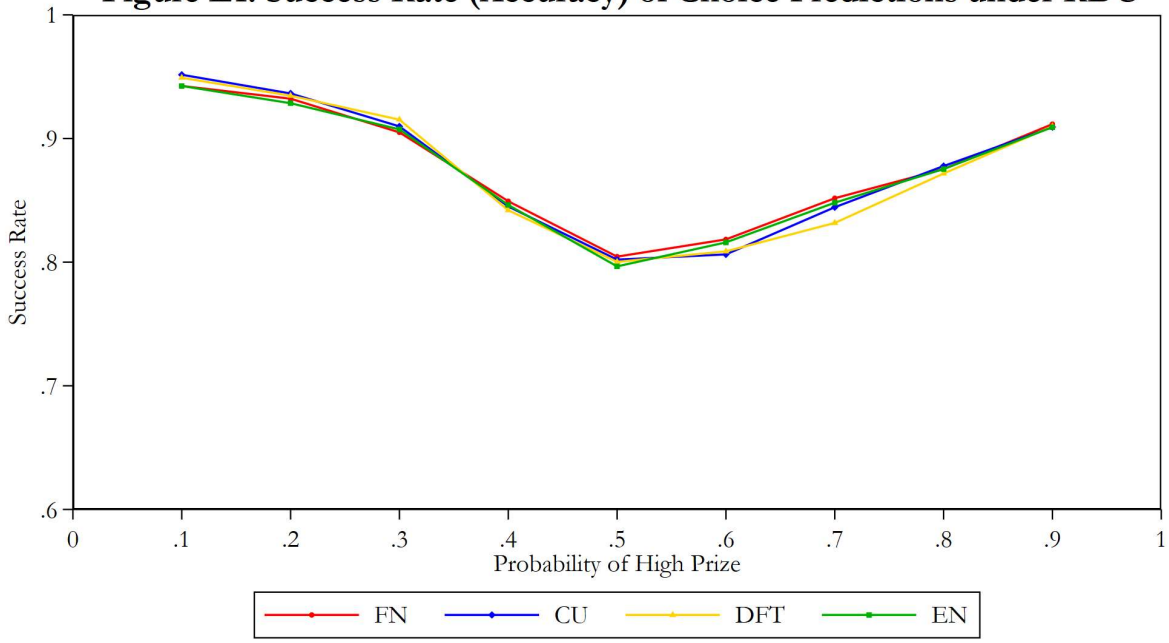
Table D3: Test Statistics of Difference in The Population Means of The η Parameter between Two Stochastic Models under RDU

(The *p*-value is provided in the parenthesis)

	FN	**CU**	**DFT**
CU	4.413 (<0.001)		
DFT	4.701 (<0.001)	0.727 (0.467)	
EN	0.806 (0.257)	−3.394 (0.001)	−3.785 (<0.001)

Appendix E: Success Rates of The In-Sample Choice Prediction Under RDU

Figure E1: Success Rate (Accuracy) of Choice Predictions under RDU



Chapter 2
An Econometric Analysis of Intertemporal Choice Heuristics

Abstract. We discuss two recent decision heuristics in time preferences. It has been claimed that heuristic models explain discounting behaviors better than their structural counterparts despite their radical departure from the standard economic theory of decision making. We analyze data from the incentivized field experiment to gain insights into their relative efficacy and compare the results to the structural discounting models. We contend that the claim neglects the importance of accounting for non-linear utility function in eliciting the individual discount rates. Our analytic observation also finds that the econometric modeling of the heuristic discounting models implies a stochastic structure akin to what one may induce by combining the structural with a finite mixture of two behavioral error stories. By allowing for the non-linear utility function or accounting for a mixture of error specifications, our in- and out-of-sample analyses refute the claim of the heuristics' superiority.

1. Introduction

Recent studies in psychology suggest that linear index models of intertemporal choice heuristics outperform structural discounting models in terms of out-of-sample predictive performance (Ericson, White, Laibson and Cohen [2015]; Wulff and van den Bos [2018]). These findings apply to all structural models widely used in economic applications, including non-constant discounting models, which have psychological origins or motivations. It remains to be seen, nevertheless, what economists may learn from these emerging findings. The available results are based on an online survey that did not have a well-defined sampling frame, and the survey respondent made hypothetical choices that had no economic consequences. The relevant analyses do not control for utility curvature, which is known to bias inferences about discount rates from structural discounting models, and the source of the seemingly superior performance by the heuristic models has not been identified yet.

We study the performance of heuristic discounting models relative to their structural counterpart, using data from an incentivized experiment with a representative sample of adult Denmark population. Our data source is the field experiment reported in Andersen, Harrison, Lau and Rutström (AHLR) [2014a], which recruited subjects from a general adult population in Denmark. The experiment administered both intertemporal choice and risky choice tasks, enabling us to apply a joint estimation approach to disentangle the effects of utility curvature from delay discounting. As with the psychological studies, we focus on two types of heuristic discounting models, namely the Difference-Ratio-Interest-Finance-Time (DRIFT) model by Read and Scholten [2013] and the Intertemporal Choice Heuristic (ITCH) model by Ericson, White, Laibson and Cohen [2015]. For comparisons, we consider a broad range of structural discounting models reviewed by AHLR. We use the Bayesian Information Criterion (BIC; Schwarz [1978]) and Vuong test of nonnested models (Vuong [1989]) to evaluate the models' in-sample performance. To

measure the predictive performance of each model, we randomly split the AHLR data into an estimation sample and a hold-out sample, and evaluate the model's log-likelihood in the latter sample at the parameter estimates obtained from the former. We repeat the same procedure multiple times so that the results are not influenced by a particular instance of the randomization.

We draw several conclusions. First, our results demonstrate the complementary roles played by the experimental design and the econometric method. Each discounting task in the AHLR experiment prompted a choice between a smaller, sooner (SS) payment and a larger, later (LL) payment. Perhaps the simplest structural modeling approach would be to model the choice probability by assuming linear utility; specifying an index function that captures the present value difference between the SS and LL payments; and plugging the index function into a probit or logit link function. With this simple approach and assuming Fechner error, we find that the two heuristic models outperform all the structural models under consideration: That is, despite the use of the data from the AHLR experiment, which used monetary rewards for the subject's choices to encourage effort, we are able to reproduce main conclusions based on hypothetical responses. However, as we summarize shortly, this superior performance by the heuristic models disappear once we apply a more modern approach to the structural analyses of discounting behavior.

Second, our results stress the importance of choosing appropriate stochastic choice models in the structural estimation of discounting behavior. Besides the Fechner error, the structural discounting models are commonly combined with Luce errors to account for behavioral noises in decision making. Our key insight is the use of both level and relative differences in the attribute comparisons make the heuristic models resemble two different decision rules under the two standard stochastic specifications. We find that the estimated coefficient of level difference of money attribute in both heuristics is not statistically significant, indicating that the Luce errors better represents the decision rule and stochastic behavior of decision-makers under the structural models.

Indeed, our results choose the structural models over their heuristic counterparts as we switch from Fechner to Luce errors while maintaining the risk neutrality assumption. The linear index model of the heuristic specifications also implies a stochastic structure that is akin to what one may induce by combining the structural discounting models with a finite mixture of Fechner and Luce errors. In accordance with this analytic observation, we empirically find a relatively weak performance of the heuristic discounting models once we adopt the mixture error specifications, regardless of whether we control for utility curvature or not.

Third, our results stress the importance of accounting for utility curvature.[1] We find that the seemingly superior performance of the heuristic discounting models vanish when we control for the effects of utility curvature in the estimation of the structural discounting models. This finding is obtained regardless of behavioral error specifications used in structural estimation. Since Andersen, Harrison, Lau and Rutström [2008] demonstrated empirically that neglecting utility curvature results in substantially biased estimates of intertemporal discount rates, adjusting for utility curvature in the analyses of discounting behavior has become a widely accepted practice in experimental economics, inspiring the development of many experimental designs (Cheung [2015]). We contribute to this literature by showing that neglecting utility curvature can completely reverse inferences about the relative performance of heuristic and structural discounting models.

Finally, our results are robust to any pair of heuristic and structural discounting models. This further stresses that the seemingly superior performance by the heuristic discounting models has nothing to do with specific ways in which choice heuristics are better at modeling particular aspects of intertemporal choice behavior than structural discounting models. It is more an artefact of their oversimplified econometrics modeling. Our results complement Stahl [2018], who finds that plain

[1] The term "utility" that repeatedly quoted in Ericson, White, Laibson and Cohen [2015] and Wulff and van den Bos [2018] is merely a metaphor than a fact. Their estimations of the structural models instead assume risk neutrality due to the absence of the binary risky choices from their experiment.

EUT outperforms the toolbox model of heuristics for choice under risk. Our results are also in line with Glöckner and Betsch [2008] and Rieger and Wang [2008], who found that the typical claim on choice heuristics as better descriptive models of individual decision-making is often built upon implicit assumptions in their econometric modeling and empirical evaluation.

2. Data

We use data from an artefactual field experiment conducted in Denmark in 2009 with a representative sample of Danish adult population. The experiment aimed to study the "magnitude effect" on discounting (Andersen, Harrison, Lau and Rutström [2013]), alternative specifications of discount functions (Andersen, Harrison, Lau and Rutström [2014a]), intertemporal risk aversion (Andersen, Harrison, Lau and Rutström [2018]), and non-constant discounting, temporal stability and dynamic consistency in the discounting behavior (Harrison, Lau and Yoo [2020]). The experiments was attended by 413 subjects recruited from the adult Danish population aged between 18 and 75 years.

Each subject in the experiment was confronted with 40 risk attitude choices and 40 discounting choices with a 10% chance of being paid for one choice in each set of 40 choices. The risk attitude choices preceded the discounting choices in one treatment and *vice versa* in another treatment. The average nominal payment was 242 kroner and 201 kroner, respectively, for the risk attitude choices and the discounting choices, for a combined average of 443 kroner. With the exchange rate of 5 kroner per U.S. dollar, the combined average income from these decision tasks was equivalent to $91 at the time of the experiment. The subjects were also paid a 300 kroner or 500 kroner fixed attendance fee, in addition to earnings from subsequent tasks.[2]

[2] The additional tasks earned subjects an average of 659 kroner, so total earnings from choices made in the session averaged 1102 kroner, or roughly $221, in addition to the fixed fee of $60 or $100.

A. Experiments to Elicit Discounting Functions

Individual discount rates are measured from data in which subjects responded to a series of choices over two outcomes that are paid at different dates. For example, a subject is asked to choose between 1000 kroner in 30 days and 1100 kroner in 90 days. If the subject is risk neutral and chooses the sooner option, we can infer the discount rate is above 10% for a time delay between the two options of 60 days. If the same subject selects the later option instead, we can deduce that the discount rate is lower than 10%. We can identify the subject's discount rate by varying the monetary payment in the later option, conditional on knowing the utility of those amounts to the subject. One can also identify the discounting function by varying the time delay between the sooner and later options and, of course, by varying the delay to the sooner option. This method has been extensively employed, among others, by Coller and Williams [1999], Harrison, Lau and Williams [2002], Eckel, Johnson and Montmarquette [2005], Andersen, Harrison, Lau and Rutström [2008][2013][2014a][2018] and Dohmen, Falk, Huffman and Sunde [2010].

Time delays between the sooner and later options vary from 2 weeks to 1 year. The experiment presented each subject with choices over four different time delays in ascending or descending order, and those time delays are randomly drawn from a set of thirteen intervals (2 weeks, and 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 and 12 months). The delay to the sooner option is also varied on a between-subjects basis, in which approximately half of the sample had decision tasks with no delay to the sooner option, and the other half had a 30-day delay. Similarly, the experiment varied the provision of implied annual interest rates for each choice on a between-subjects basis. Finally, the experiment employed two principal amounts on a between-subjects basis (1500 and 3000 kroner) to assess the significance of magnitude effects on elicited discount rates. These four treatments, the order of presentation of the time delay, the delay to the sooner option, information

on implied interest rates, and the level of the principal, give a 2×2×2×2 design. Each subject was assigned at random to one of these sixteen combinations.

The experiment presented the subjects with 40 binary choices, in four sets of 10 with the same time delay between the sooner and later option. The annual interest rate varied between 5% and 50%, in increments of 5%, on the principal of 1500 kroner or 3000 kroner. We randomly selected one decision task using numbered dice and paid the subjects their preferred smaller-sooner (SS) or larger-later (LL) option. The large incentives and budget constraints precluded us from paying all subjects, so each subject was given a 10% chance to receive the payment.[3]

B. Experiments to Elicit Utility Functions

Utility functions were evaluated by asking subjects to make a series of choices over two risky lotteries. For example, lottery A offers a 50:50 chance of receiving 1000 kroner or 875 kroner today, and lottery B has a 50:50 chance of receiving 2000 kroner or 75 kroner today. The subject picks A or B. We used the procedures of Hey and Orme [1994] and presented each binary choice to the subject as a "pie chart" with visual displays of each outcome's probability. We presented the subjects with 40 choices in four sets of 10 with the same prize combinations. The prize sets employed were: [A1: 2000 and 1600; B1: 3850 and 100], [A2: 2250 and 1000; B2: 4500 and 50], [A3: 1125 and 750; B3: 2000 and 250] and [A4: 1000 and 875; B4: 2000 and 75]. We randomized the order of the four prize sets for each subject, with probabilities of prizes varying within each set.[4] We refer to the first two prize sets as the "high stakes" lotteries compared to the last two sets. All subjects were confronted with the same set of decision tasks.

[3] The immediate payments were cash-based and paid at the end of the experiment, while the future payments were wire-transferred to the subjects' bank account, which is a common practice in Denmark.

[4] Each prize set had 10 choices presented one at a time and arranged in an orderly manner, with the probability of the high prize starting at 0.1 and increasing by 0.1 until the last choice was between two degenerated lotteries.

The experiment asked each subject to respond to all 40 risk aversion tasks and then randomly decided which one to play out using numbered dice for the actual payment for that subject, on top of a fixed show-up fee. Studies examining lottery choice experiments of this kind commonly find subjects to be averse to risk, with considerable heterogeneity in attitudes toward risk across individuals (e.g., Harrison, Lau and Rutström [2007]).

3. Theory

We first write out the standard discounting functions with the most parsimonious specification, namely the exponential discounting and simple hyperbolic discounting. We then continue with the two decision heuristic models: the Inter-temporal Choice Heuristic (Ericson et al. [2015]) and the Difference-Ratio-Interest-Finance-Time (Read et al. [2012]).

A. The Structural Discounting Specifications

In choosing between a sooner but smaller level of income and a larger but later income, the discounted utility theory assumes that, for some utility function $U(\cdot)$, there exists a discount factor $D(t)$ that equates the utility of the time-dated incomes. Economics and psychology literature have offered varying specifications of discounting functions, and we consider two of them.

The first specification is the exponential (EXP) discount function (Samuelson [1937]). With δ denotes the *baseline* discount rate, the discount function is defined as

$$D^{EXP}(t) = 1/(1 + \delta)^t, \tag{1}$$

where $\delta > 0$. We denote time delay t in years (e.g., $t = 0.5$ for a 6-month horizon) and specify δ on an annualized basis. Given the discount function, the discount rate δ in this model is constant over

time.[5] Additionally, it is apparent from Jensen's inequality that the implied discount rate δ lowers when one allows for concave utility functions.

Contradicting the constant discounting feature of EXP, experimentally elicited discount rates frequently exhibit a hyperbolically declining pattern over time, implying a diminishing impatience (see, for example, Thaler [1981], Benzion, Rapoport and Yagil [1989]). The most common functional form that represents the pattern is the simple hyperbolic (SH) discounting proposed by Herrnstein [1981], Ainslie [1992], and Mazur [1984]. It is defined as

$$D^{SH}(t) = 1 / (1 + \varkappa \times t), \tag{2}$$

where $\varkappa > 0$. The discount rate for this specification is $d^{SH}(t) = (1 + \varkappa \times t)^{(1/t)} - 1$, which is hyperbolically declining with time delay t when linear utility function is assumed. However, as one relaxes the assumption of risk neutrality, the quantitative magnitude of the decline is much smaller if U(Y) is concave in Y. Furthermore, the level of discount rates is also theoretically lower in the latter case.

Each discounting task in the experiment had the subjects to choose between option A that pays Y_t in period t and option B that pays $Y_{t+\tau}$ in period $t+\tau$, where τ is the time horizon with a positive value ($\tau > 0$). Given a discounting function D(t) and an atemporal utility function U(Y), the discounted utility of each option is specified as

$$PV_A = D(t) \times U(Y_t + \omega) + D(t + \tau) \times U(\omega) \tag{3}$$

$$PV_B = D(t) \times U(\omega) + D(t + \tau) \times U(Y_{t+\tau} + \omega) \tag{4}$$

where ω is a measure of background consumption.[6]

---

[5] Considering the required sign constraint and the highest annual interest rate in the discounting tasks, we set $\delta \in (0, 0.7)$. Recall from Section 2, the maximum annual interest rate offered in the experiment is 50%, which is equivalent to an effective interest rate of 65% *per annum* with daily compounding.

[6] Following Andersen, Harrison, Lau and Rutström [2008][2014], we exogenously set the background consumption parameter ω to be 130 kroner, which is the per capita consumption of private non-durable goods on an average daily basis at the time of the experiments.

The discounted utility predicts that the observed choice is option B when it gives the larger present value than option A and *vice versa*. One can express this decision rule as either $PV_B - PV_A > 0$ or $PV_B/PV_A > 1$. While the theory itself is agnostic about which path to follow, the choice of decision rule has an implication for stochastic modeling. Using the level difference of present values leads to the adoption of the Fechner model. Contrarily, if the ratio of present values is considered, we end up with the Luce specifications. Before we progress further to the stochastic specifications, we will discuss how the heuristic models adopt such computational rules in their specifications.

B. The Difference-Ratio-Interest-Finance-Time Heuristic

Although their decision rules are not based on present values, heuristic models generally implement the same numerical judgment for their intra-dimensional comparisons, *i.e.*, whether the two values in each dimension are compared based on their level or relative difference. The trade-off model of Scholten and Read [2010], for example, calculates the level difference of two values in both money and time dimensions. The stochastic difference model by González-Vallejo [2012], on the other hand, computes their proportional differences. Other heuristic models take a more flexible approach by combining the two numerical comparisons in their specifications: using the level difference to compare values in one dimension and proportional difference in the other dimension.[7]

Even further, Read et al. [2012] assume that the individuals may simultaneously compute both level and proportional differences of the two values in a choice attribute and integrate them into an overall judgment. Specifically, their Difference-Ratio-Interest-Finance-Time Heuristic

[7] Among the example is the interval discounting model by Scholten and Read [2006], which calculates the relative difference in monetary amounts and level differences in delivery times. The decision framework by Adriani and Sonderegger [2020] defines an interval where the two values are perceived as similar. In the money dimension, this similarity interval is based on the proportional comparison of the two prizes, while the similarity interval in the time dimension is based on their absolute difference. In the context of decision making under risk, Stahl [2018] similarly integrates both quantitative rules in his toolbox specification. The model, adapted from the priority heuristic of Brandstätter et al. [2006], computes the relative level difference in monetary prizes and the level difference in the probability of occurrence.

(DRIFT) model evaluates the money attribute based on the level difference $\Delta Y^{abs} = Y_{t+\tau} - Y_t$ as well as the relative difference $\Delta Y^{rel} = (Y_{t+\tau} - Y_t)/Y_t$. The following index gives the choice evaluation:[8]

$$W_{DRIFT} = \alpha_1 \Delta Y^{abs} + \alpha_2 \Delta Y^{rel} + \alpha_3 \tau + \alpha_4 I. \quad (5)$$

The variable I represents the annually compounded interest, given by $I = (Y_{t+\tau}/Y_t)^{(12/\tau)} - 1$.[9] Obviously, it is equal to ΔY^{rel} only when the time horizon is twelve months.

Read et al. [2012] assume that $\alpha_1, \alpha_2, \alpha_4 \geq 0$ and $\alpha_3 \leq 0$; thus, a larger W_{DRIFT} displays a greater tendency to choose the large-later outcome $Y_{t+\tau}$. Despite its inability to elicit individual discount rates directly, the model predicts the delay effect, *i.e.*, preference over the sooner choice is generally weaker as the time horizon gets longer, consistent with the continuously declining discount rate behavior. When the annual interest rate is fixed while the time horizon is extended, the monetary reward $Y_{t+\tau}$ grows faster than the time horizon τ. So do ΔY^{abs} and ΔY^{rel} relative to τ. The net effect is thus a higher W_{DRIFT}.

C. The Inter-temporal Choice Heuristic

Only calculating the level difference of time values prevents the DRIFT model from accounting for the effect of the introduction of a front-end delay, *i.e.*, a more patient behavior when

[8] Ericson et al. [2015] modified DRIFT by adding a constant term to its original formulation. However, Wulff and Van den Boss [2017] pointed out that the constant term improves the model's performance despite lacking a behavioral foundation. We, therefore, adopt the original form to measure the pure performance of the main variables in capturing the discounting behavior. The original specification of DRIFT has a dummy variable that identifies whether the decision task is an investment or consumption-framed task. We exclude the dummy variable provided that the subjects in our experiment choose between two consumption opportunities.

[9] As implied by its name and function, the variable annually compounded interest I is not an intra-dimensional comparison. Read et al. [2012] used it to capture the behavior influenced by the interest rate framing. In this frame, the task had the subjects choose between a certain amount of money now or an extra amount presented in a compounded interest rate. For example, they choose between 1500 kroner now or an additional 25% effective annual interest rate (compounded annually) over six months. The subjects thus must calculate by themselves how much the extra money they will get from the given interest rate. Note that in our data, the interest rate is additional information provided to the subjects. The main information is the nominal value of the monetary rewards and their time deliveries.

a delay is introduced to the immediate reward while keeping the time horizon τ fixed (Read and Roelofsma [2003] and Thaler [1981]). To capture that effect, Ericson et al. [2015] modified the DRIFT model by adding an explanatory variable in terms of proportional time difference $\Delta T^{nor} = \tau/((2t + \tau)/2)$.[10] The relative money difference is adjusted to its normalized form, given by $\Delta Y^{nor} = (Y_{t+\tau} - Y_t)/((Y_{t+\tau} + Y_t)/2)$. Although it looks similar, this normalized ratio increases with time horizon at a slower pace relative to the proportional difference ΔY^{rel} in DRIFT for a given interest rate. Finally, the annualized interest rate variable is removed. The resulting model is thus integrating the level and relative difference measures both for money and time dimensions with the following specification:

$$W_{ITCH} = \alpha_1 \Delta Y^{abs} + \alpha_2 \Delta Y^{nor} + \alpha_3 \tau + \alpha_4 \Delta T^{nor}. \quad (6)$$

For the same reason as DRIFT, we also remove the constant term from the original specification of the ITCH model. Ericson et al. [2015] predict that $\alpha_1, \alpha_2 \geq 0$ and $\alpha_3, \alpha_4 \leq 0$, so that a larger W_{ITCH} represents a greater propensity to select the remote choice. ITCH also captures the diminishing impatience behavior. As the time horizon is prolonged while keeping the annual interest rate fixed, ΔY^{abs} and ΔY^{nor} increase faster than τ and ΔT^{nor}. The front-end delay effect is, of course, captured by ΔT^{nor}. Postponing Y_t and $Y_{t+\tau}$ by t' period while maintaining the values of both outcomes only lowers ΔT^{nor}. As the other variables are unchanged, W_{ITCH} increases.

