ISSN (Print) 0473-453X
ISSN (Online) 2435-0982

Discussion Paper No. 1185

INVESTIGATION OF THE CONVEX TIME BUDGET EXPERIMENT BY PARAMETER RECOVERY SIMULATION

Keigo Inukai
Yuta Shimodaira
Kohei Shiozawa

Revised March 2023
August 2022

The Institute of Social and Economic Research
Osaka University
6-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan

Investigation of the Convex Time Budget Experiment by Parameter Recovery Simulation*

Keigo Inukai† Yuta Shimodaira‡ Kohei Shiozawa§

March 19, 2023

Abstract

The convex time budget (CTB) method is a widely used experimental technique for eliciting an individual's time preference in intertemporal choice problems. This paper investigates the accuracy of the estimation of the discount factor parameter and the present bias parameter in the quasi-hyperbolic discounting utility function for the CTB experiment. In this paper, we use a simulation technique called "parameter recovery." We found that the precision of present bias parameter estimation is poor within the scope of previously reported parameter estimates, making it difficult to detect the effect of present bias. Our results recommend against using a combination of the CTB experimental task and the quasi-hyperbolic discounting utility model to explore the effect of present bias. This paper contributes to addressing the replicability issue in experimental economics and highlights the importance of auditing the accuracy of parameter estimates before conducting an experiment.

Keywords: Discounting, Convex time budget, Quasi-hyperbolic discounting, Present bias, Parameter recovery

JEL codes: C91, D90

*The authors are grateful to Taisuke Imai, Nobuyuki Hanaki, Takeshi Murooka, Masaru Sasaki, and participants of the Workshop on Microeconomic Analysis of Social Systems and Institutions (at Kansai University) for helpful discussions. Y.S. acknowledges financial support provided through a Grant-in-Aid for JSPS Fellows (19J12097) from the Japan Society for the Promotion of Science. K.I. acknowledges financial support provided through a Grant-in-Aid for Young Scientists (17H04780) and a Grant-in-Aid for Transformative Research Areas (21H05070) from the Japan Society for the Promotion of Science. OnLine English (www.oleng.com.au) edited the manuscript to improve its readability.

†Department of Economics, Meiji Gakuin University. E-mail: inukai@eco.meijigakuin.ac.jp

‡Institute of Social and Economic Research, Osaka University. E-mail: shimodaira@iser.osaka-u.ac.jp

§Department of Economics, Takasaki City University of Economics. E-mail: shiozawa@tcue.ac.jp

1 Introduction

All forms of life face the trade-off between smaller, immediate rewards and larger, delayed rewards. However, most organisms, including humans, struggle to delay rewards and tend to give priority to immediate gains rather than large future rewards. The discount rate is a key factor in determining the degree to which future profits are discounted over time in intertemporal choice problems. Discount rates can be measured by various means, and interesting findings have arisen from such assessments.

It is important to note that discount rates can change over time. To illustrate this, suppose that we face the following choice: consume one chocolate now or delay gratification for a week and receive two chocolates. Many individuals would likely succumb to temptation and choose to consume one chocolate immediately rather than wait for the larger reward. However, if the choice becomes whether to consume one chocolate in a week or two chocolates in two weeks, people are more prone to wait the full two weeks. This tendency is known as *present bias*, a time inconsistency of choice associated with the choice problem between different points in time, as discussed in O'Donoghue and Rabin (2015). The existence of present bias suggests that our willpower may be weaker than we imagine. The phenomenon of procrastination regarding unpleasant tasks is among the serious issues engendered by such anomalies.

Moreover, it is imperative to recognize that intertemporal choice problems entail an element of risk. Individuals may perceive immediate gains as certain, while future gains are perceived as uncertain. The precise influence of risk on the choice problem between different points in time remains a matter of debate. Accordingly, experimental economists have proposed various methodologies to isolate choice tasks and risk preferences across various time frames.