5. Econometrics

We first outline the estimation for the heuristic models and illustrate how the level and relative differences in their attribute comparisons resemble two different decision rules under the standard stochastic specifications, namely Fechner and Luce errors. We then present the application

[10] Similarly, Leland [2002] use the similarity heuristic of Rubinstein [1988] to explain the front-end delay effect by assuming that the decision makers measure the ratio similarity of the time dimension, rather than the similarity based on the level difference.

of each of the stochastic specifications in the estimation of the structural discounting models. To be comparable with the hybrid specification of heuristic models that combines the two numerical comparisons, we consider an extension for the structural models where the two decision rules are both possible to be used in the decision making. We specify such extension as the mixture of the two stochastic specifications described under the finite mixture model. In contrast to the heuristic models that assume linear transformation of the monetary reward, literature has emphasized the need to account for the non-linear utility function to correct the upwardly biased discount rates. Therefore, we finally also write out the joint estimation of the structural discounting models with the Expected Utility Theory (EUT) and Rank-Dependent Utility (RDU) Theory.

A. Elicitation of the Decision Heuristic Models

We use an index function model for binary response to estimate the coefficients in the two heuristic models. Let h be DRIFT or ITCH specification. The latent propensity index is given by $y = \mathbf{I}[W_h + \varepsilon > 0]$, where y denotes a binary indicator of whether the observed choice is option B ($y = 1$) or option A ($y = 0$), and W_h is given by equation (5) and (6), respectively, for DRIFT and ITCH. Assume that ε is logistically distributed with a standard deviation of μ, $\varepsilon \sim \text{Logistic}(0, \mu^2)$. It follows that the choice probability is given by

$$P_h(\alpha_1,\alpha_2,\alpha_3,\alpha_4) = \Lambda(W_h/\mu)^{(1-y)} \times [1 - \Lambda(W_h/\mu)]^y, \tag{7}$$

where $\Lambda(.)$ is the CDF of logistic distribution.

Based on the numerical ways to compute a difference, let us now consider each heuristic model as a dual decision rule. Under DRIFT, the first rule is based on the level difference in money ΔY^{abs} and the time horizon τ, while the second rule is based on the relative difference in money ΔY^{rel} and annually compounded interest I. Accordingly, one can also decompose ITCH into two rules,

with the first rule being the same as that in DRIFT, and the second decision rule is determined by ΔY^{nor} and ΔT^{nor}.

The probability of observing choice B conditional on using the first rule is $P_1(y|\text{rule}_i = 1) = \Lambda[(\alpha_1(Y_{t+\tau} - Y_t) + \alpha_3 \tau)/\mu]$ for both heuristics. The probability of observing choice B conditional on applying the second criteria, is given by $P_2(y|\text{rule}_i = 2) = \Lambda[(\alpha_2(Y_{t+\tau} - Y_t)/Y_t + \alpha_4 ((Y_{t+\tau}/Y_t)^{(12/\tau)} - 1))/\mu]$ and $P_2(y|\text{rule}_i = 2) = \Lambda[(\alpha_2 (Y_{t+\tau} - Y_t)/((Y_{t+\tau} + Y_t)/2) + \alpha_4 (\tau/((2t + \tau)/2)))/\mu]$, respectively for DRIFT and ITCH. Recall that the relative difference can also be represented by the log difference. For example, in ITCH, the conditional probability with the second rule can be equivalently rewritten into $P_2(y|\text{rule}_i = 2) = \Lambda[(\alpha_2 (\ln(Y_{t+\tau}) - \ln(Y_t)) + \alpha_4 (\ln(t + \tau) - \ln(t)))/\mu]$.

Similar to the modeling of dual criteria decisions in Andersen, Harrison, Lau and Rutström [2014b], the decomposition of DRIFT and ITCH can also be modeled as the mixture model between the two decision rules. Let π be the probability of observed choices in the sample following the first rule, and $(1 - \pi)$ if the choices follow the second rule. Then, $\pi \times P_1(\cdot) + (1 - \pi) \times P_2(\cdot)$ defines the marginal probability of observing choice B under the two-rule mixture.

These decompositions make $P_1(\cdot)$ resembles a stochastic choice under Fechner specification, in which the two values are compared according to their level difference. As the probability $P_2(\cdot)$ is calculated based on the proportional difference of two values, the stochastic specification then mimics the Luce error. Next, we write out the application of the two stochastic specifications in the structural discounting models.

B. Stochastic Choices in the Structural Discounting Models

Recall that the choice in the discounted utility theory depends on the comparison of the present values of both options. First, consider the decision rule where the deterministic choice is governed by the level difference between the two present values, where the subjects choose B if PV_B

$- PV_A > 0$ and *vice versa*. A stochastic behavioral error term must be introduced to allow observed choices to deviate from deterministic theoretical predictions. So, now the observed choice depends on two parts: the deterministic choice based on the $PV_B - PV_A$ rule and the random error term $\boldsymbol{\varepsilon}$, such that $y = \mathbf{I}(PV_B - PV_A + \boldsymbol{\varepsilon} > 0)$. With $\boldsymbol{\varepsilon} \sim \text{Logistic}(0, \mu_{FC}^2)$, the probability of choosing option B is then $\Lambda(\nabla PV)$, in which the index ∇PV is given by

$$\nabla PV = (PV_B - PV_A)/\mu_{FC}. \tag{8}$$

As the noise parameter μ_{FC} approaches 0, the choice is deterministic. Conversely, as μ_{FC} gets arbitrarily large, the choice is purely random, driven entirely by noise, and both options have a 50:50 chance of being selected regardless of the underlying PV difference. This stochastic specification is popularized by Hey and Orme [1994] and is commonly known as Fechner (FC) model. It follows that the likelihood function for each choice observation takes the form

$$P^{FC}(\boldsymbol{\theta}, \mu_{FC}) = [1 - \Lambda(\nabla PV)]^{(1-y)} \times \Lambda(\nabla PV)^y, \tag{9}$$

where $\boldsymbol{\theta}$ denotes discounting parameter(s) in a structural discounting model.

The decision rule can be alternatively defined by the ratio of the two present values, where the observed choice is option B if $PV_B/PV_A > 1$ or $\ln(PV_B) - \ln(PV_A) > 0$. Similarly, a structural noise parameter must be added to allow some errors from the perspective of the deterministic theory so that the binary indicator is given by $y = \mathbf{I}(\ln(PV_B) - \ln(PV_A) + \boldsymbol{\varepsilon} > 0)$. Now represent the standard deviation of the behavioral error term $\boldsymbol{\varepsilon}$ with μ_{LC}. The probability of choosing option B is then specified by $\Lambda(\nabla \ln PV)$, where index $\nabla \ln PV$ is given by

$$\nabla \ln PV = [\ln(PV_B) - \ln(PV_A)]/\mu_{LC}. \tag{10}$$

The use of logistic cumulative density function makes the probabilistic choice function algebraically equivalent to $PV_B^{1/\mu_{LC}}/(PV_B^{1/\mu_{LC}} + PV_A^{1/\mu_{LC}})$, the common expression of Luce (LC) specification as introduced by Holt and Laury [2002]. Again, the choice is strictly determined by the PV of the two

options when $\mu_{LC} \to 0$ and essentially random when $\mu_{LC} \to \infty$. The likelihood of each choice is then given by

$$P^{LC}(\boldsymbol{\theta}, \mu_{LC}) = [1 - \Lambda(\nabla \ln PV)]^{(1-y)} \times \Lambda(\nabla \ln PV)^{y}. \quad (11)$$

C. Finite Mixture Approach for Heterogenous Decision Rules

As discussed earlier, the hybrid specification of the heuristic models implies a linear approximation of a mixture of two decision rules. Under the structural discounting models, such structure is generally seen as a mixture of two latent stochastic processes that generate each choice observation, one by the Fechner specification and another by the Luce specification.

The finite mixture model provides an ideal statistical framework to estimate the co-existence of more than one model in a population. AHLR [2008][2014a], for example, allowed a fraction of choices better characterized by exponential discounting and the other fraction better characterized by a non-constant discounting specification. We extend this mixture notion to the two discounting models with the same specification, such as the mixture of two exponential discounting models with a different stochastic process for each segment.[11] One may argue that the mixture of two structural discounting models may be advantageous over the heuristic models as they may provide some flexibility in capturing heterogeneous discounting behavior. To avoid such debate, we further restrict the model so that the two segments in the mixture specification share the same discounting parameters but have different error parameters. In other words, our mixture specification is basically the mixture of two behavioral error stories within a structural discounting model.

[11] Karp [2007] and Ekeland and Pirvu [2008], for example, adopted a similar approach in modeling non-constant discounting. The model is coined "pseudo-exponential discounting," which is basically a mixture of exponential discounting models to capture unobserved heterogeneity in the δ parameter.

Let π_{FC} denotes the probability that the Fechner specification generates a given choice observation. Accordingly, $(1 - \pi_{FC})$ is the probability that the Luce specification generates the observation. The mixture log-likelihood function is then given by

$$\ln[P(\boldsymbol{\theta}, \mu_{FC}, \mu_{LC}, \pi_{FC})] = \pi_{FC} \times \ln[P^{FC}(\boldsymbol{\theta}, \mu_{FC})] + (1 - \pi_{FC}) \times \ln[P^{LC}(\boldsymbol{\theta}, \mu_{LC})] \quad (12)$$

D. Joint Elicitation of EUT and Time Preferences

The risk aversion tasks in the experiments allow us to identify the utility function U(M) and perform the joint estimation strategy of AHLR [2008]. Consider the estimation of risk preferences under EUT as the simplest model of decision-making under risk. Each risk aversion task in the experiment presents a choice between two lotteries with two potential outcomes for each lottery. Let M_{ij} be the j^{th} outcome of lottery i, where i = A,B and j = 1,2. Assume the constant relative risk aversion (CRRA) utility function

$$U(M_{ij}) = [M_{ij}^{(1 - r)} - 1]/(1 - r) \quad (13)$$

for $M_{ij} > 1$ and $r \neq 1$, where r is the CRRA coefficient. Then, under EUT, $r = 0$ denotes linear utility or risk neutral behavior, $r > 0$ represents a concave utility function implying risk-averse behavior, and $r < 0$ characterizes a convex utility function or risk-loving attitude. The EU of lottery i is simply the probability weighted average of its outcome utilities,

$$EU_i = p(M_{i1}) \times U(M_{i1}) + p(M_{i2}) \times U(M_{i2}), \quad (14)$$

where $p(M_{i2}) = 1 - p(M_{i1})$.

Similar to the structural discounting models, the two lotteries are evaluated according to their level difference $EU_B - EU_A$ or its ratio EU_B/EU_A, where the choice of decision rules subsequently leads to different stochastic specifications. We assume that the same decision rule is used to evaluate the choices in both risk and discounting tasks.

Let υ_{FC} be the standard deviation of the behavioral error ε in the Fechner specification for the structural model of decision under risk. The probabilistic choice function is thus given by $\Phi(\nabla EU)$, where the index ∇EU is given by $\nabla EU = (EU_B - EU_A)/\upsilon_{FC}$. A multiplicative term to υ_{FC} is normally introduced to obtain a stochastic specification that is more theoretically coherent to the risk aversion concept. Specifically, the resulting stochastic specification ensures that the choice probabilities vary monotonically with the CRRA parameter.[12] Following Wilcox [2011], the multiplicative term is defined as the difference between the maximum and minimum of the potential outcome utilities in the lottery pair.[13] The index of expected utility difference is now given by

$$\nabla EU = (EU_B - EU_A)/ [(U_{max} - U_{min}) \times \upsilon_{FC}], \tag{15}$$

and the likelihood function for each choice observation takes the form

$$P^{FC}(r, \upsilon_{FC}) = \Lambda(\nabla EU)^{y} \times (1 - \Lambda(\nabla EU))^{(1 - y)}. \tag{16}$$

The assumption of non-linearity in the utility function for the discount rate estimation requires joint elicitation of risk and discounting parameter because the discounted utility of the background wealth and the monetary rewards in the discounting task is calculated using the estimate of the CRRA coefficient. The joint estimation of risk and discounting parameters can then be done by simultaneously maximizing the joint likelihood function in equations (9) and (16) so that the joint likelihood function is specified as:

$$P^{FC}(r, \theta, \mu_{FC}, \upsilon_{FC}) = P^{FC}(r, \upsilon_{FC}) \times P^{FC}(r, \theta, \mu_{FC}, \upsilon_{FC}). \tag{17}$$

The Luce specification defines the index $\nabla lnEU$ as

$$\nabla lnEU = [\ln(EU_B) - \ln(EU_A)] / \upsilon_{LC}, \tag{18}$$

---

[12] The issue of non monotone choice probabilities with the CRRA coefficient r is initially identified by Wilcox [2011] and reiterated by Apesteguia and Ballester [2018]. When one lottery is a mean-preserving spread of the other, this implies that the predicted probability of choosing the safer lottery does not always increase in the degree of risk aversion as measured by the r parameter.

[13] This generalization is operationalized in numerous studies, such as Andersen, Cox, Harrison, Lau, Rutström, and Sadiraj [2018], Andersen, Harrison, Lau, and Rutström [2018], Blavatskyy and Pogrebna [2010], and Dixit, Harb, Martínez-Correa, and Rutström [2015].

where υ_{LC} denotes the standard deviation of behavioral errors. The likelihood function for each observation is then given by

$$P^{LC}(r, \upsilon_{LC}) = \Lambda(\nabla \ln EU)^{y} \times (1 - \Lambda(\nabla \ln EU))^{(1-y)}, \tag{19}$$

while the joint likelihood function can be written as

$$P^{LC}(r, \boldsymbol{\theta}, \mu_{LC}, \upsilon_{LC}) = P^{LC}(r, \upsilon_{LC}) \times P^{LC}(r, \boldsymbol{\theta}, \upsilon_{LC}, \mu_{LC}). \tag{20}$$

We also expand our estimation by assuming a mixture of Fechner and Luce specifications in this joint estimation of EUT and structural discounting models. It is done by replacing the equation (12) with

$$\ln[P(r, \boldsymbol{\theta}, \mu_{FC}, \upsilon_{FC}, \mu_{LC}, \upsilon_{LC}, \pi_{FC})] = \pi_{FC} \times \ln[P^{FC}(r, \boldsymbol{\theta}, \mu_{FC}, \upsilon_{FC})] + (1 - \pi_{FC}) \times \ln[P^{LC}(r, \boldsymbol{\theta}, \mu_{LC}, \upsilon_{LC})]. \tag{21}$$

We estimate each structural discounting model with data that is pooled across all individuals to allow a characterization of discounting behavior of a representative individual. We use maximum likelihood (ML) estimation to estimate risk preference parameter r, discounting parameter(s) $\boldsymbol{\theta}$, the behavioral noise parameters μ_{FC}, υ_{FC}, μ_{LC}, and υ_{LC}, and the mixture probability π_{FC}.

E. Rank-Dependent Utility Theory Specifications

Each lottery in the risk aversion task can also be evaluated using the Rank-Dependent Utility (RDU) model of Quiggin [1982] such that the RDU evaluation of lottery i is

$$RDU_i = [w(p(M_{i1})) \times U(M_{i1})] + [(1 - w(p(M_{i1}))) \times U(M_{i2})], \tag{22}$$

where w(.) is the probability weighting function due to Prelec [1998]

$$w(p) = \exp\{-(-\ln p)^{\varphi}\}, \tag{23}$$

and is defined for $0 < p < 1$ and $\varphi > 0$.[14] The resulting function exhibits inverse-S probability weighting (optimism for small probability, and pessimism for large probability) for $\varphi < 1$, S-shaped probability weighting (pessimism for small probability, and optimism for large probability) for $\varphi > 1$, and linear probability weighting that reduces RDU to EUT when $\varphi = 1$. We simply have to examine all risk parameters to characterize risk preferences in the case of RDU: r and φ.

6. Results

Our first objective is to analyze the discounting attitude by each specification. For the structural discounting models, one can directly analyze the discounting behavior from the inferred discount rates. Conversely, their derivation is inexplicit in the heuristic models. One can only interpret the attitude towards intertemporal choices through the choice patterns captured by the sign and magnitude of its significant covariates.

Our second objective is to evaluate each model's goodness-of-fit. We compare the in-sample fitness by BIC and Vuong test. As in Ericson et al. [2015], we also use the cross-validation technique to compare the models' out-of-sample prediction.[15] We randomly split the subjects into estimation and hold-out samples with a 75:25 proportion. The models' parameters are estimated from each estimation sample. The estimates are then used to predict the choices in the hold-out sample. We

[14] One may extend the analysis by assuming a two-parameter Prelec probability weighting function given by $w(p) = \exp\{-\eta(-\ln p)^{\varphi}\}$, where $\eta > 0$ and $\varphi > 0$. Our main aim is to evaluate the effect of accounting for a non-linear utility function on the performance of the structural discounting models relative to their heuristic counterparts. We, therefore, argue that using RDU specification with a one-parameter Prelec probability weighting function should be sufficient for a robustness check.

[15] Ericson et al. [2015] also use mean absolute deviation (MAD) as the performance measure, which is defined as the absolute difference between the predicted choice probability and the observed choice. For example, when the observed choice is option B, and the calculated probability of choosing B is 60%, the MAD is 40%. Accordingly, if the observed choice is A, the MAD is then 60%. As rightly pointed out by Wulff and and van den Bos [2018], the use of MAD for model evaluation can be biased as it is inconsistent with the estimation procedure which use maximum likelihood.

repeat the process one hundred times for different estimation and hold-out splits and compare the arithmetic mean of the predicted log-likelihood values in the hold-out samples.

We consider 3×3 scenarios for the structural models using different combinations of stochastic models and utility assumptions. Other than Fechner and Luce errors, we also allow for a finite mixture of both stochastic specifications as a comparable configuration to the two heuristic specifications. Apart from the linear utility assumption, we also account for non-linear utility by EUT and RDU.

We also consider adding a constant term to the DRIFT and ITCH specifications as in Ericson et al. [2015]. Besides having no behavioral basis, the presence of the constant term in the estimation can produce counterintuitive signs of coefficients. Despite those concerns, however, the constant term may significantly contribute to the better fit of the heuristic models.

Finally, we compare the performance of the heuristic models to other structural discounting specifications which generalize the exponential discounting function. We consider two flexible functions, namely the Quasi-hyperbolic and Weibull hyperbolic discounting specifications.

A. Initial Estimates

Table 1 reports maximum likelihood (ML) estimates of the exponential (EXP) discounting with Fechner (FC) error specification.[16] Panel A presents the estimation results for EXP assuming risk neutrality. The estimate indicates a discount rate of 18.3% on an annualized basis, which is significantly different from 0 with a p-value < 0.001. The 95% confidence interval for this discount rate estimate is between 15.5% and 21.2%. Panel B shows the result of the joint estimation of EUT and EXP. We observe a concave utility function with the estimated CRRA coefficient of 0.546

[16] We present the estimation results of Exponential discounting and Simple Hyperbolic discounting with Luce specification in Appendix A.

(p-value < 0.001). The implied discount rate is consequently lower than when risk neutrality is assumed, which is now 10.2% (p-value < 0.001), with a 95% confidence interval between 8.6% and 11.8%. Finally, we observe a more concave utility function in the joint RDU-EXP, with an estimated r of 0.792 (p-value < 0.001), which accordingly yields to even lower annual discount rate of 7.92% (p-value < 0.001). The discount rate's 95% confidence interval is between 6.0% and 8.7%. The estimate of the φ parameter is 2.309 (p-value < 0.001), suggesting an S-shaped probability weighting function.

The parameter estimate in the Simple Hyperbolic (SH) discounting assuming linear utility (presented in Panel A of Table 2) indicates annual discount rates that are 19.8% for two weeks horizons and only decline to 18.1% for a one-year horizon. The 95% confidence interval in each horizon is roughly between 15% to 23%. When we allow for concave utility function, the observed quantitative magnitude of the declining discount rate is smaller and the elicited discount rates are lower. Assuming non-linear utility under EUT (Panel B), the discount rate for two weeks and the one-year horizon is 10.6% and 10.1%, respectively. The 95% confidence interval in each case is roughly between 8% to 12%. With RDU (Panel C), the discount rate for the two time horizons is 7.6% and 7.3%, respectively, with a 95% confidence interval is between around 6% to 9% in each case. The results suggest that there is no evidence of significantly declining discount rates.

Panel A and B of Table 3 displays the estimation results of DRIFT and ITCH, respectively. As Read and Scholten [2012] and Ericson et al. [2015] predicted, we observe positive coefficients for the money-related variables and negative coefficients for time-related variables, suggesting the trade-off between the money and time dimensions. The estimated coefficient of level difference in money (ΔY^{abs}) in DRIFT is 0.003, which is not significantly different from zero, with a p-value of 0.611. On the contrary, the estimated coefficient of proportional difference in money ΔY^{rel} is 7.848, which is statistically significant with a p-value < 0.001. The significant positive estimate implies an

increased probability of choosing the larger-later payment with the proportional difference between the two payments while keeping the level difference in money and time horizon fixed. Specifically, the average partial effect of the relative difference in money to the probability of choosing B is 1.762 (p-value < 0.001). The time horizon variable τ displays statistically significant coefficients of −0.133 (p-value < 0.001), representing a higher tendency of choosing the sooner payment as the time deliveries in both options are delayed by a common multiplicative constant. The average partial effect of the time horizon is −0.030 (p-value < 0.001). Finally, the estimate of the coefficient of the annually compounded interest I is 0.265 and not statistically significant, with a p-value of 0.325.

The coefficient of level difference in money ΔY^{abs} in ITCH is also not statistically significant, with an almost similar estimate of 0.002 (p-value = 0.791). The estimate of the coefficient and the average partial effect of the proportional difference in money ΔY^{nor} is 10.552 and 2.362, respectively, with p-values < 0.001. The time horizon τ has an estimated coefficient of −0.137 (p-value < 0.001), which is similar in magnitude to that in DRIFT. Finally, the proportional time difference ΔT^{nor} has a coefficient of −0.057, which is not statistically significant with a p-value of 0.375.

B. Assuming Fechner Specification and Risk Neutrality

We now evaluate the discounting models based on their performance in approximating the true data-generating process. First, assume risk neutrality and Fechner errors for EXP and SH. The BIC scores reported in Table 4 favor the heuristics over the two structural models with those assumptions with a ranking of ITCH $\succ_{bic}$ DRIFT $\succ_{bic}$ SH $\succ_{bic}$ EXP. The relation sign $\succ_{bic}$ implies that the BIC prefers the former model with its smaller information criterion to the latter. The BIC score of DRIFT and ITCH are 21,120 and 21,072, respectively.[17] Assuming linear utility and Fechner

[17] Recall that the estimation of both heuristic models generates two significant coefficients, which are on the relative difference in money and the time horizon. With a similar sign and magnitude of time horizon coefficient, the disparity in the goodness-of-fit between the two heuristics is caused by the different measures of the relative difference in money used in their specification. For a given interest rate, the proportional

errors, the maximum likelihood estimation of EXP yields a BIC score of 21,168. Meanwhile, the estimation of the SH generates a BIC score of 21,163.

The BIC comparison dichotomies the better and worse models arbitrarily as it makes little difference on the ranking of the models whether the BIC difference is 1 or 100. To allow for a more meaningful interpretation, we can approximate the significance of the BIC difference by comparing it to the likelihood ratio test. Recall that the BIC score is given by $BIC = -2 \times \ln L + \# \times \ln(N)$, where ln L is the log-likelihood value, # is the number of estimated parameters, and N is the total number of observations pooled over all individuals. With model 1 is nested within model 2, so that $\#_2 = \#_1 + k$, the BIC difference between the two is then a linear combination of the standard likelihood ratio statistic and the degrees of freedom penalty associated with the test, given by $\nabla BIC_{12} = [-2 \times (\ln L_1 - \ln L_2)] - [k \times \ln(N)]$. As the likelihood ratio test is asymptotically χ^2 distributed under the null hypothesis that the two models are equivalent, it implies that if the BIC difference between non-nested models exceeds the critical value of $\chi^2_{.05}(k)$, it also exceeds the BIC difference between the restricted and unrestricted models by default.

Let's now compare SH, the best performing structural model, and DRIFT, the worst-performing heuristic, by treating as if SH is nested within DRIFT. With two extra parameters in DRIFT, the critical value of $\chi^2_{.05}(2)$ equals 5.99. Since the BIC difference between the two models is 44, we can say that BIC strongly favors DRIFT over SH. Obviously, with the lowest information

---

difference in money measured by ΔY^{rel} grows faster with the time horizon τ than when it is measured by ΔY^{nor}. The resulting estimate of the coefficient is then higher in ITCH than in DRIFT. Replacing ΔY^{nor} with ΔY^{rel} deteriorates the performance of ITCH as the log-likelihood decrease to an almost similar value as in DRIFT, that is −10,544. Obviously, replacing ΔY^{rel} with ΔY^{nor} improves the log-likelihood of DRIFT to −10,521, almost similar to that of ITCH. Other functions reviewed in Tornqvist, Vartia, and Vartia [1985] can also be considered. For example, $\Delta Y^{prt} = (Y_{t+\tau} - Y_t)/\max(Y_{t+\tau}, Y_t)$, which grows with time horizon with a slower pace than ΔY^{nor}. Replacing ΔY^{rel} and ΔY^{nor} with ΔY^{prt} improve the log-likelihoods to −10,509 and −10,498, respectively, for DRIFT and ITCH. The results imply that the relative performances of the heuristic models are not robust to the measure of relative difference used in its specification, although the different measures offer the same behavioral interpretation. For a given interest rate, the slower the relative difference in money grows with the time horizon, the higher the coefficient and the better is the heuristic model. Appendix B presents the estimates of coefficients and the average partial effects in each scenario.

criterion, ITCH is significantly better than SH or EXP in approximating the data generating process. Such an analysis of BIC difference, however, is only a coarse approximation as the twice log-likelihood difference is no longer asymptotically χ^2 distributed if a model is not nested within the other. We now turn to the Vuong test as the more appropriate test statistic for non-nested models.

The Vuong test assumes that the log-likelihood differences between two specifications are asymptotically normally distributed and conducts a z-test on them. The null hypothesis is that the two models are equally close to the true data generating process, against the alternative that one model is closer. The Vuong test is directional in the sense that it favors the benchmark model if the z-statistic is considerably large and positive, and *vice versa*.

Figure 1 presents the z-statistics of the Vuong tests with the heuristic models as benchmarks.[18] The red circles and diamonds with solid lines represent the Vuong tests with DRIFT and ITCH as benchmarks, respectively. The shaded area from −1.96 to +1.96 in each panel is the inconclusive region, representing the non-rejection to the null hypothesis of indistinguishable fitness between the contending models with a 5% significance level. The first row of the figure shows that the tests consistently reject the null hypothesis of non-discrimination in performance between ITCH and EXP or SH with z-statistics of 3.728 (p-value < 0.001) and 3.585 (p-value < 0.001), respectively. With such large, positive z-statistics, the tests pick ITCH as the better model. The Vuong test also significantly favors DRIFT over EXP with z-statistics of 2.030 (p-value = 0.042). Although DRIFT provides a better fit than SH with a positive z-statistic of 1.833, the test shows that the log-likelihood difference between the two models is not statistically different from zero (p-value = 0.067).

We now compare the models' predictive ability using the cross-validation procedure. As illustrated in the first and second rows of Figure 2, we find that the heuristics outperform SH and

[18] We use the modified Vuong test that imposes a penalty on the log-likelihood difference due to less parsimony as in BIC. Following Desmarais and Harden [2013], the corrected log-likelihood difference of two models, say model 1 and model 2, in observation i is given by $\nabla \ln L_i = \ln L_{1i} - \ln L_{2i} + k \times \ln(N) / 2N$, where $k = \#_2 - \#_1$.

EXP with a ranking of ITCH $\succ_{cv}$ DRIFT $\succ_{cv}$ SH $\succ_{cv}$ EXP. The sign $\succ_{cv}$ means that the former model performs better than the latter based on the average cross-validation log-likelihoods. ITCH remains the best model with an average log-likelihood of −2,633. DRIFT is now outperformed by the two structural models, although its cross-validation log-likelihood only differ from SH and EXP in decimal values. On average, their cross-validation log-likelihood is −2,639.

Again, such a comparison does not indicate how significant the cross-validation log-likelihood difference is. To the best of our knowledge, a formal test statistic for the cross-validation comparison does not yet exist. The main issue is that the samples in the cross-validation are inherently dependent due to overlapping sets of estimation and hold-out samples across multiple replications. The log-likelihood differences are hence not normally distributed, which violates the key assumption of the standard *t*-test.[19] A test that does not impose any a priori assumption about the type of distribution is then preferred. However, as it still maintains the assumption of independent samples within each replication, the non-parametric test approach must be seen as an approximation to the cross-validation comparison in the absence of a more proper alternative.[20] We use the sign test as in Clarke [2003][2007] to compare the out-of-sample log-likelihoods. Under its null hypothesis, the median log-likelihood difference between the two models in the cross-validation

[19] Indeed, the Shapiro-Wilk normality test (Shapiro and Wilk [1965]) with a 5% significance level rejects the hypothesis that the cross-validation log-likelihood difference in each pair of models is normally distributed.

[20] Several studies proposed a modification to the t-test in response to the violation of non-normal limiting distribution, albeit lacking well-founded theoretical justification. To circumvent the issue of the underestimated variance, Nadeau and Bengio [2003], for example, proposed a variance correction factor, which is given by $[1/n + (n_e/n_h)]$, where n is the number of replications in the cross-validation, and (n_e/n_h) is the data splitting ratio of estimation and hold-out samples. In a different motivation, Harden and Desmarais [2011] adopted a modification by Johnson [1978] to correct the potential bias due to skewness in the distribution of the log-likelihood difference. Despite those corrections, the limiting distribution of the tests under the null hypothesis is still normal.

is zero.[21] Accordingly, two contending models are equivalent if the true proportion of positive or negative signs of the differences is one-half.

The Clarke tests reject the null hypothesis of equivalent cross-validation log-likelihood between ITCH and SH or EXP (p-values < 0.001) and prefers ITCH to SH and EXP. The sign tests, however, cannot discriminate the cross-validation log-likelihoods of DRIFT and SH or EXP with a p-value of 0.368. The approximations using the Clarke test is thus almost in line with the results in the in-sample evaluation using the Vuong test.

The above results imply two things. First, the disparity between our finding and the results in Ericson et al. [2015] and Wulff and Van den Boss [2018], especially on the weaker performance of DRIFT, highlights the importance of the incentivized experiment as the subject is aware of the economic consequence of their choices. Second, the better performance of at least one of the heuristic models indicates that Fechner may not be the best representation of the decision rule and the stochastic behavior of the representative agent. Next, we switch to Luce error and see if a different choice of error specification affects the relative performance of the structural discounting models.

C. Assuming Luce Specification and Risk Neutrality

The BIC scores presented in Table 4 show an improvement in the performance of the structural models by favoring SH over the heuristic models when Luce error is applied. The ranking among the four discounting models is now given by SH $\succ_{bic}$ EXP $\succ_{bic}$ ITCH $\succ_{bic}$ DRIFT. Assuming risk neutrality and applying Luce error, SH and EXP now generate lower BIC scores of 21,060. And 21,065, respectively. The gap in the information criterion with ITCH (BIC = 21,072) and DRIFT

[21] An obvious alternative is the Wilcoxon signed-rank test (Wilcoxon 1945]). However, unlike the Clarke test, the signed-rank test assumes that the distribution of the log-likelihood difference is symmetric.

(BIC = 21,120) is greater than the critical value of $\chi^2_{.05}(2)$, implying a significantly better performance of the two structural discounting models relative to ITCH and DRIFT.

The Vuong tests presented in the middle panel of Figure 1 generate a slightly different result. All z-statistics are now negative, which implies better fitness of SH and EXP relative to the heuristic models. The tests significantly reject the null hypothesis of equivalent fitness between DRIFT and SH or EXP with p-values < 0.001. The test also rejects the null hypothesis of indistinguishable performance between ITCH and SH (p-values = 0.036) but shows an insignificant fitness difference between ITCH and EXP (p-values = 0.301).

The out-of-sample prediction illustrated in the third row of Figure 3 confirms the results under the Vuong tests where SH and EXP with Luce error outperform the heuristic models with a ranking of SH $\succ_{cv}$ EXP $\succ_{cv}$ ITCH $\succ_{cv}$ DRIFT. The two structural models generate a similar mean of cross-validation log-likelihoods, which is −2,628, with SH having a slightly higher decimal value than EXP. Recall that ITCH and DRIFT generate log-likelihood predictions of −2,633 and −2,639, respectively. The Clarke tests for each pair of the structural and heuristic models reject the null hypothesis of zero medians of log-likelihood differences with p-values < 0.001 to favor SH and EXP.

To conclude, we observe an improved performance of SH and EXP as we switch from Fechner to Luce specification while maintaining the risk neutrality assumption. SH and EXP outperform ITCH and DRIFT both in the in- and out-of-sample evaluation, with a significant difference in goodness-of-fit between the structural and the heuristic discounting models, except for the performance comparison between ITCH and EXP. While the two previous evaluations assume risk neutrality for the structural model, plentiful experimental literature emphasizes the necessity to allow for a non-linear utility function to remedy an upward bias in the estimated discount rates. We

next evaluate the effects of a non-linear utility function on the fitness of EXP and SH relative to the two heuristic specifications.

D. Allowing for A Non-linear Utility Function

When allowing for a non-linear utility function, we use the adjusted log-likelihood of the joint estimation of standard discounting models with EUT or RDU to calculate BIC as well as the Vuong statistic. The adjusted log-likelihood is calculated by applying the estimated CRRA coefficient and discounting parameters to the observed choices on the discounting tasks.[22] Accordingly, the penalty term in BIC and Vuong test only considers the CRRA coefficient, the discounting parameters, and the error parameter in the discounting estimation. The probability weighting parameter is excluded from the penalty calculation as we do not use the parameter to compute the discounted utility. For instance, the BIC penalty for the joint RDU and Exponential Discounting model only considers three parameters, namely r, δ, and μ.

As presented in Table 4, allowing for a non-linear utility function only improves the performance of the structural discounting models with the Fechner error but not with the Luce error. The in-sample fit comparison yields a ranking of ITCH $\succ_{bic}$ SH $\succ_{bic}$ EXP $\succ_{bic}$ DRIFT, which is robust across the underlying theories of risky behavior and stochastic specifications. Assuming

[22] From Section 3, the adjusted log-likelihood is thus given by $\ln P^S(\hat{r}, \hat{\theta}, \hat{\mu}_S)$, where $\hat{r}$ is the estimated CRRA coefficient under EUT or RDU, $\hat{\theta}$ is the estimated discounting parameters, $\hat{\mu}_S$ is the estimated standard deviation of the behavioral errors of present value difference, and $S = \{FC, LC\}$.

One may use separate estimations, in which the risk parameters under EUT or RDU are estimated in advance, and then use the estimated CRRA coefficient in the estimation of the discounting parameters to account for the non-linear utility assumption. It is equivalent to exogenously setting the CRRA coefficient for the estimation of structural discounting models. In this case, we only need to count the discounting parameters to calculate the penalty in BIC and the Vuong test. Consequently, regardless of assuming a linear or a non-linear utility function, the number of parameters for the penalty calculation is the same. This estimation procedure consequentially affects the BIC ranking. EXP and SH now outperform DRIFT and ITCH. The only exception is when we compare ITCH to EXP with EUT and Fechner error as they have a comparable BIC score. We provide the BIC scores calculated using this assumption in Table C1 of Appendix C.

Fechner error, the BIC scores of SH are 21,079 and 21,073, respectively, with EUT and RDU. With Luce error, the BIC scores of joint EUT-SH and RDU-SH are 21,073 and 21,072, respectively, higher than when we assume linear utility (BIC = 21,060). EXP with Fechner error generates BIC scores of 21,082 and 21,075, respectively, when jointly estimated with EUT and RDU. The BIC scores of EXP with Luce error are 21,076 and 21,075, respectively, with a non-linear utility assumption under EUT and RDU. Those BIC scores are higher than that of ITCH (BIC = 21,072) but lower than that of DRIFT (BIC = 21,120).

When the restricted and unrestricted models only differ by one parameter, the comparison of the two models based on BIC is equivalent to applying a two-sided t-test with a critical value of $\sqrt{\ln(N)}$ (Raftery [1995] and Cameron and Trivedi [2009, p. 279]), which equals 2.45 in our case. Roughly compared to the critical value of the two-sided t-test, the differences in BIC scores between ITCH and EXP/SH are significant. The only exceptions are when we compare ITCH to SH with a non-linear utility function and Luce error and ITCH to SH with RDU and Fechner error. We also find that the BIC difference between DRIFT and EXP or SH is at least 38, indicating a significantly better fitness of the latter over the former.

The top and middle panels of Figure 1 illustrate the Vuong tests for the scenario of allowing for a non-linear utility for the structural discounting models. While ITCH provides a better fit than SH and EXP, the log-likelihood differences, according to the Vuong tests, are not sufficiently large to reject the null hypothesis of equivalent fitness (p-values ≥ 0.434). The z-statistics of tests with ITCH as the benchmark fall in the inconclusive region, with values varying between 0.051 to 1.950. The lowest z-statistics is found for the test against SH with EUT and Luce error, while the highest z-statistic is generated in the test against EXP with EUT and Luce error. With DRIFT as the benchmark, all tests reject the null hypothesis of non-discrimination in fitness to the data between

DRIFT and EXP or SH (p-values < 0.001) to favor the two structural models with a non-linear utility, regardless of the stochastic specification.[23]

Contrarily to the in-sample results, our out-of-sample predictions select SH and EXP over the heuristic models with a ranking of SH $\succ_{cv}$ EXP $\succ_{cv}$ ITCH $\succ_{cv}$ DRIFT. The second and third rows of Figure 2 illustrate this. Regardless of the stochastic model, the means of cross-validation log-likelihoods of SH and EXP with a non-linear utility function are around −2,628, thus higher than ITCH (−2,633) and DRIFT (−2,639). All Clarke tests reject the null hypothesis that the median log-likelihood difference between ITCH/DRIFT and SH/EXP equals zero and favor SH and EXP over the two heuristics with a 5% significance level.

To sum up, by allowing for the non-linear utility assumption to correct the upward bias in the elicited discount rates, we conversely find that the in-sample evaluations pick SH and EXP over DRIFT. Although ITCH remains the best model under BIC, the Vuong tests do not show it to be significantly superior to SH and EXP. Instead, we find a significantly better performance of the two structural models than the heuristic models in the out-of-sample evaluations. As illustrated earlier in Section 5, integrating level and proportional differences make the heuristic models resemble the mixture of Fechner and Luce specifications. Next, we expand our analysis by comparing the performance of DRIFT and ITCH to those of the structural discounting models assuming heterogeneous stochastic behavior.

E. The Mixture of Fechner and Luce Specifications

With the mixture of Fechner and Luce errors, both in- and out-of-sample evaluations now agree to the superiority of SH and EXP with a ranking of SH $\succ_{bic}$ EXP $\succ_{bic}$ ITCH $\succ_{bic}$ DRIFT.[24] The

[23] The results are robust even if we exclude the CRRA coefficient from the number of parameters in the penalty calculation (see Figure C1 in Appendix C).

[24] The maximum likelihood estimations of the mixture models for all structural discounting specifications are provided in Table D1 and D2 of Appendix D. The results show that the probability of the

BIC scores presented in Table 4 show that the exponential discounting model with linear utility and the mixture of the two error specifications generate an information criterion of 20,955. With non-linear utility under EUT and RDU, the BIC score of EXP increases to 20,973 and 20,987, respectively. Under the assumption of heterogeneous behavioral error, the BIC score of SH with linear utility is 20,933. Relaxing the risk neutrality assumption, the BIC scores of SH are rather higher, which are 20,959 and 20,974, respectively, under EUT and RDU. With the BIC score of 21,072 in ITCH, the smallest BIC difference is 85, given by the comparison between ITCH against joint RDU-EXP. The difference is far above the critical points of $\chi^2_{.05}(3)$, implying a strong preference towards SH and EXP with the mixture of Fechner and Luce specifications.

The Vuong tests illustrated in the bottom panel of Figure 1 confirm the results. All tests consistently reject the null hypothesis of equivalent fitness (p-values ≤ 0.001). The generated z-statistics vary from -3.212 to -6.968, thus significantly favoring SH and EXP with the mixture of the stochastic models, regardless of the linear or non-linear utility curvature.

The bottom panel of Figure 2 presents the distribution of the log-likelihood values of the cross-validations. SH and DRIFT respectively provide the best and worst fit to the data, generating a ranking of SH $\succ_{cv}$ EXP $\succ_{cv}$ ITCH $\succ_{cv}$ DRIFT. Similar to the in-sample evaluations, we also observe lower log-likelihood values when turning from linear utility assumption to EUT or RDU. The average log-likelihood of EXP with the heterogeneous stochastic behavior varies between $-2{,}613$ and $-2{,}616$. Meanwhile, the average log-likelihood of SH with the mixture of the two error models is between $-2{,}610$ and $-2{,}615$. As ITCH and DRIFT generate lower means of the out-of-sample log-likelihoods, which are $-2{,}633$ and $-2{,}639$, respectively, both SH and EXP significantly outperform the heuristic models, regardless of the linearity assumption of the utility functions, with sign test p-values < 0.001.

Luce specification generating the data is at least 66% in each estimation. In addition, we also observe that the estimates of SH in the mixture model show robust evidence of almost constant discounting.

In summary, when we assume the mixture of two stochastic specifications, SH and EXP now outperform their heuristic counterparts regardless of the assumption of linear or non-linear utility in all evaluations. This result thus supports our analytic observation, which discovers the similarity between the stochastic structure in the heuristic models to the finite mixture of Fechner and Luce errors in the structural discounting models.

F. Other Non-Constant Discounting Specifications

We now consider comparing heuristic models with two other non-constant discounting specifications. The front-end delay treatment in the experiments allows us to test the immediate temptation behavior. While ITCH uses its relative difference in time ΔT^{nor}, the discounting theory commonly represents such attitude with the Quasi-hyperbolic (QH) discounting (Elster [1979], Laibson [1997], and Phelps and Pollak [1968]). Another flexible non-constant discounting specification to consider is the Weibull Hyperbolic (WEI) discounting (Read [2001] and Prelec [2004]).[25]

We find a similar ranking to the previous comparisons in every change of assumption of linear/non-linear utility curvature and the stochastic specification. When we assume risk neutrality and Fechner error, we find that all heuristic models also outperform QH and WEI.[26] As recollected

[25] The discount function of the quasi-hyperbolic discounting is given by $D^{QH}(t) = \beta \times 1/(1+\delta)^t$ when $t > 0$ and $D^{QH}(t) = 1$ at time $t = 0$, where β is present bias parameter. Under the Weibull hyperbolic discounting, it is defined as $D^{WEI}(t) = [1/(1+\delta)]\, t^{(1/\zeta)}$, where $\zeta > 0$. The EXP specification is nested within the two specifications as a special case of $\beta = 0$ and $\zeta = 1$, respectively, under the quasi-hyperbolic and Weibull hyperbolic models. We present the estimation results of the two discounting specifications in Appendix D. The discount rates elicitation under the Quasi-hyperbolic specification also show robust evidence of constant discounting. Assuming linear or non-linear utility, we find that the estimate of the present bias parameter is not statistically different from 1, suggesting that there is no quasi-hyperbolic discounting. The estimate of the δ parameter is also close to the estimate of δ from the Exponential discounting. We also observe that the Weibull discounting also collapses to the Exponential model, as the estimate of parameter which characterizes the decreasing impatience is also not statistically different from 1 in each estimation.

[26] Appendix D provides more detailed analyses.

from earlier evaluations, the BIC scores of ITCH and DRIFT are 21,072 and 21,120, respectively. On the other hand, QH generates a BIC score of 21,174, while WEI yields 21,176. With those gaps, the raw approximation using the likelihood ratio test suggests significant differences in performance, thus favoring ITCH and DRIFT over the two non-constant discounting models.

Accounting for Luce error and maintaining the linear utility assumption also improve the BIC scores of QH and WEI to 21,066 and 21,073, respectively, making QH better than ITCH and DRIFT. The difference with the BIC score of ITCH is insignificant if we roughly compare it to the likelihood ratio test. While being insignificantly outperformed by ITCH, WEI performs significantly better than DRIFT.

The non-linear utility assumption only improves the performance of the QH and WEI when Fechner error is used, but not with Luce error. Their BIC scores vary from 21,076 to 21,090. The range of information criteria thus implies a significantly inferior performance of QH and WEI relative to ITCH, and at the same time, a significantly superior performance relative to DRIFT.[27]

With a mixture of the two stochastic specifications, QH and WEI now significantly outperformed the two heuristics, regardless of linear or non-linear utility assumptions. The BIC scores vary between 20,936 and 20,996. The difference between the BIC score of ITCH, which equals 21,072, and the highest score in the range is significant when roughly compared to the critical value of $\chi^2_{.05}(2)$, which is 5.99.

The results in BIC comparisons are confirmed by the Vuong tests. The only exception is when we account for non-linear utility, in which the Vuong tests can not differentiate the fitness of ITCH to QH and WEI with a 5% significance level, despite the positive z-statistics. All tests

[27] We find a different ranking when we ignore the CRRA coefficient in the number of parameters to calculate the penalty of BIC. ITCH significantly outperforms QH with EUT and Fechner Error. ITCH is also significantly better than WEI with EUT and any error models but is not significantly better than WEI with RDU and Fechner or Luce error. ITCH and DRIFT have a significantly inferior fitness relative to QH with RDU and Fechner and QH with EUT/RDU and Luce. We presented the details in Table E8.

significantly favor the structural models with the mixture of Fechner and Luce errors over the two heuristic models with p-values < 0.001.

Our out-of-sample analyses find a consistent result to the Vuong tests. The two heuristic models significantly outperform QH and WEI with risk neutrality and Fechner errors. The Clarke tests, however, can not reject the null hypothesis that the median log-likelihood difference between ITCH and QH or WEI in the cross-validation equals zero when we use Luce error and/or assume the non-linear utility function. The tests, however, significantly favor QH and WEI over DRIFT with a 5% significance level. Finally, the cross-validation log-likelihood of QH and WEI with the mixture of the stochastic specifications are significantly higher than those of ITCH and DRIFT. Indeed, the sign tests reject the null hypothesis of equivalent log-likelihoods to favor QH or WEI with p-values < 0.001.