One method is called the convex time budget (CTB), which was developed by Andreoni and Sprenger (2012a, henceforth AS). The CTB method attempts to elicit simultaneously the effects of time discounting and risk attitude by directly estimating the curvature of the utility function using one single instrument.[1] When analyzing

[1]Although studies using the CTB method are still being conducted, the interpretation of the interplay between time preferences and risk preferences is still a matter of debate (Andreoni and Sprenger, 2012b; Harrison et al., 2013; Andersen et al., 2014; Cheung, 2015; Epper and Helga, 2015; Miao and Zhong, 2015; Andreoni et al., 2015; Andreoni and Sprenger, 2015).

preferences from behavioral data collected by the CTB method, researchers do not merely compare the intertemporal allocations across conditions, but also estimate the parameters of the quasi-hyperbolic discounting utility function (Laibson, 1997; O'Donoghue and Rabin, 1999). In experimental economics, many experiments on intertemporal choice problems now adopt the CTB method in both laboratory and field settings (e.g., Augenblick et al., 2015; Carvalho et al., 2016; Blumenstock et al., 2018; Cheung et al., 2022; Dantas et al., 2022). Imai et al. (2020) have identified 67 articles that employed the CTB method and presented a meta-analysis on these studies. This meta-analysis showed that, on average, the experiment participants discounted the future payoff by 0.95–0.97 over the payoff available now.[2]

While researchers typically assess the reliability of estimates post hoc based on the magnitude of standard errors associated with the estimates, it is uncommon to examine the trueness and precision of the estimates prior to conducting an experiment. The degree to which we can accurately estimate an individual's utility function using a CTB experiment remains unclear. For instance, if an individual's present bias parameter estimate is 0.97, can we truly claim that this individual's behavior is biased? To address this issue, we use a simulation technique called "parameter recovery" (Wilson and Collins, 2019) to examine the accuracy of parameter estimates. The process of parameter recovery simulation involves three steps: first, generating artificial decision data using assumed parameter values (referred to as "ground-truth values"); second, estimating the parameters from the artificial data using the software intended for the real data; and finally, comparing the estimated parameters to the true values to assess the level of precision in their recovery. This paper audits the accuracy of parameter estimation in a conventional CTB experimental design. Our findings demonstrate that the performance of present bias parameter estimation is suboptimal within the scope of previously reported parameter estimates.

According to Imai et al.'s (2020) meta-analysis, there may be a tendency for selective reporting of present bias parameter estimates of less than one, particularly in studies using real effort tasks. Additionally, our investigation has revealed imprecision in estimating present bias parameters, which can exacerbate the problem of selective reporting of the parameter estimate by reducing the power of a statistical test based

[2]Cheung et al. (2021), while performing a meta-analysis on present bias parameter estimates that were not limited to CTB method papers, observed that estimates derived from data collected through the CTB experiment tended to be closer to 1 compared with those obtained from other methods, including the double multiple price list method proposed by Andersen et al. (2008).

on it, regardless of its true value (van Zwet and Cator, 2021). Consequently, our results imply serious caution against the use of the CTB method at least in its conventional form—more precisely, a combination of the CTB experimental task and the quasi-hyperbolic discounting utility model—for exploring the effect of present bias that is uncertain and possibly insignificant.

In psychology, the replicability of experimental findings can often be problematic, and in experimental economics, it is a crucial issue to that should also be considered. While it has been recognized that the replication rate of experimental studies in economics is somewhat superior to that in psychology (Camerer et al., 2016), there is still heterogeneity in outcomes across experiments. This variability in experimental outcomes may be attributed to participants' demographic and cultural backgrounds, but it could also be contingent on the measurement technique and parameter estimation method used. To ensure replicability of experimental results, it is imperative that we audit our experimental methods by carrying out simulations at the experimental design phase.