In summary, we find the same results as the previous evaluations when comparing the heuristic models to QH and WEI models. The two heuristic models outperform the structural discounting models when assuming risk neutrality and Fechner error. However, QH and WEI consistently outperform DRIFT once we move away from Fechner error or risk neutrality assumption. The in- and out-of-sample evaluations exhibit an equivalent fitness of QH and WEI to ITCH when the structural models are combined with Luce error or when combined with non-linear utility assumption and any stochastic specification. Finally, the structural discounting models with the mixture of the stochastic specifications consistently outperform their heuristic counterparts.

G. Extension: The Effect of Constant Term on the Performance of Heuristic Models

Now we add a constant term to the DRIFT and ITCH specifications as in Ericson et al. [2015]. Here we aim to re-analyze the finding by Wulff and Van den Boss [2018], who show the significant contribution of the constant term to the heuristics' goodness-of-fit despite its lack of

underlying behavior. We denote the heuristic models containing the constant term by DRIFT$^+$ and ITCH$^+$, respectively.

Panel C of Table 3 provides the estimates of DRIFT$^+$ with α_0 representing the constant term. We observe three variables with statistically significant coefficients, which are the constant term (−1.039, *p*-value < 0.001), the annually compounded interest I (3.188, *p*-value < 0.001), and the relative difference in money ΔY^{rel} (3.006, *p*-value = 0.046). The coefficient of the time horizon τ is unexpectedly positive, that is 0.002, and statistically insignificant (*p*-value = 0.869). Finally, the coefficient of the level difference in money ΔY^{abs} remains insignificant, that is, 0.003 (*p*-value = 0.609). The results imply that the trade-off is represented by the relative difference in money and annually compounded interest against the constant term, although the latter is less meaningful in a behavioral term.[28]

From the estimation result presented in Panel D of Table 3, the variables with significant coefficients in ITCH$^+$ are the constant term, the proportional difference in money ΔY^{nor}, and the time horizon τ. The constant term is significant at 5% level with an estimate of −0.436 (*p*-value = 0.037). The estimated coefficient of the relative difference in money ΔY^{nor} is 10.798, while the coefficient of time horizon τ is estimated at −0.138, both with a *p*-value < 0.001. The estimated coefficient of level difference in money ΔY^{abs} is 0.001, which is not significantly different from zero (*p*-value = 0.876). Unlike what is expected by Ericson et al. [2015], the sign of the coefficient of the

[28] The constant term in the logit regression by itself has no intrinsic meaning as, in our experiment data, all DRIFT covariates never equal zero. One may then interpret the estimated constant term as a decision threshold. However, using only one decision threshold for the entire comparisons confounds the interpretation as it is not obvious how the models scale the money and time dimensions into a common currency. In that case, it is more sensible to allow the thresholds to vary across dimensions as in the lexicographic semi-order model of Dai et al. [2018] or the stochastic difference model of González-Vallejo [2012].

relative difference in time ΔT^{nor} is positive, that is 0.190, although it is not statistically significant (p-value = 0.176).[29]

Despite the counterintuitive sign of coefficient, we find that DRIFT$^+$ fits data the best, followed by ITCH$^+$ in second place when compared against EXP, SH, QH, and WEI with homogeneous stochastic behavior and any decision theory of risky behavior. As shown in Table 5, the BIC score of DRIFT$^+$ is 20,801, while that of ITCH$^+$ is 21,010. Recall from the previous evaluations that the information criteria of EXP, SH, QH, and WEI with Fechner or Luce error vary between 21,060 and 21,176. With the likelihood ratio test analogy, all BIC differences between DRIFT$^+$/ITCH$^+$ and any structural discounting model are considered significant at the 5% level. However, when we assume a mixture of Fechner and Luce errors, the four structural discounting models outperform ITCH$^+$. The BIC scores of constant and non-constant discounting models with the discrete heterogeneous stochastic behavior span between 20,933 and 20,996. With this range, the BIC difference between ITCH$^+$ and any structural discounting model is significant when approximated using the likelihood ratio test with a 5% significance level.

The blue circles in Figure 1 represent the χ-statistic of Vuong tests with DRIFT$^+$ as the benchmark against EXP and SH, while the blue diamonds represent the χ-statistic of Vuong tests with ITCH$^+$ as the benchmark. The tests against QH and WEI are presented in Figure E1 of

[29] Without the constant term, a positive coefficient of the proportional difference in money ΔT^{nor}, combined with a positive coefficient of the level and proportional differences in money (ΔY^{abs} and ΔY^{rel}), and a negative coefficient of time horizon τ, make ITCH exhibit heterogeneous discounting behaviors. Suppose that in the choices between immediate and delayed payments with a fixed interest rate, we observe an increasing probability of choosing the delayed options with a longer time horizon. Then also assume that when we add a front-end delay to the immediate rewards and keep the time horizon fixed, the observed choices are always the sooner options. The positive coefficient of level and proportional difference in money and negative coefficient of time horizon capture the first behavior because the monetary difference grows faster than the time horizon. The positive coefficient of the proportional time difference captures the second behavior. Under a structural discounting model, a representative agent may only exhibit one of the above behaviors but not both unless we account for unobserved individual heterogeneity. The first behavior can only be captured by a decreasing discount rates but not by the exponential specification. The second behavior can only be represented by an increasing discount rates.

Appendix E. The Vuong tests with DRIFT$^+$ as the benchmark always generate large and positive z-statistics, ranging from 4.167 to 8.991, thus consistently rejecting the null hypothesis of equivalent fitness to favor the heuristic model over EXP, SH, QH, and WEI in any scenario. The tests also significantly favor ITCH$^+$ over the four structural discounting models, except when we assume heterogeneous stochastic behavior. The tests against EXP, SH, QH, and WEI with the mixture of Fechner and Luce errors are generally inconclusive, although the z-statistics point to the direction of the four structural models as they vary between -0.417 to -1.738. The only exceptions are found in SH and WEI with risk neutrality and heterogeneous stochastic behavior, where the tests now significantly reject the null hypothesis of non-discrimination with respective z-statistics of -2.255 and -2.227 and favor SH and WEI with p-values of 0.024 and 0.026, respectively.

A slightly different result is observed again in the out-of-sample evaluations. The cross-validations generate an average log-likelihood of $-2{,}599$ for DRIFT$^+$ and $-2{,}630$ for ITCH$^+$. EXP, SH, QH, and WEI are still outperformed by DRIFT$^+$, although we assume the co-existence of Fechner and Luce specifications. The sign tests significantly reject the null hypothesis of equivalent cross-validation log-likelihood distribution to favor DRIFT$^+$ with p-values < 0.001. The sign tests with a 5% significance level show that ITCH$^+$ has a significantly better cross-validation log-likelihood than the four structural discounting models when assuming risk neutrality and Fechner error. The tests cannot reject the null hypothesis of equivalent log-likelihood between ITCH$^+$ and the four structural discounting models with risk neutrality and Luce error or with non-linear utility and any stochastic specification. However, ITCH$^+$ is significantly outperformed by EXP, SH, QH, and WEI with the mixture of the two stochastic specifications.

In conclusion, the constant term introduced by Ericson et al. [2015] to the heuristic models counts as beyond necessity. It lacks behavioural justifications and, at the same time, potentially

confound the overall behavioural interpretation. Despite those drawbacks, it appears to be too dominant in supporting the heuristics' goodness-of-fit.[30]

7. Conclusion

Using data from a field experiment, we evaluate the predictive performance of heuristic and structural discounting models. We find that the seemingly superior performance of the heuristic discounting models is due to its oversimplified structural econometrics, not because the heuristic models are inherently superior. When we adopt a mixture error specification and/or control for utility curvature, even the most inflexible structural model, the exponential discounting model, does better than heuristic models.

The constructive implications of our analyses are two-fold. First, one may consider the heuristic discounting models as a simple diagnostic tool for choosing between Fechner and Luce error specifications. Second, future research should exert more effort on proposing stochastic choice models for intertemporal choice analyses.

[30] We find that the constant term in DRIFT^{+} lends the heuristic model the flexibility to capture a particular choice pattern in the experiment. From Section 2, a choice set in the experiment varies the interest rate from 5% to 50% but keeps the time horizon fixed. One of the treatments applied in the experiment is that the time horizon varies between choice sets. When examining the choices across different sets with the same principal payment, we find that the proportion of choosing larger-later (LL) increases with the time horizon for a given interest rate. This choice pattern is not problematic to the identification and estimation of the individual discount rate by the structural discounting models because the objective is to infer the individual discount rates at which the subject would be indifferent between the smaller-sooner (SS) payment and larger-later (LL) payment in each choice set. All discounting models, except DRIFT^{+}, predict an increasing probability of choosing LL with the time horizon only when the offered interest rate is above the individual discount rates but decreasing when it is lower than the individual discount rates. As presented in Panel C of Table 3, the estimated coefficients for DRIFT^{+} imply an increasing tendency to choose LL as the time horizon is prolonged and the interest rate is fixed, regardless of whether it is lower or higher than the individual discount rates. We elucidate the finding in a more detailed analysis in Appendix F.

Table 1: Estimates of Exponential Discounting with Fechner Error

Variable	Estimate	St. Error	*p*-value	95% Confidence Interval	
		A. Assuming Linear Utility (Log-likelihood = −10,574.2)			
δ	0.183	0.015	<0.001	0.155	0.212
μ_{FC}	22.118	1.503	<0.001	19.172	25.065
		B. Assuming Non-linear Utility under EUT (Adjusted Log-likelihood = −10,526.2)			
r	0.546	0.031	<0.001	0.486	0.606
δ	0.102	0.008	<0.001	0.086	0.118
υ_{FC}	0.179	0.011	<0.001	0.158	0.201
μ_{FC}	1.031	0.186	<0.001	0.666	1.396
		C. Assuming Non-linear Utility under RDU (Adjusted Log-likelihood = −10,523.0)			
r	0.792	0.046	<0.001	0.702	0.882
φ	2.039	0.104	<0.001	1.836	2.243
δ	0.073	0.007	<0.001	0.060	0.087
υ_{FC}	0.231	0.014	<0.001	0.204	0.258
μ_{FC}	0.265	0.069	<0.001	0.130	0.400

Table 2: Estimates of Simple Hyperbolic Discounting with Fechner Error

Variable	Estimate	St. Error	*p*-value	95% Confidence Interval	
		A. Assuming Linear Utility (Log-likelihood = −10,572.0)			
$\varkappa$	0.181	0.014	<0.001	0.153	0.209
μ_{FC}	21.999	1.511	<0.001	19.037	24.961
		B. Assuming Non-linear Utility under EUT (Adjusted Log-likelihood = −10,524.7)			
r	0.546	0.031	<0.001	0.486	0.606
$\varkappa$	0.101	0.008	<0.001	0.085	0.117
υ_{FC}	0.179	0.011	<0.001	0.158	0.201
μ_{FC}	1.028	0.186	<0.001	0.663	1.392
		C. Assuming Non-linear Utility under RDU (Adjusted Log-likelihood = −10,521.9)			
r	0.792	0.046	<0.001	0.702	0.882
φ	2.039	0.104	<0.001	1.836	2.243
$\varkappa$	0.073	0.007	<0.001	0.060	0.087
υ_{FC}	0.231	0.014	<0.001	0.204	0.258
μ_{FC}	0.264	0.069	<0.001	0.130	0.399

Table 3: Parameter Estimates and Average Partial Effects of Decision Heuristics

Variable	Estimate/ Avg. Partial Effect	St. Error	*p*-value	95% Confidence Interval	
		A. Difference-Ratio-Interest-Finance-Time (DRIFT) (Log-likelihood = −10,540.4)			
Parameter Estimates					
α_1 (ΔY^{abs})	0.003	0.006	0.611	−0.009	0.016
α_2 (ΔY^{rel})	7.848	1.611	<0.001	4.691	11.005
α_3 (τ)	−0.133	0.013	<0.001	−0.159	−0.107
α_4 (I)	0.265	0.270	0.325	−0.263	0.794
Average Partial Effects					
α_1 (ΔY^{abs})	0.001	0.001	0.611	−0.002	0.004
α_2 (ΔY^{rel})	1.762	0.358	<0.001	1.060	2.464
α_3 (τ)	−0.030	0.003	<0.001	−0.036	−0.024
α_4 (I)	0.060	0.060	0.323	−0.059	0.178
		B. Inter-temporal Choice Heuristic (ITCH) (Log-likelihood = −10,516.3)			
Parameter Estimates					
α_1 (ΔY^{abs})	0.002	0.006	0.791	−0.011	0.014
α_2 (ΔY^{nor})	10.552	1.867	<0.001	6.892	14.212
α_3 (τ)	−0.137	0.016	<0.001	−0.168	−0.106
α_4 (ΔT^{nor})	−0.057	0.064	0.375	−0.182	0.069
Average Partial Effects					
α_1 (ΔY^{abs})	0.000	0.001	0.791	−0.002	0.003
α_2 (ΔY^{rel})	2.362	0.408	<0.001	1.563	3.161
α_3 (τ)	−0.031	0.003	<0.001	−0.038	−0.024
α_4 (ΔT^{nor})	−0.013	0.014	<0.375	−0.041	0.015
		C. Difference-Ratio-Interest-Finance-Time with The Constant Term (DRIFT $^+$) (Log-likelihood = −10,376.0)			
Parameter Estimates					
α_1 (ΔY^{abs})	0.003	0.006	0.609	−0.009	0.016
α_2 (ΔY^{rel})	3.006	1.504	0.046	0.059	5.953
α_3 (τ)	0.002	0.014	0.869	−0.026	0.031
α_4 (I)	3.188	0.255	<0.001	2.688	3.688
α_0 (const)	−1.039	0.118	<0.001	−1.270	−0.807

Average Partial Effects

α_1 (ΔY^{abs})	0.001	0.001	0.609	−0.002	0.003
α_2 (ΔY^{rel})	0.660	0.328	0.044	0.016	1.303
α_3 (τ)	0.001	0.003	0.869	−0.006	0.007
α_4 (I)	0.699	0.053	<0.000	0.595	0.804

D. Inter-temporal Choice Heuristic with The Constant Term (ITCH$^+$)
(Log-likelihood = −10,480.7)

Parameter Estimates

α_1 (ΔY^{abs})	0.001	0.007	0.876	−0.012	0.014
α_2 (ΔY^{nor})	10.798	1.885	<0.001	7.104	14.493
α_3 (τ)	−0.138	0.016	<0.001	−0.169	−0.107
α_4 (ΔT^{nor})	0.190	0.141	0.176	−0.085	0.466
α_0 (const)	−0.436	0.206	0.034	−0.839	−0.033

Average Partial Effects

α_1 (ΔY^{abs})	0.000	0.001	0.876	−0.003	0.003
α_2 (ΔY^{rel})	−0.031	0.003	<0.001	−0.037	−0.024
α_3 (τ)	2.406	0.410	<0.001	1.603	3.210
α_4 (ΔT^{nor})	0.042	0.031	<0.174	−0.019	0.104

Table 4: Bayesian Information Criteria (BIC)

Discounting Models	Utility	Error Models	No. of Parameters	Adj. Log-Likelihood	BIC
		A. Heuristic Discounting Models			
DRIFT			4	−10,540	21,120
ITCH			4	−10,516	21,072
DRIFT^{+}			5	−10,376	20,801
ITCH^{+}			5	−10,481	21,010
		B. Exponential Discounting (EXP)			
EXP	Linear	Fechner	2	−10,574	21,168
EXP	EUT	Fechner	3	−10,526	21,082
EXP	RDU	Fechner	3	−10,523	21,075
EXP	Linear	Luce	2	−10,523	21,065
EXP	EUT	Luce	3	−10,523	21,076
EXP	RDU	Luce	3	−10,523	21,075
EXP	Linear	Fechner+Luce	4	−10,458	20,955
EXP	EUT	Fechner+Luce	5	−10,462	20,973
EXP	RDU	Fechner+Luce	5	−10,469	20,987
		C. Simple Hyperbolic Discounting (SH)			
SH	Linear	Fechner	2	−10,572	21,163
SH	EUT	Fechner	3	−10,525	21,079
SH	RDU	Fechner	3	−10,522	21,073
SH	Linear	Luce	2	−10,520	21,060
SH	EUT	Luce	3	−10,522	21,073
SH	RDU	Luce	3	−10,521	21,072
SH	Linear	Fechner+Luce	4	−10,447	20,933
SH	EUT	Fechner+Luce	5	−10,455	20,959
SH	RDU	Fechner+Luce	5	−10,463	20,974

Notes: The adjusted log-likelihood value is calculated by applying the elicited CRRA and discounting parameters only to the discounting tasks. Accordingly, the number of parameters is the number of discounting parameters added with the number of CRRA parameter. For instance, the number of parameters in the joint RDU and Exponential Discounting model is 3, consisting of parameter r, δ and γ.

Figure 1: The Vuong's z-Statistics with Heuristic Discounting Models as Benchmarks

Against Exponential and Simple Hyperbolic Discounting Models

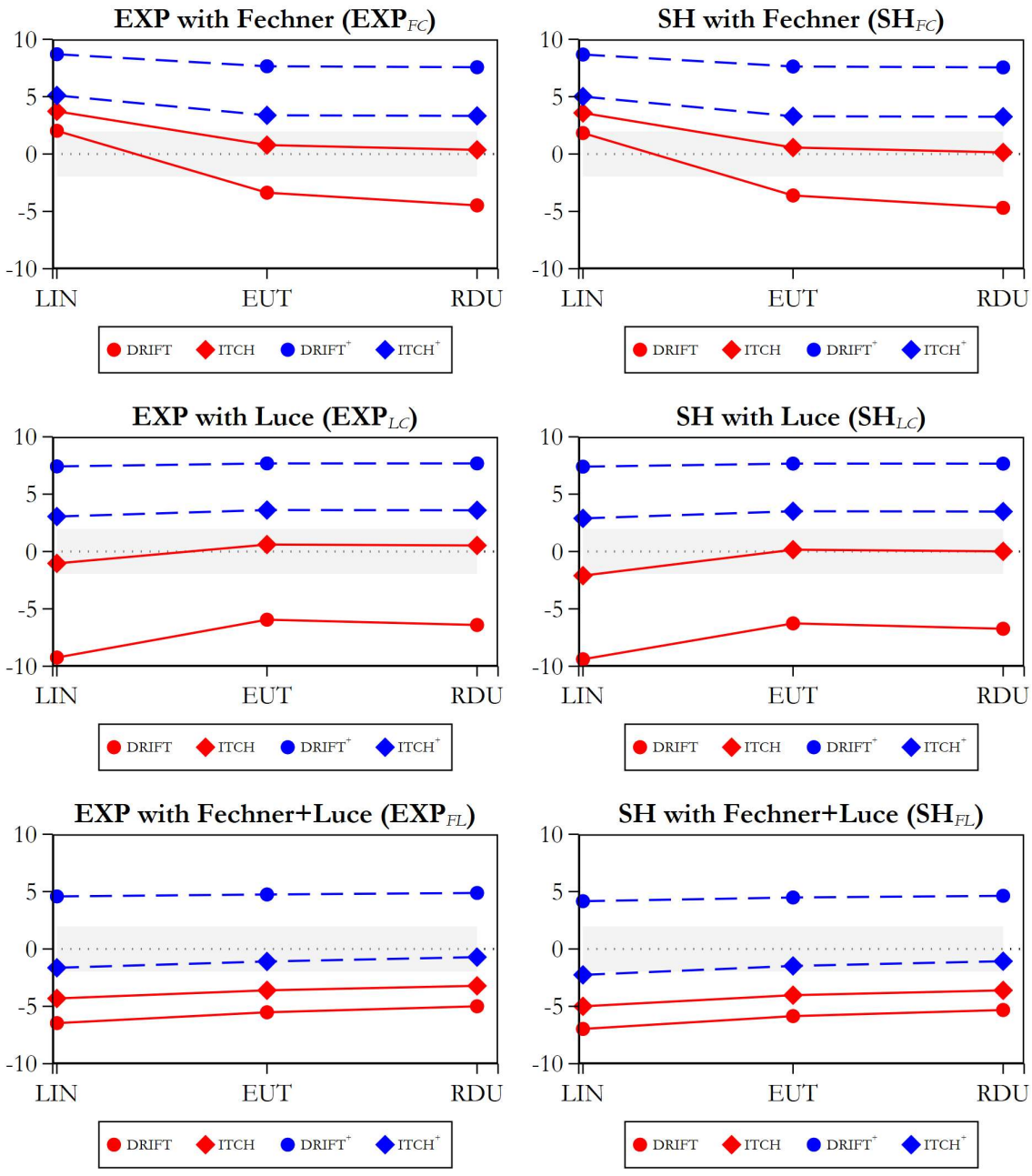
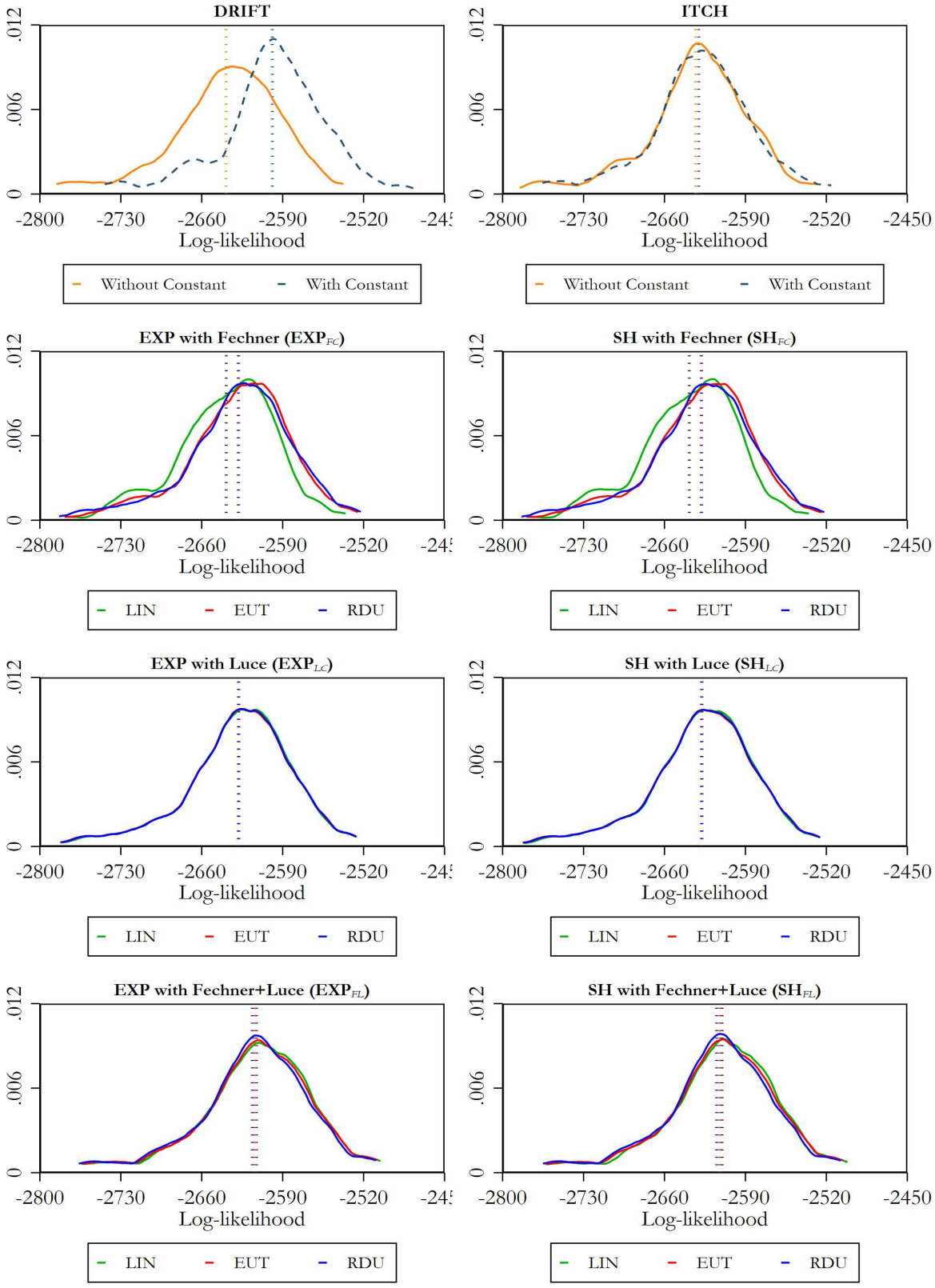


Figure 2: Distributions of Cross Validation Log-likelihoods



Appendix A: Maximum Likelihood Estimation with Luce Specification

Table A1: Estimates of Exponential Discounting with Luce Specification

Variable	Estimate	St. Error	*p*-value	95% Confidence Interval	
		A. Exponential Discounting Assuming Linear Utility			
δ	0.179	0.014	<0.001	0.152	0.206
μ_{LC}	0.082	0.005	<0.001	0.073	0.092
		B. EUT + Exponential Discounting			
r	0.414	0.027	<0.001	0.361	0.468
δ	0.118	0.009	<0.001	0.100	0.136
υ_{LC}	0.232	0.014	<0.001	0.205	0.258
μ_{LC}	0.046	0.004	<0.001	0.039	0.053
		C. RDU + Exponential Discounting			
r	0.336	0.023	<0.001	0.291	0.382
φ	0.704	0.027	<0.001	0.651	0.758
δ	0.129	0.010	<0.001	0.109	0.148
υ_{LC}	0.201	0.009	<0.001	0.184	0.219
μ_{LC}	0.053	0.004	<0.001	0.045	0.060

Table A2: Estimates of Simple Hyperbolic Discounting with Luce Error

Variable	Estimate	St. Error	*p*-value	95% Confidence Interval	
		A. Simple Hyperbolic Discounting Assuming Linear Utility			
$\varkappa$	0.177	0.014	<0.001	0.151	0.204
μ_{LC}	0.082	0.005	<0.001	0.072	0.092
		B. EUT + Simple Hyperbolic Discounting			
r	0.414	0.027	<0.001	0.361	0.468
$\varkappa$	0.117	0.009	<0.001	0.099	0.135
υ_{LC}	0.232	0.014	<0.001	0.205	0.258
μ_{LC}	0.046	0.004	<0.001	0.039	0.053
		C. RDU + Simple Hyperbolic Discounting			
r	0.336	0.023	<0.001	0.291	0.382
φ	0.704	0.027	<0.001	0.651	0.758
$\varkappa$	0.128	0.010	<0.001	0.108	0.147
υ_{LC}	0.201	0.009	<0.001	0.184	0.219
μ_{LC}	0.053	0.004	<0.001	0.045	0.060

Appendix B: Augmenting Heuristic Specifications

Table B1: Parameter Estimates and Average Partial Effects of The Augmented Decision Heuristics

Variable	Estimate/ Avg. Partial Effect	St. Error	*p*-value	95% Confidence Interval	
		A. DRIFT: replace ΔY^{rel} *with* ΔY^{nor} (Log-likelihood = −10,521.1)			
Parameter Estimates					
α_1 (ΔY^{abs})	0.002	0.006	0.757	−0.011	0.015
α_2 (ΔY^{nor})	10.223	1.958	<0.001	6.386	14.061
α_3 (τ)	−0.149	0.014	<0.001	−0.177	−0.120
α_4 (I)	0.091	0.279	0.744	−0.456	0.638
Average Partial Effects					
α_1 (ΔY^{abs})	0.000	0.001	0.758	−0.002	0.003
α_2 (ΔY^{nor})	2.289	0.433	<0.001	1.441	3.137
α_3 (τ)	−0.033	0.003	<0.001	−0.040	−0.027
α_4 (I)	0.020	0.062	0.743	−0.102	0.143
		B. DRIFT: replace ΔY^{rel} *with* ΔY^{prt} (Log-likelihood = −10,508.8)			
Parameter Estimates					
α_1 (ΔY^{abs})	0.002	0.006	0.802	−0.011	0.014
α_2 (ΔY^{prt})	12.512	2.270	<0.001	8.062	16.961
α_3 (τ)	−0.162	0.016	<0.001	−0.193	−0.132
α_4 (I)	−0.077	0.291	0.792	−0.646	0.493
Average Partial Effects					
α_1 (ΔY^{abs})	0.000	0.001	0.802	−0.002	0.003
α_2 (ΔY^{prt})	2.795	0.499	<0.001	1.816	3.774
α_3 (τ)	−0.036	0.004	<0.001	−0.043	−0.029
α_4 (I)	−0.017	0.065	0.792	−0.144	0.110
		C. ITCH: replace ΔY^{nor} *with* ΔY^{rel} (Log-likelihood = −10,543.6)			
Parameter Estimates					
α_1 (ΔY^{abs})	0.004	0.007	0.597	−0.010	0.017
α_2 (ΔY^{rel})	−0.126	0.016	<0.001	−0.157	−0.096
α_3 (τ)	8.287	1.590	<0.001	5.170	11.403

α_4 (ΔT^{nor})	−0.033	0.064	0.601	−0.158	0.091
Average Partial Effects					
α_1 (ΔY^{abs})	0.001	0.001	0.597	−0.002	0.004
α_2 (ΔY^{rel})	1.863	0.351	<0.001	1.175	2.550
α_3 (τ)	−0.028	0.003	<0.001	−0.035	−0.022
α_4 (ΔT^{nor})	−0.007	0.014	0.602	−0.035	0.021

D. ITCH: replace ΔY^{nor} *with* ΔY^{prt}
(Log-likelihood = −10,498.3)

Parameter Estimates					
α_1 (ΔY^{abs})	0.001	0.006	0.860	−0.011	0.013
α_2 (ΔY^{nor})	12.583	2.094	<0.001	8.479	16.687
α_3 (τ)	−0.145	0.016	<0.001	−0.177	−0.114
α_4 (ΔT^{nor})	−0.082	0.065	0.208	−0.209	0.045
Average Partial Effects					
α_1 (ΔY^{abs})	0.000	0.001	0.860	−0.002	0.003
α_2 (ΔY^{nor})	2.808	0.453	<0.001	1.919	3.697
α_3 (τ)	−0.032	0.006	<0.001	−0.039	−0.026
α_4 (ΔT^{nor})	−0.018	0.014	0.208	−0.047	0.010

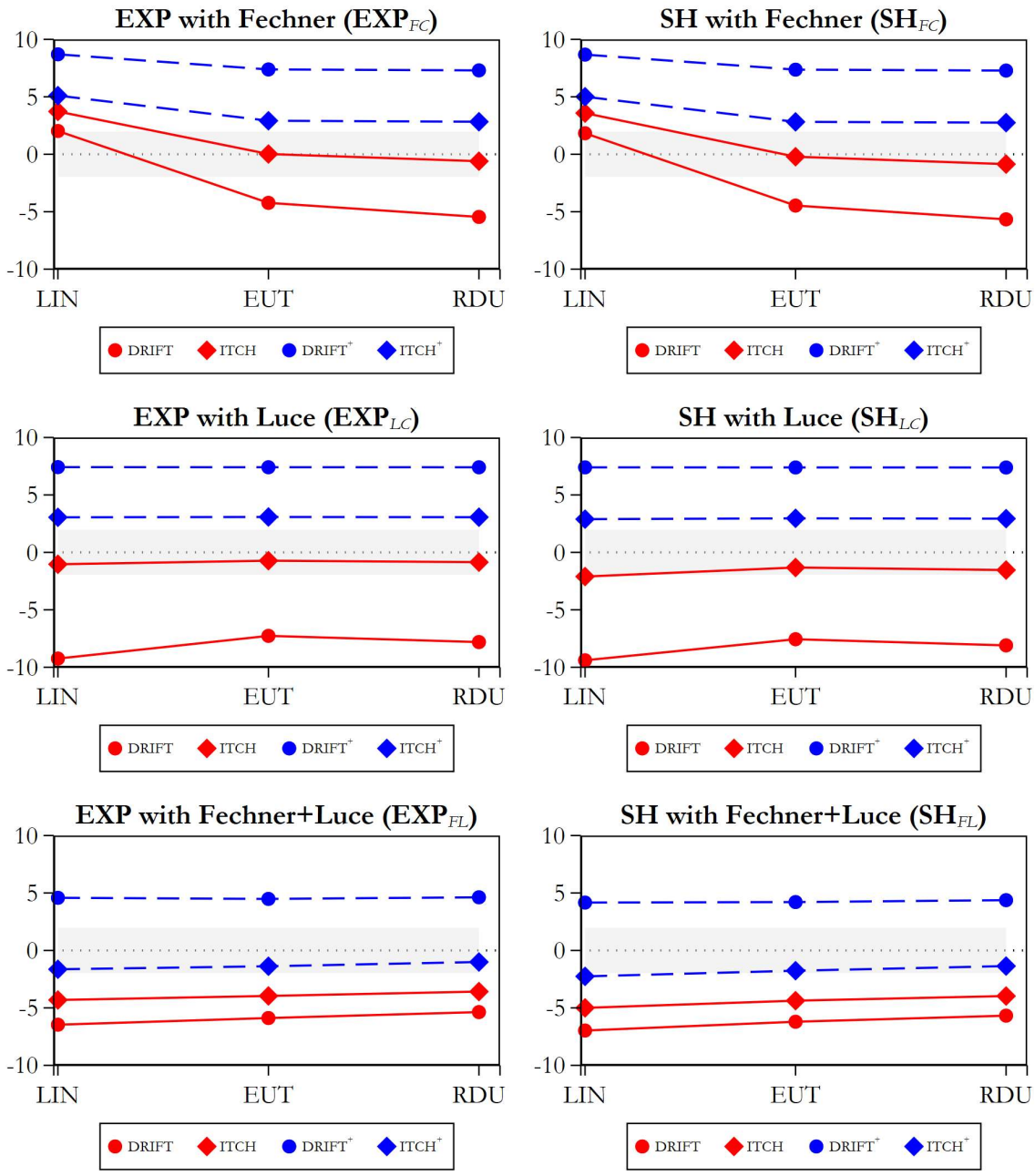
Appendix C: Augmenting Heuristic Specifications

Table C1: Bayesian Information Criteria (BIC) for Exponential and Simple Hyperbolic Discounting Models without Risk Parameters

Discounting Models	Utility	Error Models	No. of Parameters	Adj. Log-Likelihood	BIC
		A. Exponential Discounting (EXP)			
EXP	Linear	Fechner	2	−10,574	21,168
EXP	EUT	Fechner	2	−10,526	21,072
EXP	RDU	Fechner	2	−10,523	21,066
EXP	Linear	Luce	2	−10,523	21,065
EXP	EUT	Luce	2	−10,523	21,066
EXP	RDU	Luce	2	−10,523	21,066
EXP	Linear	Fechner+Luce	4	−10,458	20,955
EXP	EUT	Fechner+Luce	4	−10,462	20,963
EXP	RDU	Fechner+Luce	4	−10,469	20,977
		B. Simple Hyperbolic Discounting (SH)			
SH	Linear	Fechner	2	−10,572	21,163
SH	EUT	Fechner	2	−10,525	21,069
SH	RDU	Fechner	2	−10,522	21,063
SH	Linear	Luce	2	−10,520	21,060
SH	EUT	Luce	2	−10,522	21,063
SH	RDU	Luce	2	−10,521	21,062
SH	Linear	Fechner+Luce	4	−10,447	20,933
SH	EUT	Fechner+Luce	4	−10,455	20,949
SH	RDU	Fechner+Luce	4	−10,463	20,965

Figure C1: The Vuong's z-Statistic with Heuristic Discounting Models as Benchmarks

Against Exponential and Simple Hyperbolic Discounting Models (without Risk Parameters)



Appendix D: The Mixture of Fechner and Luce Specifications

Table D1: Estimates of Exponential Discounting with A Mixture of Fechner and Luce Specifications

Variable	Estimate	St. Error	*p*-value	95% Confidence Interval	
			A. Assuming Linear Utility		
δ	0.192	0.010	<0.001	0.172	0.212
μ_{FC}	0.740	0.503	0.141	−0.246	1.726
μ_{LC}	0.111	0.011	<0.001	0.090	0.133
π_{FC}	0.172	0.029	<0.001	0.114	0.229
π_{LC}	0.828	0.029	<0.001	0.771	0.886
			B. Assuming Non-linear Utility under EUT		
r	0.478	0.024	<0.001	0.431	0.525
δ	0.113	0.009	<0.001	0.095	0.131
υ_{FC}	0.488	0.096	<0.001	0.301	0.676
υ_{LC}	0.130	0.007	<0.001	0.117	0.143
μ_{FC}	0.161	0.100	0.109	−0.036	0.357
μ_{LC}	0.066	0.009	<0.001	0.048	0.085
π_{FC}	0.245	0.039	<0.001	0.168	0.322
π_{LC}	0.755	0.039	<0.001	0.678	0.832
			C. Assuming Non-linear Utility under RDU		
r	0.612	0.052	<0.001	0.509	0.714
φ	1.374	0.125	<0.001	1.128	1.619
δ	0.094	0.010	<0.001	0.075	0.114
υ_{FC}	0.361	0.057	<0.001	0.250	0.473
υ_{LC}	0.117	0.009	<0.001	0.100	0.134
μ_{FC}	0.159	0.051	0.002	0.059	0.259
μ_{LC}	0.069	0.014	<0.001	0.041	0.096
π_{FC}	0.366	0.066	<0.001	0.237	0.494

π_{LC}	0.634	0.066	<0.001	0.506	0.763

Table D2: Estimates of Simple Hyperbolic Discounting with A Mixture of Fechner and Luce Specifications

Variable	Estimate	St. Error	*p*-value	95% Confidence Interval	
			A. Assuming Linear Utility		
$\varkappa$	0.190	0.007	<0.001	0.175	0.204
μ_{FC}	0.639	0.347	0.065	−0.041	1.319
μ_{LC}	0.112	0.011	<0.001	0.090	0.134
π_{FC}	0.176	0.027	<0.001	0.124	0.229
π_{LC}	0.824	0.027	<0.001	0.771	0.876
			B. Assuming Non-linear Utility under EUT		
r	0.477	0.024	<0.001	0.430	0.525
$\varkappa$	0.113	0.009	<0.001	0.095	0.131
υ_{FC}	0.493	0.101	<0.001	0.295	0.691
υ_{LC}	0.130	0.007	<0.001	0.117	0.144
μ_{FC}	0.144	0.103	0.162	−0.058	0.345
μ_{LC}	0.066	0.010	<0.001	0.047	0.085
π_{FC}	0.243	0.042	<0.001	0.161	0.324
π_{LC}	0.757	0.042	<0.001	0.676	0.839
			C. Assuming Non-linear Utility under RDU		
r	0.601	0.053	<0.001	0.498	0.704
φ	1.347	0.127	<0.001	1.097	1.597
$\varkappa$	0.095	0.010	<0.001	0.075	0.116
υ_{FC}	0.369	0.061	<0.001	0.249	0.489
υ_{LC}	0.119	0.009	<0.001	0.101	0.136
μ_{FC}	0.153	0.051	0.003	0.052	0.253
μ_{LC}	0.068	0.014	<0.001	0.042	0.095
π_{FC}	0.353	0.068	<0.001	0.220	0.486
π_{LC}	0.647	0.068	<0.001	0.514	0.780

Appendix E: Quasi-hyperbolic and Weibull Hyperbolic Discounting Specifications

There is evidence that individuals exhibit extremely high discount rate when choosing over the proximate choices, but exhibit a discount rate that is relatively low and constant when choosing over the remote choices (Frederick, Loewenstein, and O'Donoghue (2002) and Harrison, Lau, and Williams [2002]). Such sharp drop in discount rates is most popularly characterized by quasi-hyperbolic (QH) structures (Elster [1979], Laibson [1997], and Phelps and Pollak [1968]), where the discount function is given by:

$$D^{QH}(t) = 1 \qquad \text{if } t = 0 \qquad \text{(A1)}$$

$$= \beta \times 1/(1+\delta)^t \qquad \text{if } t > 0$$

where $\beta < 1$ implies present bias and $\beta > 1$ implies future bias. When $\beta = 1$, the QH discount function nests EXP as a special case of no present bias. The defining characteristic of the QH specification is that the discount factor has a jump discontinuity at $t = 0$, and it is thereafter exactly the same as the EXP specification. The discount rate for the QH specification is $d^{QH}(t) = [\beta/(1+\delta)^{\eta}]^{(-1/t)} - 1$ for $t > 0$. Thus for $\beta < 1$, we observe a sharply declining discount rate in the very short run, and then the discount rate asymptotes towards δ as the effect of the initial drop in the discount factor diminishes. Using the individual-level maximum likelihood estimates in Andersen, Harrison, Lau and Rutström [2014], we set present bias parameter β in QH specification to be $\beta \in (0.7, 1.1)$.

Rather than constant after the discontinue and sudden drop at the initial time, the earliest hyperbolic specifications assumed that the discount rates continuously decline with the time horizon. One of them is a hyperbolic specification introduced by Read [2001] and Prelec [2004]:

$$D^{WEI}(t) = [1/(1+\delta)]^{t^{(1/\zeta)}} \qquad \text{(A2)}$$

where $\zeta > 0$. Ones can think of ζ as characterizing the decreasing impatience of the individual, a smoother and inverse counterpart of the notion of a present-biased preference in the QH specification. The parameter ζ can also be viewed as reflecting the "slowing down" or "speeding up"

of time as perceived by the individual. The discount function takes on the familiar shape of earlier hyperbolic specification when $\zeta > 1$. The discount rate is given by $d^{WEI}(t) = [(1/(1 + \delta))^{t^{(1/\zeta)}}]^{(1/t)} - 1$, which collapses to δ as ζ tends to 1. The results are theoretically the same if the assumption of linear utility is relaxed: allowing for concave utility functions leads to lower discount rates.

Table C1 displays maximum likelihood estimates of the Quasi-hyperbolic (QH) discounting with Fechner (FC) error specification. Assuming linear utility where the result is presented in Panel A, the estimate of parameter β is 1.005 and not statistically different from 1 (p-value = 0.639), suggesting that there is no quasi-hyperbolic discounting. The estimate of parameter δ is 0.188 (p-value < 0.001), close to the estimate of δ from EXP assuming linear utility. Similarly, we also observe that the estimated β is 1.005, which is not statistically different from 1 (p-value = 0.561) when assuming non-linear utility under EUT. The estimate of δ is also close to that of the joint estimation of EUT-EXP. The estimate of δ in the joint EUT-QH is 0.105 (p-value < 0.001).With a more risk-averse profile, we observe a lower estimate of δ, which is 0.105 (p-value < 0.001), close to that estimated in the joint EUT-EXP. The estimate of β is 1.004 and also not statistically different from 1 (p-value = 0.561).[31] With a higher estimate of the r parameter in RDU, the estimate of δ is even lower at (p-value < 0.001). The estimate of β is similar to that in EUT-QH, given by 1.003. Again, the estimate is not statistically different from 1 wit a p-value of 0.525.

The Weibull discounting model in Table C2 also collapses to the Exponential model, as the estimate of the ζ parameter is also not statistically different from 1 in each of the three estimations. The p-value on the test of the hypothesis that $\zeta = 1$ is 0.712, 0.726, and 0.730, respectively, in the estimation assuming linear utility, EUT, and RDU. Under linear utility assumption, the discount rates slightly decline from 22% in two weeks to 18.2% in one year. The 95% confidence interval for

[31] We present the ML estimation results of both structural models with Luce specification and the mixture of Fechner and Luce in Table C3-C6. We find that the results are robust across alternative stochastic models.

a two-week horizon is between 1.9% and 41.5%, while the one-year horizon is between 15.5% and 20.9%. Assuming non-linear utility under EUT, the discount rate for the two-week horizon is 11.8%, with a 95% confidence interval between 1.8% and 21.8%. Meanwhile, the discount rate for a one-year horizon is 10.1%, with a 95% confidence interval between 8.6% and 11.6%. Finally, with non-linear utility under RDU, the discount rate only declines from 8.5% in two weeks to 7.3% in one year. The 95% confidence interval for the discount rate in a two-week horizon is between 1.4% to 15.6%, while the one-year horizon is between 6.0% to 8.6%.

We now turn to the model evaluations. We start the models' evaluations with the assumption of risk neutrality and stochastic choice under Fechner specification for the Quasi-hyperbolic (QH) and Weibull (WEI) hyperbolic discounting. Similarly, all heuristic models outperform the two hyperbolic discounting models with an in-sample ranking of DRIFT$^+$ $\succ_{bic}$ ITCH$^+$ $\succ_{bic}$ ITCH $\succ_{bic}$ DRIFT $\succ_{bic}$ QH $\succ_{bic}$ WEI. As reported in Table E7, the ML estimations of QH and WEI assuming linear utility and Fechner error generate BIC scores of 21,163 and 21,165, respectively. Recall that the BIC score of DRIFT$^+$ is 20,782. Without the constant term, the BIC score is higher at 21,105. The BIC scores of ITCH$^+$ and ITCH are 20,991 and 21,057, respectively. As illustrated in the first row of Figure E1, the Vuong tests also reject the null hypothesis of equivalent fitness between two models to favor each heuristic specification with z-statistics between 2.360 and 8.991 (p-values $\leq$ 0.001). By comparing the first and the second rows of Figure E1, we observe a slightly different ranking in the out-of-sample predictions, where QH and WEI swap their positions. However, the heuristic models are still superior to the hyperbolic models. Recall that the cross-validation's log-likelihoods have a sample average of −2,599 in DRIFT$^+$ and −2,630 in ITCH$^+$. Without the constant term, the sample average of cross-validation's log-likelihoods in DRIFT is −2,639, while that of ITCH is −2,633. The means of cross-validation's log-likelihoods of WEI and QH are −2,641 and −2,642, respectively.

Assuming Luce specification, the in-sample ranking is now given by DRIFT$^+$ $\succ_{bic}$ ITCH$^+$ $\succ_{bic}$ QH $\succ_{bic}$ ITCH $\succ_{bic}$ WEI $\succ_{bic}$ DRIFT. As shown in Table E7, the BIC score of QH assuming linear utility and Luce error is 21,055, while that of WEI is 21,062. The results of the Vuong tests provided in Figure E1 show that the tests still favor DRIFT$^+$ and ITCH$^+$ with p-values $\leq$0.003 of them not being the better models. The z-statistics vary between 3.021 to 7.771. With DRIFT as the benchmark, the tests reject the null hypothesis of non-discrimination in performance to favor QH and WEI with z-statistics of −7.631 and −7.127 (p-values < 0.001), respectively. The Vuong tests with ITCH as the benchmark fall in the inconclusive region, implying statistically equivalent performance between ITCH and QH (p-value = 0.639) or WEI (p-value = 0.343). In the out-of-sample prediction, QH and WEI swap their ranking, and additionally, both models now outperform ITCH$^+$, ITCH, and DRIFT. As illustrated in the third row of Figure E1, the sample averages of cross-validation log-likelihoods are−2,630 and −2,631, respectively, for WEI and QH with risk neutrality and Luce error.