The remainder of this paper is organized as follows. Section 2 describes the virtual design of a CTB experiment and a behavioral model for the CTB experiment, as well as the parameter recovery simulation procedures. Section 3 contains the results of the parameter recovery simulation. In this paper, we perform simulations to 1) analyze whether discounting behaviors can be detected based on the standard errors associated with the estimates and 2) evaluate the resolution of the parameter estimates from the distribution of the estimates. Then, we show that the combination of the CTB method and the quasi-hyperbolic discounting model cannot obtain estimates of the present bias parameter with small errors or correctly detect the bias if the actual effect size is small. In the last part of Section 3, we discuss the reasons for the low resolution of the present bias parameter estimation. Section 4 concludes.

2 Methods

To conduct a parameter recovery simulation, we will clarify how to generate synthetic decision data in a CTB experiment—the definition of the demand function, the specification of the experimental task, and the selection of the ground-truth values of the parameters—and how to estimate parameters.

2.1 Behavioral Model

We now consider the decision-making problems associated with allocating the initial endowment m between the sooner and later periods. Let (c_t, c_{t+k}) denote an allocation bundle where c_t is the payoff for the sooner period t and c_{t+k} is for the k days later period. It only matters whether the sooner period t is 0 (i.e., present) or not; and for $t > 0$, the value of t does not matter, at least in the model we use. The exchange rate from tokens to material payoffs varies between the sooner and later periods, and we normalize the rate for the later period to be 1. We denote the exchange rate for the sooner payoff as $1 + r$, where r is interpreted as an interest rate. We assume that income is exhausted, or that the budget constraint binds the allocation bundle. Here, we can obtain the budget constraint for the decision problem as follows:

$$(1 + r)c_t + c_{t+k} = m. \tag{1}$$

To measure an individual's time preference, the experimenter asks the participants for their allocation $(c_t,\ c_{t+k})$ by changing t, k, $1 + r$, and m.

Here, we discuss a theoretical model of participants' behavior (c_t, c_{t+k}) for a given CTB experiment task $(t, k, 1 + r, m)$. For the intertemporal decision-making task described above, we suppose that each individual's time preference is represented by the following constant intertemporal elasticity of substitution and quasi-hyperbolic discounting (CES-QHD) utility function (Laibson, 1997; O'Donoghue and Rabin, 1999):

$$U(c_t, c_{t+k}) = \frac{1}{\rho}\, c_t^{\rho} + \beta^{\mathbf{1}_{t=0}} \delta^k \frac{1}{\rho}\, c_{t+k}^{\rho}. \tag{2}$$

The variable $\mathbf{1}_{t=0}$ is an indicator of whether the earlier period is the present period. The parameter δ (> 0) is the one-day discount factor, and the parameter β (> 0) represents the present/future bias. The parameter ρ controls the curvature of the utility function and characterizes the intertemporal elasticity of substitution $\sigma = (1 - \rho)^{-1}$.[3]

[3]Laibson (1997) specified that an individual's utility function is a function of the summation of instantaneous utility characterized by constant relative risk aversion. Following Laibson (1997), Andreoni and Sprenger (2012a) interpreted the parameter ρ as a risk attitude measure. They compared the parameter ρ to the within-subject Holt and Laury's (2002) risk preference measure elicited by the multiple price list tasks—the components of the double multiple price list task developed by Andersen et al. (2008)—and found that the two measures are virtually uncorrelated. The relationship between the curvature of utility under risk and utility over time is highly controversial

We assume that an individual whose preferences are represented by the CES-QHD utility function (2) faces the utility maximization problem subject to the budget constraint (1). By solving this utility maximization problem, we obtain the following demand function:

$$g(t,k,1+r,m \mid \delta,\beta,\sigma) = \begin{cases} \dfrac{1}{1+\left(\beta\delta^k\right)^{\sigma}\left(1+r\right)^{\sigma-1}} & \text{for} \quad t=0, \\ \dfrac{1}{1+\left(\delta^k\right)^{\sigma}\left(1+r\right)^{\sigma-1}} & \text{for} \quad t>0. \end{cases} \tag{3}$$

Note that the value of the demand function g corresponds to the sooner allocation c_t divided by its upper limit $m/(1+r)$, and therefore the function g maps onto the interval $[0,1]$. For mathematical tractability, the elasticity of substitution, σ, is used instead of the parameter ρ (details are provided in Section 2.3).

We perturbed the generated normalized sooner allocation $g(\bullet)$ by adding a random number ϵ, which follows a normal distribution with mean 0 and standard deviation $s \in \{0.01,\ 0.05,\ 0.10,\ 0.15,\ 0.20\}$. As the ratio of mean absolute deviation to standard deviation is $\sqrt{2/\pi} \approx 0.8$, the generated data have, on average, a 0.8% error for the interval length allowed as a decision c_t for $s = 0.01$. In the original experiment by Andreoni and Sprenger (2012a, henceforth AS), participants were asked to select an integer in the interval from 0 to 100 as a normalized allocation, which corresponds to the value of g multiplied by 100. Given that forcing discrete choice causes rounding errors in decision-making, an error size of $s = 0.01$ is inevitable. We obtained the root mean squared error (RMSE) for the parameter estimation of AS's experimental dataset: the first quartile is 0.019, the median is 0.14, and the third quartile is 0.22. Given the RMSE distribution, we believe that $s = 0.20$ is not necessarily too large.

We truncated the noise-added value to the interval $[0,1]$: we draw a random number from the distribution $\mathcal{N}\big(g(\bullet), s\big)$ and accept it as a synthesized decision if it is in $[0,1]$; otherwise, we repeatedly draw a random number again. This is because the decision task that we are considering here involves the allocation of endowment between two periods. In this scenario, decision-makers are not allowed to borrow money to consume more than their endowment in the sooner period and to repay

(Abdellaoui et al., 2013; Andersen et al., 2014; Cheung, 2015; Harrison et al., 2013; Takeuchi, 2012). We then refrain from interpreting the parameter ρ as a risk measure and instead refer to it as the mathematically straightforward interpretation; namely, the elasticity of substitution between two periods.

it in the later period. When the actual decision is at the endpoint of the budget constraint line, noise can cause the decision to move toward the inside but not toward the outside.[4]

2.2 Experimental Tasks

We have two experimental situations (defined as a combination of an early period date t and a delay length k) for the experimental task: $t = 0$ (i.e., present) and $k = 70$ (days), and $t = 1$ (i.e., not present) and $k = 70$ (days). The delay length k is typically on the scale of weeks to months, and is rarely shorter than one week (Imai et al., 2020). In each situation, we set 21 uniformly spaced prices chosen from $0.6 \leq 1 + r \leq 2$. We fixed income m at 20 for simplicity, because it does not affect behavior in the model. The number of tasks, i.e., the number of data points for each individual, is 42.

There are three critical differences between our problem set and AS's problem set. The first difference is that we chose the price $1 + r$ from the range where the interest rate r is not only positive but also negative. Few studies using CTB experiments, including that by AS, ask participants about negative interest rates. However, without asking about negative interest rates, it is impossible to estimate the discount factor for an individual who does not discount the future payoffs, but who does place a premium on them (for such an individual, the discount factor δ will be greater than 1). We also conducted simulations using the AS's original problem set and summarized the results in Appendix F.

Second, the delay k was set equal to 70 in this paper to simplify the discussion. Note that for AS's problem set, there were three conditions of k: 35, 70, and 98. We also investigated the effect of the number of conditions on k in Appendix F. Increasing or decreasing the variation in k may affect the parameter estimation error.

Third, we reduced the number of early period dates t to two, i.e., present or not present. Note that in AS's experiment, participants made decisions about the allocation between the future and a later future for $t = 7$ and $t = 35$ separately.