We find that the non-constant models fit data better when we move away from linear utility assumptions. However, their in-sample performance can only exceed DRIFT and are still outperformed by other heuristic specification with a ranking of DRIFT$^+$ $\succ_{bic}$ ITCH$^+$ $\succ_{bic}$ ITCH $\succ_{bic}$ QH $\succ_{bic}$ WEI $\succ_{bic}$ DRIFT. The best in-sample fit among the non-constant discounting models is given by QH assuming non-linear utility under RDU and Luce error with a BIC score of 21,062. WEI with EUT and Fechner error is the worst-fit non-discounting model in this comparison with a BIC score of 21,075. The Vuong tests also point to an improved fitness of the non-constant discounting models. The Vuong tests with DRIFT as the benchmark yield negative z-statistics and reject the null hypothesis of statistically equivalent performance to favor QH and WEI with p-values ≤ 0.010. The Vuong tests cannot statistically discriminate between ITCH and the non-constant discounting models with non-linear utility (p-values ≥ 0.127), except when we assume Luce error for

the Weibull discounting specification. The tests, however, pick ITCH over WEI with non-linear utility and Luce error by rejecting the null hypothesis of equivalent performance with p-values of 0.034. DRIFT$^+$ and ITCH$^+$ remain the better models than the non-constant discounting specifications with z-statistics ranging from 3.362 to 8.032 and p-values $\leq$ 0.001. As shown in Figure E2, the out-of-sample predictions also generate different rankings where QH and WEI now outperform ITCH and DRIFT.Additionally, both hyperbolic models swap their position in the ranking. Their sample average of cross-validation log-likelihoods varies between −2,630 and −2,631. DRIFT$^+$ and ITCH$^+$ are still better than QH and WEI in predicting choices in the out-of-sample data.

The two non-constant discounting models perform even better when we embed the models with the mixture of Fechner and Luce specifications. The in-sample and out-of-sample evaluations agree to give a ranking of DRIFT$^+$ $\succ$ WEI $\succ$ QH $\succ$ ITCH$^+$ $\succ$ ITCH $\succ$ DRIFT. Similar to EXP and SH, the performance of both models is better when we assume risk-neutrality and worst when RDU is assumed. As presented in Table E7, WEI assuming risk neutrality generates a BIC score of 20,918, the highest among any utility scenario for WEI and QH. Meanwhile, the lowest BIC score is 20,974, generated by QH with RDU. The Vuong tests still select DRIFT$^+$ as the best model by rejecting the null hypothesis of equivalent performance between the heuristic model and WEI or QH with z-statistics between 3.949 to 4.542 (p-values $\leq$ 0.001). All tests with ITCH$^+$ as the benchmark generate negative z-statistics which favor WEI and QH over ITCH. However, the magnitude of the z-statistics is not sufficiently strong to reject the null-hypothesis of different performance between ITCH and WEI or QH at a 5% significant level, except for the test that compares ITCH and WEI with risk neutrality (p-value = 0.026). The cross-validations of WEI and QH with a mixture of Gechner and Luce errors generate average log-likelihoods between −2,612 and −2,620. The highest

average log-likelihood value is generated by WEI with linear utility, while QH with RDU generates the lowest value. The log-likelihood distributions are illustrated in the bottom panel of Figure E2.

Table E1: Estimates of Quasi-hyperbolic Discounting with Fechner Error

Variable	Estimate	St. Error	*p*-value	95% Confidence Interval	
		A. Assuming Linear Utility			
δ	0.188	0.016	<0.001	0.157	0.220
β	1.005	0.011	<0.001	0.983	1.027
μ_{FC}	22.194	1.526	<0.001	19.203	25.185
H_0: $\beta = 1$, *p*-value = 0.639					
		B. Assuming Non-linear Utility under EUT			
r	0.546	0.031	<0.001	0.486	0.606
δ	0.105	0.009	<0.001	0.087	0.123
β	1.004	0.006	<0.001	0.991	1.016
υ_{FC}	0.179	0.011	<0.001	0.158	0.201
μ_{FC}	1.034	0.187	<0.001	0.667	1.401
H_0: $\beta = 1$, *p*-value = 0.561					
		C. Assuming Non-linear Utility under RDU			
r	0.792	0.046	<0.001	0.702	0.883
φ	2.040	0.104	<0.001	1.836	2.244
δ	0.076	0.008	<0.001	0.061	0.091
β	1.003	0.005	<0.001	0.994	1.012
υ_{FC}	0.231	0.014	<0.001	0.205	0.258
μ_{FC}	0.266	0.069	<0.001	0.130	0.401
H_0: $\beta = 1$, *p*-value = 0.525					

Table E2: Estimates of Weibull Hyperbolic Discounting with Fechner Error

Variable	Estimate	St. Error	*p*-value	95% Confidence Interval	
		A. Assuming Linear Utility			
δ	0.182	0.014	<0.001	0.155	0.209
ς	1.053	0.144	<0.001	0.771	1.336
μ_{FC}	22.051	1.528	<0.001	19.056	25.046
H_0: $\varsigma = 1$, *p*-value = 0.712					
		B. Assuming Non-linear Utility under EUT			
r	0.546	0.031	<0.001	0.486	0.606
δ	0.101	0.009	<0.001	0.086	0.116
ς	1.049	0.138	<0.001	0.777	1.320
υ_{FC}	0.179	0.011	<0.001	0.158	0.201
μ_{FC}	1.023	0.187	<0.001	0.663	1.394
H_0: $\varsigma = 1$, *p*-value = 0.726					
		C. Assuming Non-linear Utility under RDU			
r	0.792	0.046	<0.001	0.702	0.882
φ	2.039	0.104	<0.001	1.836	2.243
δ	0.073	0.007	<0.001	0.060	0.086
ς	1.048	0.138	<0.001	0.778	1.318
υ_{FC}	0.231	0.014	<0.001	0.204	0.258
μ_{FC}	0.265	0.069	<0.001	0.130	0.399
H_0: $\varsigma = 1$, *p*-value = 0.730					

Table E3: Estimates of Quasi-hyperbolic Discounting with Luce Error

Variable	Estimate	St. Error	*p*-value	95% Confidence Interval	
			A. Assuming Linear Utility		
δ	0.186	0.015	<0.001	0.156	0.216
β	1.008	0.011	<0.001	0.985	1.030
μ_{LC}	0.083	0.005	<0.001	0.073	0.092
H_0: $\beta = 1$, *p*-value = 0.502					
			B. Assuming Non-linear Utility under EUT		
r	0.414	0.027	<0.001	0.361	0.468
δ	0.122	0.010	<0.001	0.102	0.143
β	1.005	0.008	<0.001	0.990	1.020
υ_{LC}	0.232	0.014	<0.001	0.205	0.258
μ_{LC}	0.046	0.004	<0.001	0.039	0.054
H_0: $\beta = 1$, *p*-value = 0.498					
			C. Assuming Non-linear Utility under RDU		
r	0.336	0.023	<0.001	0.291	0.382
φ	0.704	0.027	<0.001	0.651	0.758
δ	0.133	0.011	<0.001	0.111	0.155
β	1.006	0.008	<0.001	0.989	1.022
υ_{LC}	0.201	0.009	<0.001	0.184	0.219
μ_{LC}	0.053	0.004	<0.001	0.045	0.061
H_0: $\beta = 1$, *p*-value = 0.499					

Table E4: Estimates of Weibull Hyperbolic Discounting with Luce Error

Variable	Estimate	St. Error	*p*-value	95% Confidence Interval	
			A. Assuming Linear Utility		
δ	0.178	0.013	<0.001	0.152	0.204
ς	1.047	0.139	<0.001	0.775	1.319
μ_{LC}	0.082	0.005	<0.001	0.072	0.092
H_0: $\varsigma = 1$, *p*-value = 0.735					
			B. Assuming Non-linear Utility under EUT		
r	0.414	0.027	<0.001	0.361	0.468
δ	0.117	0.009	<0.001	0.100	0.135
ς	1.047	0.138	<0.001	0.775	1.318
υ_{LC}	0.232	0.014	<0.001	0.205	0.258
μ_{LC}	0.046	0.004	<0.001	0.039	0.053
H_0: $\varsigma = 1$, *p*-value = 0.737					
			C. Assuming Non-linear Utility under RDU		
r	0.336	0.023	<0.001	0.291	0.382
φ	0.704	0.027	<0.001	0.651	0.758
δ	0.128	0.010	<0.001	0.109	0.147
ς	1.047	0.139	<0.001	0.775	1.318
υ_{LC}	0.201	0.009	<0.001	0.184	0.219
μ_{LC}	0.053	0.004	<0.001	0.045	0.060
H_0: $\varsigma = 1$, *p*-value = 0.736					

Table E5: Estimates of Quasi-hyperbolic Discounting with A Mixture of Fechner and Luce Specifications

Variable	Estimate	St. Error	*p*-value	95% Confidence Interval	
			A. Assuming Linear Utility		
δ	0.189	0.012	<0.001	0.166	0.213
β	0.998	0.004	<0.001	0.991	1.006
μ_{FC}	0.769	0.383	0.044	0.019	1.519
μ_{LC}	0.112	0.012	<0.001	0.090	0.135
π_{FC}	0.175	0.029	<0.001	0.119	0.232
π_{LC}	0.825	0.029	<0.001	0.768	0.881
H_0: $\beta = 1$, *p*-value = 0.667					
			B. Assuming Non-linear Utility under EUT		
r	0.478	0.024	<0.001	0.431	0.525
δ	0.113	0.009	<0.001	0.095	0.130
β	0.999	0.003	<0.001	0.993	1.005
υ_{FC}	0.498	0.108	<0.001	0.287	0.710
υ_{LC}	0.131	0.007	<0.001	0.117	0.145
μ_{FC}	0.143	0.122	0.244	−0.097	0.382
μ_{LC}	0.066	0.010	<0.001	0.047	0.085
π_{FC}	0.241	0.046	<0.001	0.151	0.330
π_{LC}	0.759	0.046	<0.001	0.670	0.849
H_0: $\beta = 1$, *p*-value = 0.762					
			C. Assuming Non-linear Utility under RDU		
r	0.611	0.055	<0.001	0.503	0.719
φ	1.372	0.136	<0.001	1.107	1.638
δ	0.094	0.010	<0.001	0.075	0.114
β	1.000	0.004	<0.001	0.992	1.008
υ_{FC}	0.362	0.062	<0.001	0.240	0.484
υ_{LC}	0.117	0.009	<0.001	0.099	0.136

μ_{FC}	0.159	0.059	0.008	0.042	0.275
μ_{LC}	0.069	0.014	<0.001	0.041	0.096
π_{FC}	0.365	0.076	<0.001	0.215	0.514
π_{LC}	0.635	0.076	<0.001	0.486	0.785

H_0: $\beta = 1$, p-value = 0.982

Table E6: Estimates of Weibull Hyperbolic Discounting with A Mixture of Fechner and Luce Specifications

Variable	Estimate	St. Error	*p*-value	95% Confidence Interval	
			A. Assuming Linear Utility		
δ	0.180	0.019	<0.001	0.143	0.218
ς	1.151	0.107	<0.001	0.941	1.361
μ_{FC}	1.032	0.611	0.091	−0.165	2.228
μ_{LC}	0.120	0.015	<0.001	0.091	0.148
π_{FC}	0.196	0.040	<0.001	0.118	0.274
π_{LC}	0.804	0.040	<0.001	0.726	0.882
H_0: $\varsigma = 1$, *p*-value = 0.160					
			B. Assuming Non-linear Utility under EUT		
r	0.478	0.024	<0.001	0.431	0.526
δ	0.108	0.010	<0.001	0.089	0.128
ς	1.166	0.093	<0.001	0.984	1.348
υ_{FC}	0.489	0.087	<0.001	0.318	0.660
υ_{LC}	0.130	0.006	<0.001	0.118	0.143
μ_{FC}	0.129	0.061	0.036	0.008	0.249
μ_{LC}	0.067	0.009	<0.001	0.049	0.086
π_{FC}	0.245	0.032	<0.001	0.182	0.307
π_{LC}	0.755	0.032	<0.001	0.693	0.818
H_0: $\varsigma = 1$, *p*-value = 0.074					
			C. Assuming Non-linear Utility under RDU		
r	0.594	0.050	<0.001	0.495	0.692
φ	1.325	0.121	<0.001	1.089	1.562
δ	0.093	0.010	<0.001	0.074	0.112
ς	1.174	0.097	<0.001	0.984	1.363
υ_{FC}	0.385	0.063	<0.001	0.261	0.508
υ_{LC}	0.120	0.009	<0.001	0.103	0.137

μ_{FC}	0.128	0.048	0.007	0.035	0.221
μ_{LC}	0.068	0.013	<0.001	0.042	0.094
π_{FC}	0.335	0.063	<0.001	0.213	0.458
π_{LC}	0.665	0.063	<0.001	0.542	0.787

H_0: $\varsigma = 1$, p-value = 0.072

Table E7: The Ranking of Bayesian Information Criteria (BIC) from Best to Worst

Rank	Discounting Models	Utility	Error Models	No. of Parameters	(Adj.) Log-likelihood	BIC
1	DRIFT^{+}			5	−10,376	20,801
2	WEI	Linear	Fechner+Luce	5	−10,444	20,936
3	WEI	EUT	Fechner+Luce	6	−10,446	20,950
4	QH	Linear	Fechner+Luce	5	−10,457	20,962
5	WEI	RDU	Fechner+Luce	6	−10,452	20,963
6	QH	EUT	Fechner+Luce	6	−10,461	20,981
7	QH	RDU	Fechner+Luce	6	−10,469	20,996
8	ITCH^{+}			5	−10,481	21,010
9	QH	Linear	Luce	3	−10,518	21,066
10	ITCH			4	−10,516	21,072
11	WEI	Linear	Luce	3	−10,522	21,073
12	QH	RDU	Luce	4	−10,519	21,076
13	QH	EUT	Luce	4	−10,519	21,077
14	QH	RDU	Fechner	4	−10,519	21,078
15	WEI	RDU	Fechner	4	−10,522	21,084
16	WEI	RDU	Luce	4	−10,522	21,084
17	WEI	EUT	Luce	4	−10,523	21,085
18	QH	EUT	Fechner	4	−10,523	21,085
19	WEI	EUT	Fechner	4	−10,526	21,090
20	DRIFT			4	−10,540	21,120
21	QH	Linear	Fechner	3	−10,572	21,174
22	WEI	Linear	Fechner	3	−10,573	21,176

Table E8: Bayesian Information Criteria (BIC) for Quasi-hyperbolic and Weibull Hyperbolic Discounting Models without Risk Parameters

Discounting Models	Utility	Error Models	No. of Parameters	Adj. Log-Likelihood	BIC
		A. Quasi-hyperbolic Discounting (QH)			
QH	Linear	Fechner	3	−10,572	21,174
QH	EUT	Fechner	3	−10,523	21,075
QH	RDU	Fechner	3	−10,519	21,068
QH	Linear	Luce	3	−10,518	21,066
QH	EUT	Luce	3	−10,519	21,067
QH	RDU	Luce	3	−10,519	21,067
QH	Linear	Fechner+Luce	5	−10,457	20,962
QH	EUT	Fechner+Luce	5	−10,461	20,971
QH	RDU	Fechner+Luce	5	−10,469	20,986
		B. Weibull Hyperbolic Discounting (WEI)			
WEI	Linear	Fechner	3	−10,573	21,176
WEI	EUT	Fechner	3	−10,526	21,080
WEI	RDU	Fechner	3	−10,522	21,074
WEI	Linear	Luce	3	−10,522	21,073
WEI	EUT	Luce	3	−10,523	21,075
WEI	RDU	Luce	3	−10,522	21,074
WEI	Linear	Fechner+Luce	5	−10,444	20,936
WEI	EUT	Fechner+Luce	5	−10,446	20,940
WEI	RDU	Fechner+Luce	5	−10,452	20,954

Figure E1: The Vuong's z-Statistics with Heuristic Discounting Models as Benchmarks

Against Non-constant Discounting Models

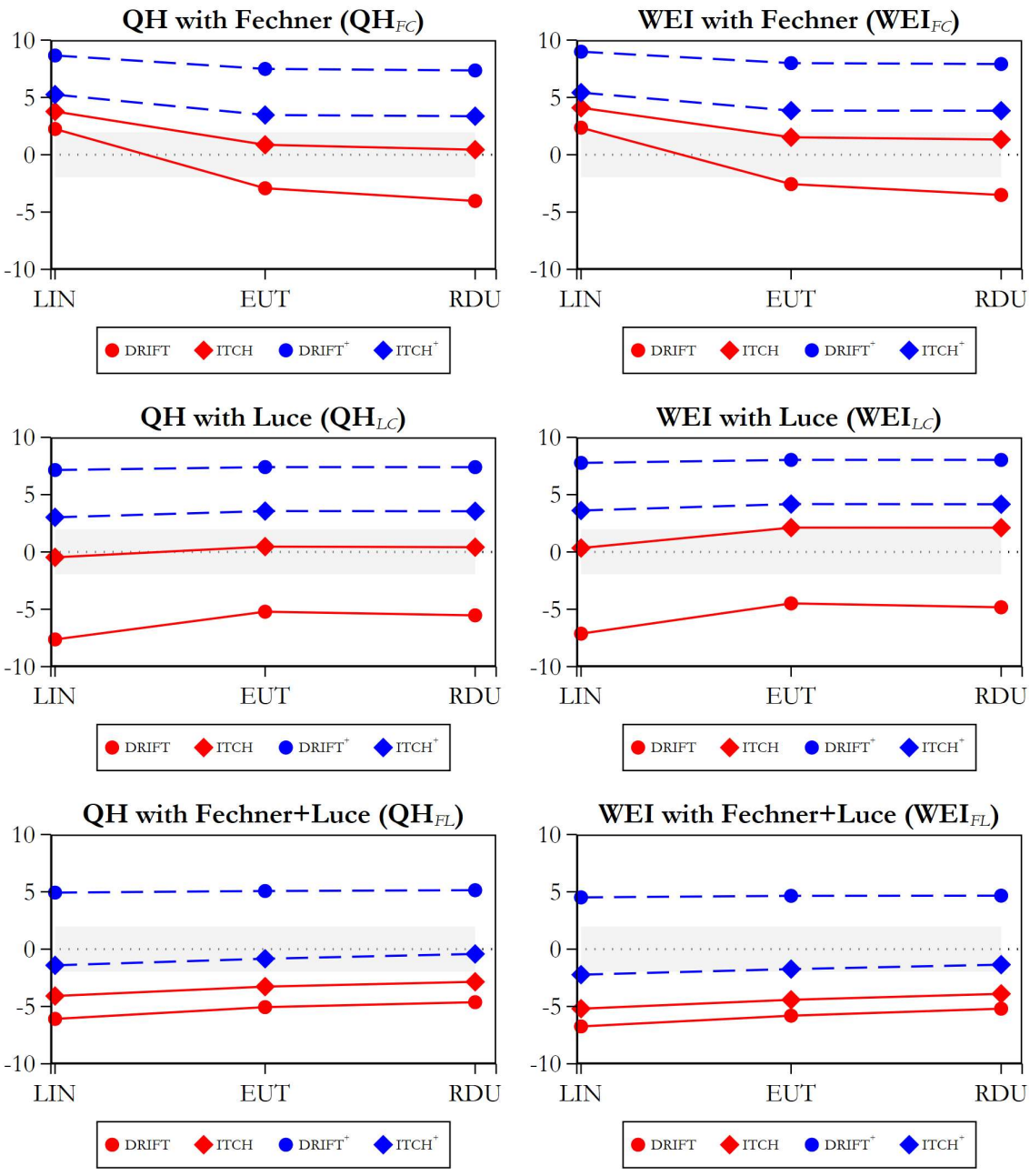


Figure E2: Distributions of Cross Validation Log-likelihoods (Heuristics vs Non-constant Discounting Models)