[4]The truncated-noised data are always the interior points of the budget constraint line, and no corners are chosen. As Harrison et al. (2013) pointed out, it is known that corners are easily chosen in CTB experiments. Therefore, it could be a more realistic assumption that the noise is censored at the corners—a noise-added value is shifted to 0 or 1 if a random number drawn from $\mathcal{N}(g(\bullet), s)$ is outside of $[0, 1]$. We also conducted parameter estimation using data with censored noise (see Appendix H).

Regarding the CES-QHD utility function model, there is no difference in decisions between $t = 7$ and $t = 35$, but it can affect real behavior. For a real experimental design, it may be helpful to designate the variation in t to treat bias in the parameter estimates, but we discarded that option.

2.3 Ground-truth Values

For the ground-truth values, we used 10 equally spaced values for δ and β from the range $0.9912 \leq \delta \leq 1.0025$ and $0.85 \leq \beta \leq 1.12$, respectively. For the curvature parameter, we use $\ln \sigma = \ln(1/1 - \rho)$ instead of the commonly used notation ρ for mathematical clarity. For the ground-truth curvature $\ln \sigma$, we used seven equally spaced values from the range $0.33 \leq \ln \sigma \leq 5.00$. Table 1 shows the ground-truth values that generate the decision data. We chose values for δ, β, and $\ln \sigma$; these values are evenly spaced as if from a uniform distribution. By combining the ground-truth values of the three parameters listed in Table 1, there are 700 synthetic individuals. As mentioned above, there are five levels of noise, s, and we generate 10 sets of data for each s, resulting in decision data for 35,000 agents.

For the discounting parameters δ and β, we selected a range that covers the distribution of the estimates reported in AS's paper. Usually, the discount factor δ and the present bias β are assumed to be less than 1. However, because some studies report individuals with estimates greater than 1,[5] we also included these values in our set of possible ground-truth values.

We next describe in detail how we selected the range of the curvature parameter $\ln \sigma$: from 0.33 ($\rho = 0.283$; nearly the Cobb–Douglas utility curvature) to 5 ($\rho = 0.993$; nearly linear curvature). Recall that the domain of $\ln \sigma$ is all real numbers. Let us assume that $\ln \sigma = 0$ (or $\rho = 0$), which corresponds to the Cobb–Douglas utility function, is the center of the curvature parameter space. For $\ln \sigma > 0$, the intertemporal allocations become substitutive and complementary otherwise. As $\ln \sigma \longrightarrow -\infty$ (or $\rho \longrightarrow -\infty$), the utility function goes to a Leontief function: $U = \min\{c_t, c_{t+k}\}$, whose indifference curve is L-shaped and is known as the perfect complement utility function. As $\ln \sigma \longrightarrow +\infty$ (or $\rho \longrightarrow 1$), the utility function goes

[5]We obtained estimates from AS's experimental dataset. For the distribution of the δ estimates, the 5th percentile is 0.9917, the median is 0.9989, and the 95th percentile is 1.0018. For the distribution of the β estimates, the 5th percentile is 0.89, the median is 1.01, and the 95th percentile is 1.15.

Table 1: Ground-truth values

δ	0.9912	0.9925	0.9937	0.9950	0.9962	0.9975	0.9987	1.0000	1.0012	1.0025
β	0.85	0.88	0.91	0.94	0.97	1.00	1.03	1.06	1.09	1.12
$\ln \sigma$	0.33	1.11	1.89	2.67	3.44	4.22	5.00			
(ρ)	(0.283)	(0.671)	(0.849)	(0.931)	(0.968)	(0.985)	(0.993)			
s	0.01	0.05	0.10	0.15	0.20					

Note: For the curvature parameter $\ln \sigma$, the corresponding ρ are listed.

to a linear function: $U = c_t + \beta^{\mathbf{1}_{t=0}}\delta^k c_{t+k}$, which is the perfect substitution utility function.

AS have reported that the curvature of participants' preferences in CTB experiments is generally (but not completely) linear. Regardless of the distribution of the actual parameter values, we should also check the estimation errors for individuals who behave relatively complementarily, because the utility function model does not explicitly exclude such individuals. However, it is known that for the standard CES utility function $U(x, y) = (x^\rho + \phi y^\rho)^{1/\rho}$, when the curvature ρ is negative, the share parameter ϕ—which corresponds to the discounting part $\beta^{\mathbf{1}_{t=0}}\delta^k$ in the CES-QHD utility—cannot be accurately estimated for mathematical reasons (Inukai et al., 2022; Thöni, 2015). As the estimation errors of δ and β are inevitably large for $\ln\sigma < 0$, we excluded them from our analysis. Consequently, we chose ground-truth values for $\ln\sigma$ from 0.33 to 5. We also investigated the estimation errors in the same way for $\ln\sigma < 0$ and reported the results in Appendix G. Note that previous studies on the curvature of time preferences report that it is rare to observe individuals for whom $\ln\sigma$ is negative, regardless of whether or not the CTB method is used (Andersen et al., 2008; Andreoni and Sprenger, 2012a; Andreoni et al., 2015; Cheung, 2020).

2.4 Estimation Methods

As we described above, for all individuals i characterized by $(\delta_i, \beta_i, \ln\sigma_i)$, and for all budget constraint lines $j \in \{1, \ldots, 42\}$, we obtain the decision data $\widetilde{c}_i^j = g(t_j, k_j, 1 + r_j, m_j \mid \delta_i, \beta_i, \ln\sigma_i) + \epsilon$. Given the generated data, we estimate the three parameters using a nonlinear least squares method.[6] Following AS, we used the "`nl`" command in Stata. Mathematically, the values of $\widehat{\delta}$, $\widehat{\beta}$, and $\widehat{\ln\sigma}$ minimize the sum of squared residuals:

$$\sum_{j=1}^{42}\Big[\widetilde{c}_i^j - g\big(t_j, k_j, 1 + r_j, m_j \mid \delta_i, \beta_i, \ln\sigma_i\big)\Big]^2. \tag{4}$$

To prevent estimation failures because of nonconvergence of the calculations, we

[6] In AS, the error term was assumed to follow a censored normal distribution, and a two-limit Tobit model was used for estimation. However, the two-limit Tobit model may result in unexpected interpretations when the error scale s is moderately large. For example, when $g(\bullet) = 0.8$ and $s = 0.1$, the decision is more likely to be in the corner ($\widetilde{c} = 1$) rather than in a position closer to the theoretical decision. For this reason, we specify an error term with a truncated normal distribution rather than a censored distribution.

transformed $\ln \sigma$ using a sigmoid function f as $\ln \sigma = f(\theta) = 4 \tanh(\theta) + 1.5$ and estimated the latent variable θ.[7]

For the parameter estimation, we set the convergence criterion as 10^{-5} and the maximum number of iterations as 200. By combining all parameters $(\delta_i, \beta_i, \ln \sigma_i)$ and s, there are 3,500 synthetic individuals, and we regenerated the decision data 10 times for each synthetic individual. Of the 3,500 synthetic individuals, 3,483 converged all 10 times, and the remaining 17 individuals had only one failure to converge.

3 Results

3.1 Detectability of Time Discounting

As a first measure to discuss the estimation error, we examine whether the estimated discount factor $\widehat{\delta}$ and the present/future bias parameter $\widehat{\beta}$ are distinguishable from 1, indicating that the individual does not discount (or place a premium on) future payoffs. In previous studies, most attention has been paid to whether present-biased behavior exists. We examine here how far the true β is away from 1 to determine whether it can be distinguished from 1.