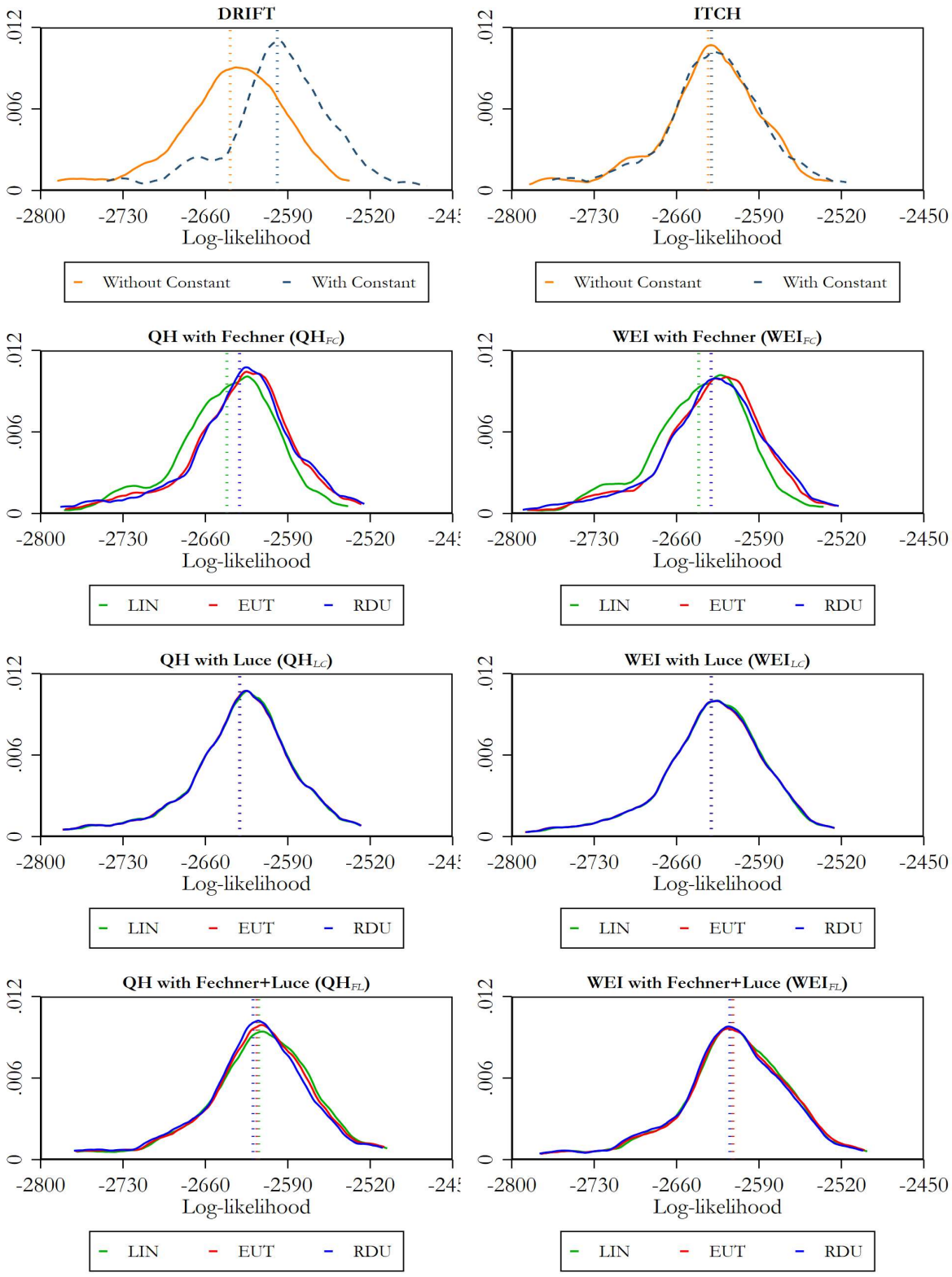
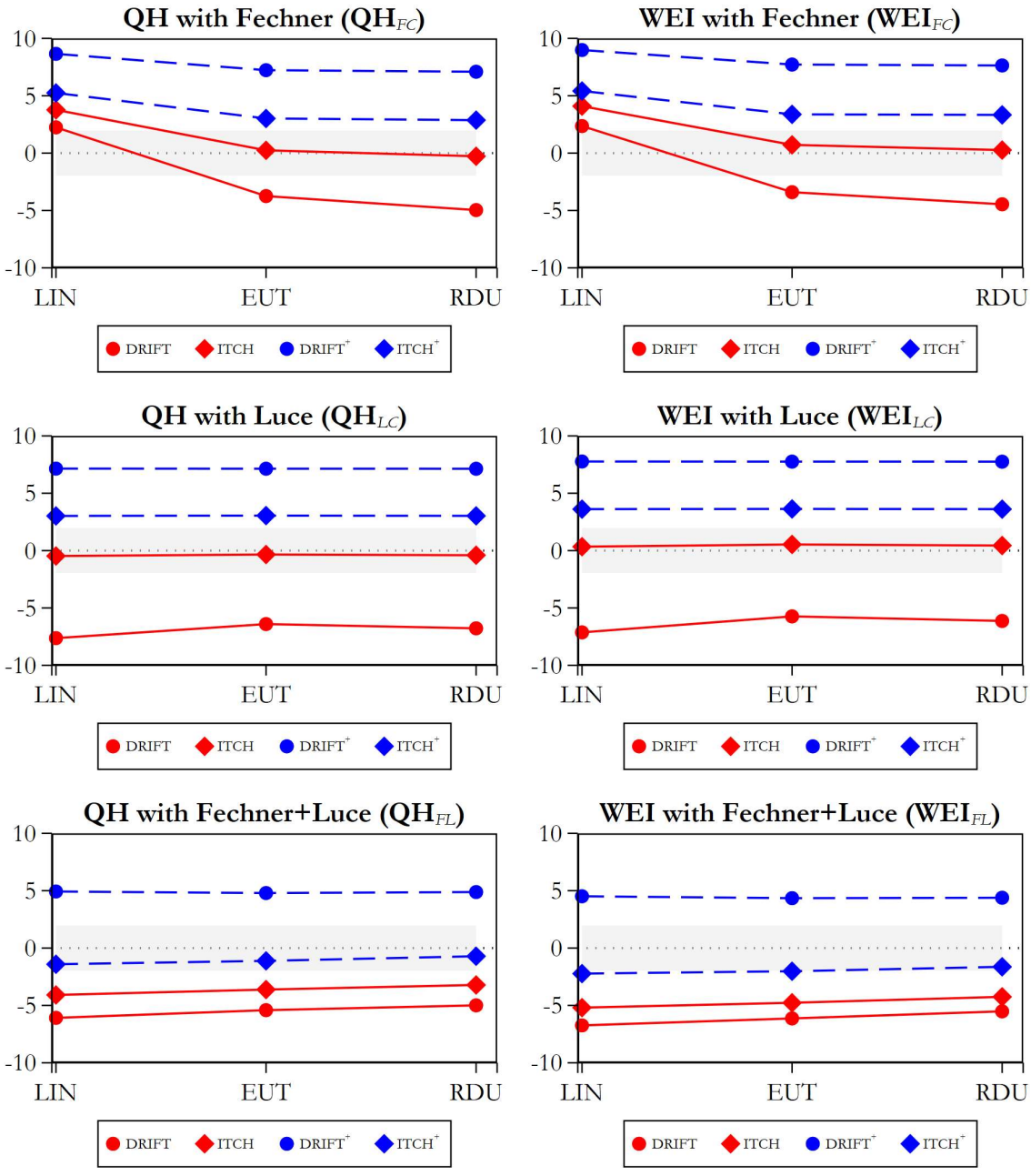


Figure E3: The Vuong's z-Statistics with Heuristic Discounting Models as Benchmarks

Against Non-constant Discounting Models (without Risk Parameters)



Appendix F: A Non-Parametric Analysis of the Goodness-of-fit of DRIFT$^+$

Table F1 deciphers the discounting models' out-of-sample performance on predicting choices of smaller-sooner (SS) and larger-later (LL) payments. All discounting models generally have a better performance in predicting the larger-later (LL) payments, as reflected in the higher means of cross-validation log-likelihoods compared to the smaller-sooner (SS) payments. We find that the performance of DRIFT$^+$ is supported by its superiority in predicting the choices of the smaller-sooner (SS) payment. With the constant term, the heuristic model generates a cross-validation log-likelihood of −1,347. However, DRIFT$^+$ is the second-worst model in predicting the larger-later (LL) choice, with a cross-validation log-likelihood of −1,252.

Figure F1 and F2 illustrate the probability of choosing LL with a given interest rate estimated by the heuristic and structural discounting models. The probabilities are calculated by applying the estimated parameters to the in-sample data. Recall from the estimation results presented in Table 3, the choice probability is significantly affected by the negative constant term, the compound annual interest, and the relative difference in monetary payments. The latter two variables have a positive estimate. The coefficient for the time horizon variable t in DRIFT$^+$ is statistically insignificant. The estimated coefficients imply that if we held interest rate constant while the time horizon t increases, DRIFT$^+$ predicts increasing probability of choosing LL. Then, the difference between DRIFT$^+$ and other models is that, in DRIFT$^+$, the probability of choosing LL always increases with the time horizon t. The other discounting models predict a decreasing probability of choosing LL when the offered interest rate is below 20%. The probability of choosing LL is estimated to increase with the time horizon when the offered interest rate is 20% or above. The estimated discount rates by the structural discounting models imply that the switching points from choosing SS to LL are roughly around 15-20% interest rates. Stated equivalently, the

representative agent chooses SS when the offered interest rate is 15% or lower and chooses LL when the offered interest rate is 20% or higher.

Figure F3 maps the proportion of observed choice LL for a given interest rate. For any offered interest rates, the proportion of observed choice LL always increases with the time horizon. Comparing Figure F3 to Figure F1 and F2, the predicted probability of choosing LL in DRIFT^{+} is in line with the choice patterns, especially in the tasks that offered an interest rate of 15% or lower. As in those tasks, the representative agent is more likely to choose SS; the conformity between the predicted choice probability and the choice pattern may explain the superior performance of DRIFT^{+}.

Table F1: The Ranking of Out-of-Sample Prediction
(The Top and Bottom Three)

Ranking	Discounting Models	Utility	Error Models	Cross-Validation Log-Likelihood
		A. Smaller-Sooner Payment		
1	DRIFT+			−1,347
2	ITCH+			−1,361
3	WEI	Linear	Fechner+Luce	−1,401
38	QH	EUT	Luce	−1,444
39	QH	RDU	Fechner	−1,446
40	DRIFT			−1,446
		B. Larger-Later Payment		
1	EXP	RDU	Fechner+Luce	−1,183
2	QH	RDU	Fechner+Luce	−1,184
3	EXP	EUT	Fechner+Luce	−1,184
38	WEI	Linear	Fechner	−1,236
39	DRIFT+			−1,252
40	ITCH+			−1,268

Figure F1: Estimated Probability of Choosing Larger Later Option by The Heuristic Discounting Models

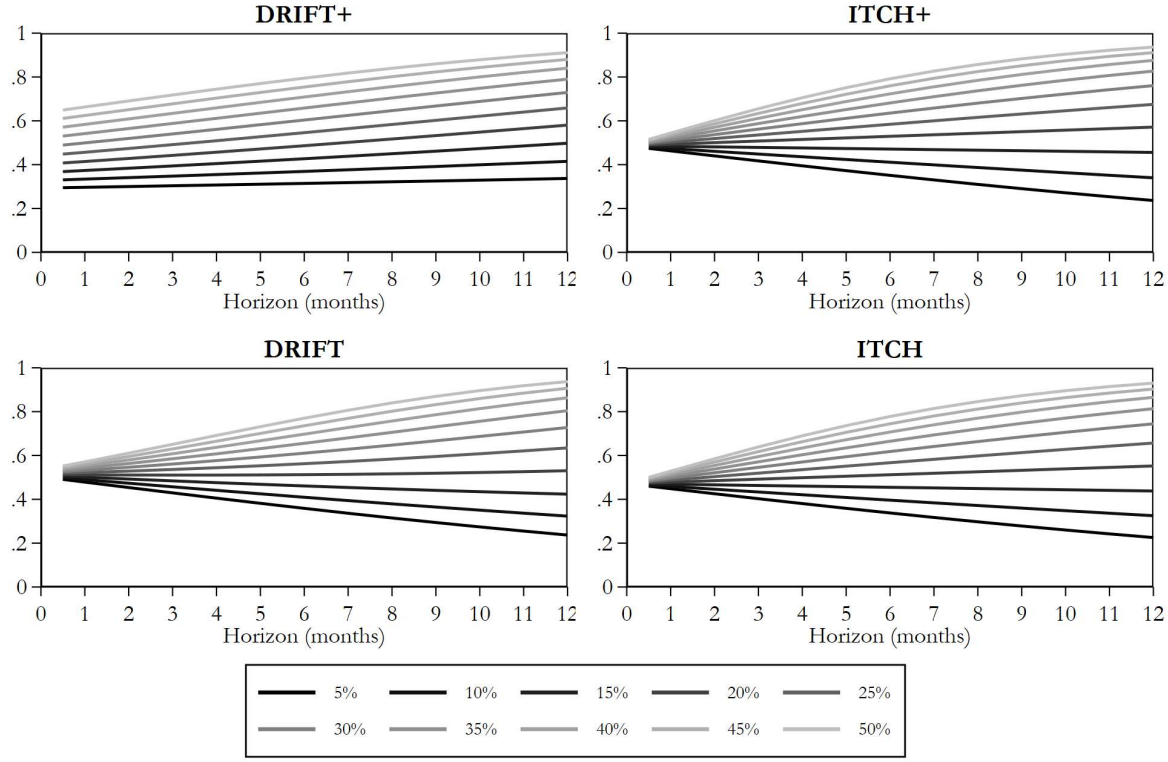


Figure F2: Estimated Probability of Choosing Larger-Later Option by The Structural Discounting Models

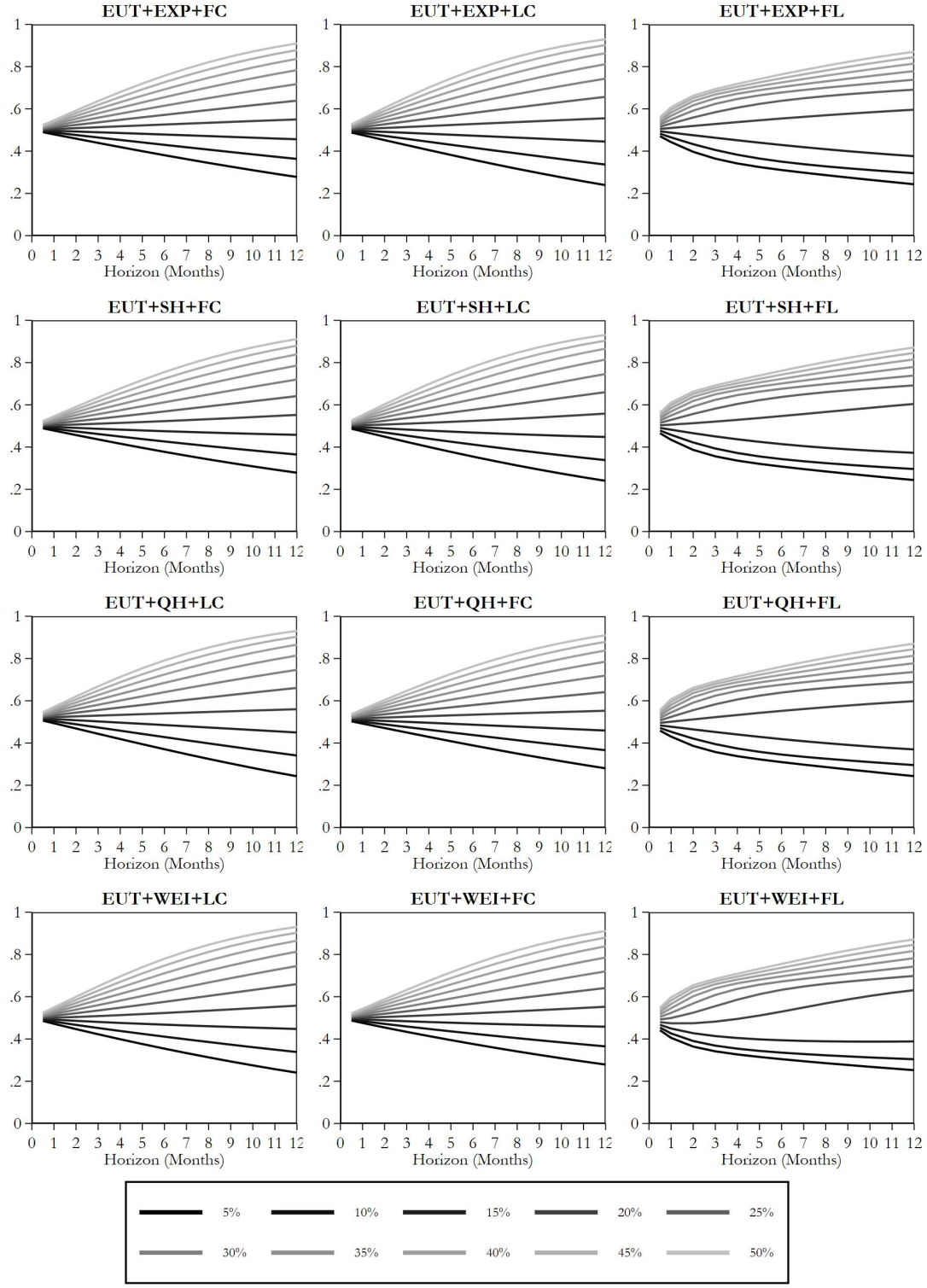
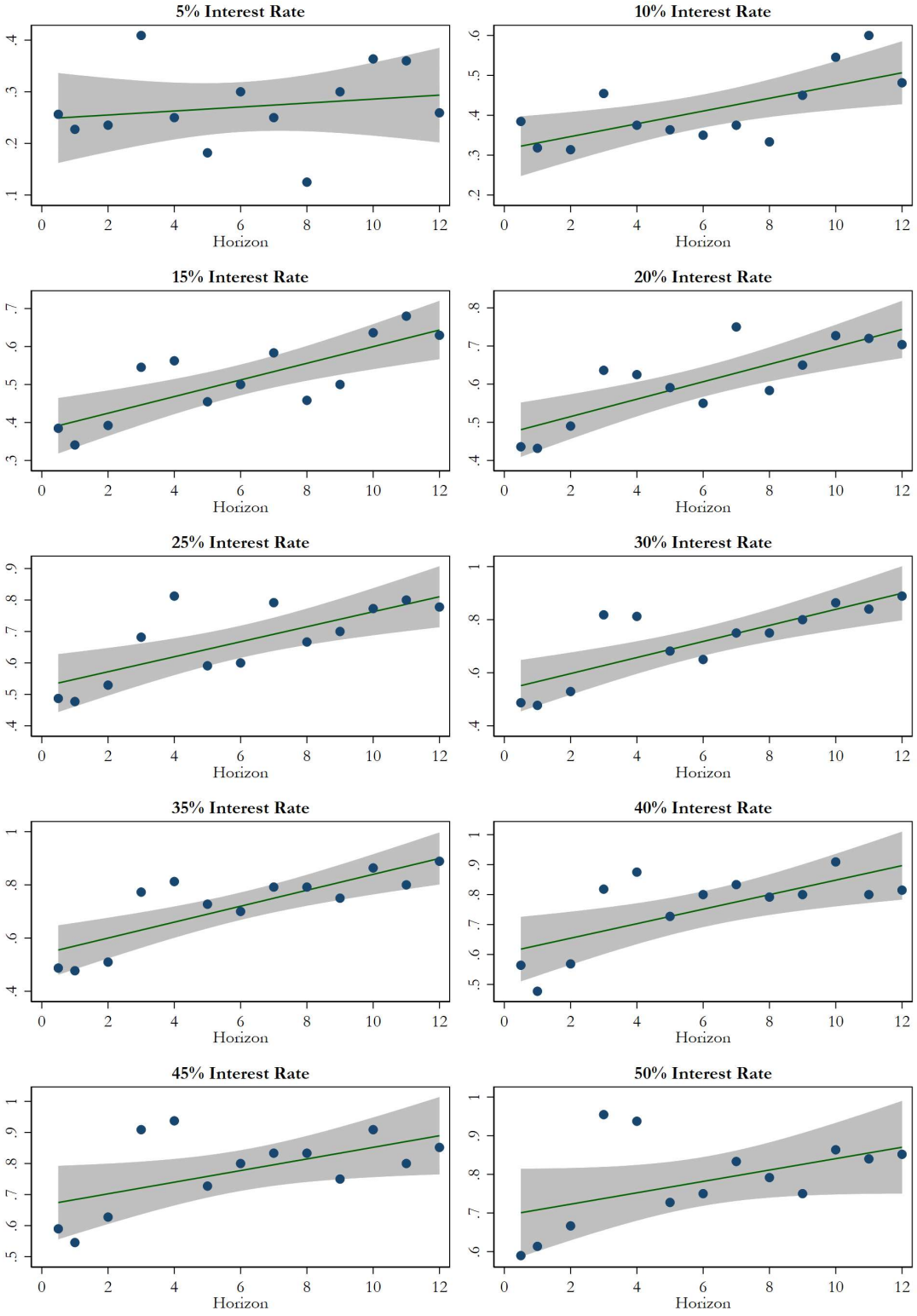


Figure F3: The Proportion of Obversed Choice of the Larger-Later Option



Concluding Remarks

This thesis discusses two topics related to behavioral errors in risk and time preferences. In the context of the decision making under risk, we examine the relationship between the task complexity and choice randomness and analyze how the relationship is better specified. On the subject of intertemporal choice, we critically evaluate the decision heuristic specifications that appeared to be related to the choice of behavioral error stories. The results of these analyses reveal a multitude of observations.

The first essay empirically evaluates the performance of three stochastic models, which consider the task complexity of the source of choice randomness, when they are embedded in the deterministic models of decision making under risks. The three error models, namely Contextual Utility, Decision Field Theory, and Entropy model, differ in measuring the task complexity and incorporating it into their heteroskedastic specification. We use the homoskedastic specification of Fechner error as the benchmark in the model evaluations. We find that the inferred risk attitudes under the Expected Utility theory are robust across alternative stochastic models, whereas the Rank-Dependent Utility theory results show some sensitivity. Our results highlight the need to accommodate the effects of task complexity on behavioral errors by rejecting homoskedasticity to favor at least two heteroskedastic models. Under the Expected Utility theory, we find that an inverse-U function better represents the relationship between task complexity and behavioral errors. It implies that behavioral errors increase up to moderate levels of task complexity but then decrease as the complexity levels continue to increase because the decision-maker devotes more attention to the decision tasks and deliberates her choices more thoroughly. Under the Rank-Dependent Utility, the monotonic function postulated by Decision Field Theory and Contextual Utility is better in characterizing the relationship between the task complexity and behavioral errors.

We investigate the rationale of the models' ranking in goodness-of-fit by looking into the components of the index of utility difference that enters the probit link function. Our analytical results show that the ability of a stochastic model to capture a more diverse behavior at an individual level explains its superior performance. With more heterogeneous individuals, the resulting distributions of expected utility and rank utility differences are more dispersed and larger in their absolute values. It implies that the more superior stochastic model assumes a more contrast preference between two options, which is translated into a higher probability of choosing a particular option unless the decision maker is indifferent. The higher choice probabilities then lead to higher log-likelihood values and, accordingly, a better performance. We also find that a more diverse individual-level parameter that controls the curvature of the probability weighting function contributes to the much-improved performance of the Decision Field Theory. Specifically, the higher estimated probability weighting parameter lowers its standard deviation of behavioral errors at decision tasks where the individuals have a clearer preference toward a particular option. The result is in line with the finding of less noisy choices at points far from indifference between two options.

In the second essay, we discuss the studies that claim a better accuracy of the decision heuristic models in predicting the individual's time preferences relative to the standard parametric discounting specifications based on a hypothetical experiment. Using the data from an incentivized field experiment, which provided monetary rewards for responses made by the subject to encourage effort, we find evidence supporting the claim. However, our further analyses find that the performance superiority of heuristic models is more due to the artefact of their oversimplified econometrics modeling than a specific ability to capture particular characteristics of discounting behavior. Using a more advanced econometrics analysis, we are able to reverse the inference and, at the same time, highlight two important issues.

The first issue is the importance of the appropriate choice of behavioral error stories. The heuristic specifications, which combine the level and proportional differences for their intra-attribute evaluations, exhibit a linear approximation of a finite mixture of two decision rules advocated by Fechner and Luce errors. When evaluating them against the structural discounting models, it is then as if we compare a discrete representation of unobserved heterogeneity in decision rules with a representative agent model. When we comparably apply a mixture specification of the two behavioral errors to the structural discounting models, the relative performance of heuristic models is completely overturned. Nonetheless, one can still use the heuristic discounting models as a simple diagnostic tool for choosing between the two stochastic specifications.

The second issue is the significance of accounting for utility curvature. While plentiful experimental literatures show the benefit to allow for non-linear utility function to correct the bias in the estimated discount rates, the studies of decision heuristics in intertemporal choices assumes risk neutrality. The experiment data that we use provides us the opportunity to allow for a non-linear utility function in the discount rates elicitation by jointly estimating risk attitudes and discounting behaviors. As we control for the effects of utility curvature in the estimation of structural discounting models, we observe a diminished relative performance of heuristic discounting models.

Overall, both essays highlight the crucial role played by behavioral error stories in the elicitation of decision making behaviors which has important implications for policymaking and future research. The first essay demonstrates how the choice of behavioral error stories may affect the ability of the deterministic models of risk attitude in capturing a more diverse behavior, which is useful for distinguishing decision makers in the population. With the heterogeneous decision makers, who are identified through their specific risk attitudes, a policy that targets particular individuals can then be effectively formulated. The second essay illustrates the impact of the choice of behavioral error specifications on the ability of deterministic intertemporal choice models in

explaining the observed discounting behaviors. For academic researchers, the flexible form of heuristic models which allows for choosing between two behavioral error stories challenges future research to put more effort into proposing alternative stochastic specifications for intertemporal choice analyses. The second essay also re-emphasizes the need to account for utility curvature in eliciting individual discount rates for the design of a policy. It is done by showing how the structural discounting models that account for a non-linear utility function are generally better at explaining the observed choice behavior of a decision maker than the heuristic models.

References

Chapter 1

Agranov, Marina, and Pietro Ortoleva, "Stochastic Choice and Preferences for Randomization," *Journal of Political Economy*, 125(1), 2017, 40-68.

Agranov, Marina, and Pietro Ortoleva, "Ranges of Preferences and Randomization," *Mimeo*, Princeton University, 2020.

Akaike, Hirotogu, *Information Theory and an Extension of the Maximum Likelihood Principle*, In B. N. Petrov & F. Caski (Eds.), Proceedings of the Second International Symposium on Information Theory (Budapest: Akademiai Kiado, 1973).

Alekseev, Aleksandr G., Glenn W. Harrison, Morten I. Lau,, and Don Ross, "Deciphering the Noise: The Welfare Costs of Noisy Behavior," Center for Economic Analysis of Risk (CEAR) Working paper No. 2018-01, 2019.

Allenby, Greg M., and Peter E. Rossi, "Marketing Models of Consumer Heterogeneity," *Journal of Econometrics*, 89(1–2), 1998, 57-78.

Alós-Ferrer, Carlos, and Michele Garagnani, "Choice Consistency and Strength of Preference," *Economics Letters*, 198, 2021, 109672.

Andersen, Steffen, Glenn W. Harrison, Morten I. Lau, and E. Elisabet Rutström, "Elicitation Using Multiple Price List Formats" *Experimental Economics*, 9, 2006, 383-405.

Andersen, Steffen, Glenn W. Harrison, Morten I. Lau, and E. Elisabet Rutström, "Eliciting Risk and Time Preferences," *Econometrica*, 76(3), 2008, 583-618.

Andersen, Steffen, Glenn W. Harrison, Morten I. Lau, and E. Elisabet Rutström, "Behavioral Econometrics for Psychologists" *Journal of Economic Psychology*, 31(4), 2010, 15-33.

Andersen, Steffen, Glenn W. Harrison, Morten I. Lau, and E. Elisabet Rutström, "Discounting Behavior: A Reconsideration" *European Economic Review*, 71(1), 2014, 15-33.

Ballinger, T. P., and Nathaniel T. Wilcox, "Decisions, Error and Heterogeneity," *The Economic Journal*, 107(443), 1997, 1090-1105.

Bhatia, Sudeep, and Timothy L. Mullett, "Similarity and Decision Time in Preferential Choice," *Quarterly Journal of Experimental Psychology*, 71(6), 2018, 1276–1280.

Blavatskyy, Pavlo R., "Stochastic Expected Utility Theory," *Journal of Risk and Uncertainty*, 34, 2007, 259-286.

Blavatskyy, Pavlo R., and Ganna Pogrebna, "Models of Stochastic Choice and Decision Theories: Why Both Are Important for Analyzing Decisions," *Journal of Applied Econometrics*, 25(6), 2009, 963-986.

Bruhin, Adrian, Helga Fehr-Duda, and Thomas Epper, "Risk and Rationality: Uncovering Heterogeneity in Probability Distortion," *Econometrica*, 78(4), 2010, 1375-1412.

Buschena, David E., and David Zilberman, "Generalized Expected Utility, Heteroscedastic Error, and Path Dependence in Risky Choice," *Journal of Risk and Uncertainty*, 20(1), 2000, 67–88.

Busemeyer, Jerome R., and James T. Townsend, "Decision Field Theory: A Dynamic-Cognitive Approach to Decision Making in an Uncertain Environment," *Psychological Review*, 100(3), 1993, 432-459.

Cheung, Stephen L., "Risk Preferences Are Not Time Preferences: On the Elicitation of Time Preference under Conditions of Risk: Comment," *American Economic Review*, 105(7), 2015, 2242-2260.

Conte, Anna, John D. Hey, and Peter G. Moffatt, "Mixture Models of Choice Under Risk," *Journal of Econometrics*, 162(1), 2011, 79–88.

Daly, Andrew, Stephen Hess, and Kenneth Train, "Assuring Finite Moments for Willingness to Pay in Random Coefficient Models," *Transportation*, 2012, 39(1), 19-31.

Dashiell, John F., "Affective Value-Distances as a Determinant of Esthetic Judgment-Times," *The American Journal of Psychology*, 50(1/4), 1937, 57-67.

Dave, Chetan, Catherine C. Eckel, Cathleen A. Johnson, and Christian Rojas, "Eliciting risk preferences: When is simple better?" *Journal of Risk and Uncertainty*, 41(3), 2010, 219-243.

Desmarais, Bruce A., and Jeffrey J. Harden, "Testing for Zero Inflation in Count Models: Bias Correction for the Vuong test," *The Stata Journal*, 13(4), 2013, 810-835.

Dickhaut, John, Vernon Smith, Baohua Xin, and Aldo Rustichini, "Human Economic Choice as Costly Information Processing," *Journal of Economic Behavior & Organization*, 94, 2013, 206-221.

Dwenger, Nadja, Dorothea Kübler, and Georg Weizsäcker, "Flipping a Coin: Evidence from University Applications," *Journal of Public Economics*, 167, 2018, 240-250.

Galarza, Francisco, "Choices under Risk in Rural Peru," MPRA Paper No. 17708, 2009. Available at https://mpra.ub.uni-muenchen.de/17708/.

Greene, William H., *Econometric Analysis* (Essex, UK: Pearson, Eight Edition, 2020).

Glöckner, Andreas, and Tilmann Betsch, "Do People Make Decisions Under Risk Based on Ignorance? An Empirical Test of the Priority Heuristic Against Cumulative Prospect Theory," *Organizational Behavior and Human Decision Processes*, 107(1), 2008, 75-95.

Harless, David W., and Colin F. Camerer, "The Predictive Utility of Generalized Expected Utility Theories," *Econometrica*, 62(6), 1994, 1251-1289.

Harrison, Glenn W., "Experimental Design and Bayesian Interpretation," in H. Kincaid and D. Ross (eds.), *Modern Guide to the Philosophy of Economics* (Cheltenham, UK: Elgar, forthcoming 2021).

Harrison, Glenn W., and Jia M. Ng, "Evaluating The Expected Welfare Gain From Insurance," *Journal of Risk and Insurance*, 83(1), 2016, 91–120.

Harrison, Glenn W., Morten I Lau, and Hong Il Yoo, "Risk Attitudes, Sample Selection and Attrition in a Longitudinal Field Experiment," *Review of Economics and Statistics*, 102(3), 2020, 552-568.

Hey, John D., "Experimental investigations of errors in decision making under risk," *European Economic Review*, 39(3-4), 1995, 633-640.

Hey, John D., "Does Repetition Improve Consistency?" *Experimental Economics*, 4(1), 2001, 5-54.

Hey, John D., and Chris Orme, "Investigating Generalizations of Expected Utility Theory Using Experimental Data," *Econometrica*, 62(6), 1994, 1291-1326.

Hey, John D., Gianna Lotito, and Anna Maffioletti, "The Descriptive and Predictive Adequacy of Theories of Decision Making under Uncertainty/Ambiguity," *Journal of Risk and Uncertainty*, 41(2), 2010, 81–111.

Lau, Morten I., Hong Il Yoo, and Hongming Zhao, "The Reflection Effect and Fourfold Pattern of Risk Attitudes: A Structural Econometric Analysis ," *Mimeo*, 2019, Available at SSRN: https://ssrn.com/abstract=3458881.

Lau, Morten I., Hong Il Yoo, and Hongming Zhao, "Temporal Stability of Cumulative Prospect Theory," in G. W. Harrison & D. Ross (eds.), *Prospect Theory as a Model of Risky Choice: Descriptive and Normative Assessments*. (Bingley, UK: Emerald, Research in Experimental Economics, 2021).

Lau, Morten I., and Hong Il Yoo, "Structural Estimation of Higher Order Risk Preferences," Center for Economic Analysis of Risk (CEAR) Working paper No. 2021-09, 2021.

Leland, Jonathan W., "Generalized Similarity Judgments: An Alternative Explanation for Choice Anomalies," *Journal of Risk and Uncertainty*, 9(2), 1994, 151-172.

Loomes, Graham, and Robert Sugden, "Incorporating A Stochastic Element Into Decision Theories," *European Economic Review*, 39(3-4), 1995, 641–648.

Loomes, Graham, Peter G. Moffatt, and Robert Sugden, "A Microeconometric Test of Alternative Stochastic Theories of Risky Choice," *Journal of Risk and Uncertainty*, 24(2), 2002, 103-130.

Luce, Robert D., "A Choice Theory Analysis of Similarity Judgments," *Psychometrika*, 26(2), June 1961, 151–163.

Moffat, Peter G., "Stochastic Choice and the Allocation of Cognitive Effort," *Experimental Economics*, 8(4), December 2005, 369-388.

Monroe, Brian A., "The Statistical Power of Individual-level Risk Preference Estimation," *Journal of the Economic Science Association*, 6, November 2020, 168–188.

Nosofsky, Robert M., "Choice, Similarity, and the Context Theory of Classification," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10(1), January 1984, 104-114.

Nosofsky, Robert M., "Relations between Exemplar-Similarity and Likelihood Models of Classification," *Journal of Mathematical Psychology*, 34(4), December 1990, 393-418.

Prelec, Drazen, "The Probability Weighting Function," *Econometrica*, 66(3), May 1998, 497-527.

Quiggin, John, "A Theory of Anticipated Utility," *Journal of Economic Behavior & Organization*, 3(4), December 1982, 323-343.

Revelt, David, and Kenneth Train, "Customer-Specific Taste Parameters and Mixed Logit: Households' Choice of Electricity Supplier," *Working paper, Department of Economics, University of California, Berkeley*, 2000.

Rubinstein, Ariel, "Similarity and Decision Making Under Risk: Is There a Utility Theory Resolution to the Allais Paradox?" *Journal of Economic Theory*, 46(1), 1988, 145-153.

Sandorf, Erlend D., Danny Campbell, and Nick Hanley, "Disentangling The Influence of Knowledge on Attribute Non-attendance," *Journal of Choice Modelling*, 24, 2017, 36-50.

Sarias, Mauricio, "Individual-specific Posterior Distributions from Mixed Logit Models: Properties, Limitations and Diagnostic Checks," *Journal of Choice Modelling*, 36, September 2020, 100224.

Schwarz, Gideon, "Estimating the Dimension of a Model," *The Annals of Statistics*, 6(2), March 1978, 461-464.

Shanon, Claude E., "A Mathematical Theory of Communication," *The Bell System Technical Journal*, 27, July 1948, 379-423.

Swait, Joffre and Wiktor Adamowicz, "Choice Environment, Market Complexity, and Consumer Behavior: A Theoretical and Empirical Approach for Incorporating Decision Complexity into Models of Consumer Choice," *Organizational Behavior and Human Decision Processes*, 86(2), 2001, 141–167.

Train, Kenneth, *Discrete Choice Models with Simulation* (Cambridge, UK: Cambridge University Press, Second Edition, 2009).

Von Gaudecker, Hans-Martin, Arthur van Soest, and Erik Wengström, "Heterogeneity in Risky Choice Behavior in a Broad Population," *The American Economic Review*, 101(2), 2011, 664-694.

Wilcox, Nathaniel T., "Lottery Choice: Incentives, Complexity and Decision Time," *The Economic Journal*, 103(421), 1993, 1397-1417.

Wilcox, Nathaniel T., "Theories of Learning in Games and Heterogeneity Bias," *Econometrica*, 74(5), 2006, 1271-1292.

Wilcox, Nathaniel T., "'Stochastically More Risk Averse:' A Contextual Theory of Stochastic Discrete Choice under Risk," *Journal of Econometrics*, 162(1), 2011, 89-104.

Wilcox, Nathaniel T., "Error and Generalization in Discrete Choice Under Risk," *ESI Working Paper 15-11*, 2015.

Wooldridge, Jeffrey, *Econometric Analysis of Cross Section and Panel Data* (Cambridge, USA: MIT Press, Second Edition, 2010).

Yaari, Menahem E., "The Dual Theory of Choice under Risk," *Econometrica*, 55(1), 1987, 95-115.

Zhou, Xiao-Hua, Sujuan Gao, and Siu L. Hui, "Methods for Comparing the Means of Two Independent Log-Normal Samples," *Biometrics*, 1997, 53(3), 1129-1135.

Chapter 2

Adriani, Fabrizio and Silvia Sonderegger, "Optimal Similarity Judgments in Intertemporal Choice (and Beyond)," *Journal of Economic Theory*, 190, 2020, 105097.

Ainslie, George, "Specious Reward: a Behavioral Theory of Impulsiveness and Impulse Control," *Psychological Bulletin*, 92(4), 1975, 463–496.

Andersen, Steffen, James C. Cox, Glenn W. Harrison, Morten I. Lau, E. Elisabet Rutström, and Vjollca Sadiraj, "Asset Integration and Attitudes toward Risk: Theory and Evidence," *The Review of Economics and Statistics*, 100(5), 2018, 816–830.

Andersen, Steffen, Glenn W. Harrison, Morten I. Lau, and E. Elisabet Rutström, "Eliciting Risk and Time Preferences," *Econometrica*, 76(3), 2008, 583-618.

Andersen, Steffen, Glenn W. Harrison, Morten I. Lau, and E. Elisabet Rutström, "Discounting Behavior and the Magnitude Effect: Evidence from a Field Experiment in Denmark," *Economica*, 80(320), 2013, 670-697.

Andersen, Steffen, Glenn W. Harrison, Morten I. Lau, and E. Elisabet Rutström, "Discounting Behavior: A Reconsideration," *European Economic Review*, 71(1), 2014a, 15-33.

Andersen, Steffen, Glenn W. Harrison, Morten I. Lau, and E. Elisabet Rutström, "Dual criteria decisions," *Journal of Economic Psychology*, 41, 2014b, 101-113.

Andersen, Steffen, Glenn W. Harrison, Morten I. Lau, and E. Elisabet Rutström, "Multiattribute Utility Theory, Intertemporal Utility, and Correlation Aversion," *International Economic Review*, 59(2), 2018, 537-555.

Andreoni, James, and Charles Sprenger, "Estimating Time Preferences from Convex Budgets," *American Economic Review*, 102(7), 2012, 3357–3376.

Apesteguia, José, and Miguel A. Ballester, "Monotone Stochastic Choice Models: The Case of Risk and Time Preferences," *Journal of Political Economy*, 126(1), 2018, 74-106.

Benzion, Uri, Amnon Rapoport, and Joseph Yagil, "Discount Rates Inferred from Decisions: An Experimental Study," *Management Science*, 35(3), 1989, 270-284.

Blavatskyy, Pavlo R., and Ganna Pogrebna, "Models of Stochastic Choice and Decision Theories: Why Both are Important for Analyzing Decisions," *Journal of Applied Econometrics*, 25(6), 2010, 963-986.

Brandstätter, Eduard, Gerd Gigerenzer, and Ralph Hertwig, "The Priority Heuristic: Making Choices Without Trade-Offs," *Psychological Review*, 113(2), 2006, 409-432.

Cameron, Adrian C., and Pravin K. Trivedi, *Microeconometrics: Methods and Applications* (Cambridge, UK: Cambridge University Press, 2009).

Cheung, Stephen, "Risk Preferences are Not Time Preferences: On the Elicitation of Time Preference under Conditions of Risk: Comment." *American Economic Review*, 105(7), 2015, 2242-2260.

Clarke, Kevin A., "Nonparametric Model Discrimination in International Relations," *The Journal of Conflict Resolution*, 47(1), 2007, 72-93.

Clarke, Kevin A., "A Simple Distribution-Free Test for Nonnested Model Selection," *Political Analysis*, 15(3), 2007, 347-363.

Coller, Maribeth, and Melonie B Williams, "Eliciting Individual Discount Rates," *Experimental Economics*, 2, 1999, 107-127.

Desmarais, Bruce A., and Jeffrey J. Harden, "Testing for Zero Inflation in Count Models: Bias Correction for the Vuong test," *The Stata Journal*, 13(4), 2013, 810-835.

Dixit, Vinayak V., Rami C. Harb, Jimmy Martínez-Correa, and Elisabet E. Rutström, "Measuring Risk Aversion to Guide Transportation Policy: Contexts, Incentives, and Respondents," *Transportation Research Part A: Policy and Practice*, 80, 2015, 15-34.

Dohmen, Thomas, Armin Falk, David Huffman, and Uwe Sunde, "Are Risk Aversion and Impatience Related to Cognitive Ability?," *American Economic Review*, 100(3), 2010, 1238-12

Eckel, Catherine, Cathleen Johnson, and Claude Montmarquette, "Savings Decisions of the Working Poor: Short- and Long-term Horizons," in J. Carpenter, G. W. Harrison and J. A. List (eds.), Field Experiments in Economics (Greenwich, CT: JAI Press, *Research in Experimental Economics*, Volume 10, 2005).

Ekeland, Ivar, and Traian A. Pirvu, "Investment and Consumption without Commitment," *Mathematics and Financial Economics*, 2, 2008, 57-86.

Ericson, Keith M. M., John M. White, David Laibson, and Jonathan D. Cohen, "Money Earlier or Later? Simple Heuristics Explain Intertemporal Choices Better Than Delay Discounting Does," *Psychological Science*, 26(6), 2015, 826-833.

González-Vallejo, Claudia, "Making Trade-Offs: A Probabilistic and Context-Sensitive Model of Choice Behavior," *Psychological Review*, 109(1), 2002, 137-155.

Harden, Jeffrey J., and Bruce A. Desmarais, "Linear Models with Outliers: Choosing between Conditional-Mean and Conditional-Median Methods," *State Politics & Policy Quarterly*, 11(4), 2011, 371-389.

Harrison, Glenn W., Morten I. Lau, and Melonie B. Williams, "Estimating Individual Discount Rates in Denmark: A Field Experiment," *American Economic Review*, 92(5), 2002, 1606-1617.

Harrison, Glenn W., Morten I. Lau, and E. Elisabet Rutström, "Estimating Risk Attitudes in Denmark: A Field Experiment," *Scandinavian Journal of Economics*, 109(2), 2007, 341-368.

Harrison, Glenn W., Morten I. Lau, and Hong Il Yoo, "Constant Discounting, Temporal Instability and Dynamic Inconsistency in Denmark: A Longitudinal Field Experiment," Center for Economic Analysis of Risk (CEAR) Working paper No. 2020-09, 2020.

Harrison, Glenn W., and E. Elisabet Rutström, "Expected Utility And Prospect Theory: One Wedding and A Decent Funeral," *Experimental Economics*, 12(2), 2009, 133-158.

Harrison, Glenn W., and Todd Swarthout, "Cumulative Prospect Theory in the Laboratory: A Reconsideration," Forthcoming in G.W. Harrison and D. Ross (eds.), *Models of Risk Preferences: Descriptive and Normative Challenges* (Bingley, UK: Emerald, Research in Experimental Economics, 2021).

Holt, Charles A., and Susan K. Laury, "Risk Aversion and Incentive Effects," *American Economic Review*, 92(5), 2002, 1644-1655.

Karp, Larry, "Non-constant Discounting in Continuous Time," *Journal of Economic Theory*, 132, 2007, 557 – 568.

Laibson, David, "Golden Eggs and Hyperbolic Discounting," *The Quarterly Journal of Economics*, 112(2), 1997, 443–477.

Leland, Jonathan W., "Similarity Judgments and Anomalies in Intertemporal Choice," *Economic Inquiry*, 40(4), February 2002, 574-581

Luce, Robert D., *Individual Choice Behavior: A Theoretical Analysis* (New York: Wiley, 1959).

Mazur, James E., "Tests of an Equivalence Rule for Fixed and Variable Reinforcer Delays," *Journal of Experimental Psychology: Animal Behavior Processes*, 10(4), 1984, 426–437.

Nadeau, Claude, and Yoshua Bengio, "Inference for the Generalization Error," *Machine Learning*, 52, 2003, 239–281.

Phelps, Edmund S., and Robert A. Pollak, "On Second-Best National Saving and Game-Equilibrium Growth," *The Review of Economic Studies* , 35(2), 1968, 185-199.

Prelec, Drazen, "The Probability Weighting Function," *Econometrica*, 66, 1998, 497-527.

Prelec, Drazen, "Decreasing Impatience: A Criterion for Non-stationary Time Preference and "Hyperbolic" Discounting," *Scandinavian Journal of Economics*, 106(3), 2004, 511-532.

Quiggin, John, "A Theory of Anticipated Utility," *Journal of Economic Behavior & Organization*, 3(4), 1982, 323-343.

Raftery, Adrian E., "Bayesian Model Selection in Social Research," *Sociological Methodology*, 25, 1995, 111-163 .

Read, Daniel, "Is Time-Discounting Hyperbolic or Subadditive?" *Journal of Risk and Uncertainty*, 23(1), 2011 5–32.

Read, Daniel, and Peter H.M.P Roelofsma, "Subadditive Versus Hyperbolic Discounting: A Comparison of Choice And Matching," *Organizational Behavior and Human Decision Processes*, 91(2), 2003, 140-153.

Read, Daniel, Frederick Shane, and Marc Scholten, "DRIFT: An Analysis of Outcome Framing in Intertemporal Choice," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39(2), 2012, 573-588.

Rieger, Mark O., and Mei Wang, "What is Behind the Priority Heuristic? A Mathematical Analysis and Comment on Brandstätter, Gigerenzer, and Hertwig (2006)," *Psychological Review*, 115(1), 2008, 274–280.

Rubinstein, Ariel, "Similarity and Decision-making under Risk (Is There a Utility Theory Resolution to the Allais Paradox?)," *Journal of Economic Theory*, 46(1), October 1988, 145-153.

Salmon, Timothy C., "An Evaluation of Econometric Models of Adaptive Learning," *Econometrica*, 69(6), 2001, 1597-1628.

Samuelson, Paul A., "A Note on Measurement of Utility," *The Review of Economic Studies*, 4(2), 1937, 155-161.

Shapiro, S. S., and M. B. Wilk, "An Analysis of Variance Test for Normality (Complete Samples)," *Biometrika*, 52(3/4), 1965, 591-611.

Scholten, Marc, and Daniel Read, "Discounting by Intervals: A Generalized Model of Intertemporal Choice," *Management Science*, 52(9), 2006, 1424–1436.

Scholten, Marc, and Daniel Read, "The Psychology of Intertemporal Tradeoffs," *Psychological Review*, 117(3), 2010, 925–944.

Schwarz, Gideon, "Estimating the Dimension of a Model," *The Annals of Statistics*, 6(2), 1978, 461-464.

Stahl, Dale 0., "An Empirical Evaluation of the Toolbox Model of Lottery Choices," *The Review of Economics and Statistics*, 100(3), 2018, 528–534.

Thaler, Richard, "Some Empirical Evidence on Dynamic Inconsistency," *Economics Letters*, 8(3), 1981, 201-207.

Tornqvist, Leo, Pentti Vartia, and Yrjo O. Vartia, "How Should Relative Changes Be Measured?" *The American Statistician*, 1985, 39(1), 43-46.

Tversky, Amos, "Intransitivity of preferences," *Psychological Review*, 76(1), 1969, 31-48.

Wilcox, Nathaniel T., "Stochastic Models for Binary Discrete Choice under Risk: A Critical Stochastic Modeling Primer and Econometric Comparison," In J. C. Cox and G. W. Harrison (eds.), *Risk Aversion in Experiments*. (Bingley, UK: Emerald, Research in Experimental Economics, 2008).

Wilcox, Nathaniel T., "'Stochastically More Risk Averse:' A Contextual Theory of Stochastic Discrete Choice under Risk," *Journal of Econometrics*, 162(1), 2011, 89-104.

Wilcoxon, Frank, "Individual Comparisons by Ranking Methods," *Biometrics Bulletin*, 1(6), 1945, 80-83.

Wulff, Dirk U., and Wouter van den Bos, "Modeling Choices in Delay Discounting," *Psychological Science*, 29(11), 2018, 1890-1894.