Figure 1 shows the percentage of successfully rejected null hypotheses such that $\widehat{\delta} = 1$ and $\widehat{\beta} = 1$. We conducted two-tailed Student's t-tests at the 5% significance level to examine whether the null hypothesis could be rejected for each simulation agent.[8] Each point summarizes 10 replications of all combinations of ground-truth

[7]For $\ln \sigma > 5.5$, we cannot observe differences in the decision data with a practical significant figure for an additional decrease in $\ln \sigma$; in other words, we cannot observe an increase in substitutability as a behavior. For $\ln \sigma < -2.5$, we also cannot observe an increase in complementarity for an additional increase in $\ln \sigma$. Then, we assume that $\ln \sigma$ greater than 5.5 means perfect substitutes and $\ln \sigma$, less than -2.5 means perfect complements because the effect of $\ln \sigma$ variation on behavior $g(\bullet)$ saturates (see the demand curves in Appendix J). In the saturating range, the parameter estimations sometimes do not converge. We can prevent calculation failures using the S-shaped function $f(\theta)$. If a ground-truth $\ln \sigma$ is positive, as discussed in the main analysis, this transformation had little effect: out of 35,000 agents, it failed 15 (0.004%) without and 17 (0.005%) with the transformation. However, when ground-truth $\ln \sigma$ is negative (-2.00, -1.22, and -0.44), the calculations failed 367 (2%) out of 15,000 agents without, but never with the transformation. See Appendix K for a comparison of estimates with and without the transformation.

[8]The test statistics are computed using the standard error of the estimate, which is estimated by the jackknife method. We found that estimation using the bootstrap method overestimates the standard error of the estimate (see Appendix I). Therefore, we chose the jackknife method to avoid undervaluing the precision, i.e., to be conservative about what we are trying to conclude.

values of β (δ) and $\ln \sigma$, i.e., 700 simulation agents.

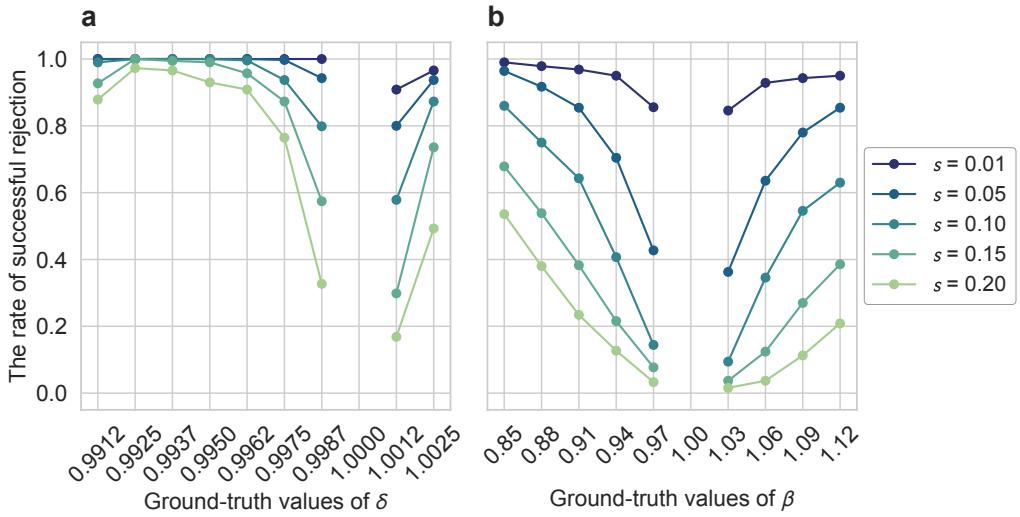


Figure 1: Rate of successful rejection of the null hypothesis

Notes: Tests on **a**) $\widehat{\delta} = 1$ and **b**) $\widehat{\beta} = 1$. Each point summarizes 10 replications of all combinations of ground-truth values of β (δ) and $\ln \sigma$, i.e., 700 simulation agents.

An inspection of Figure 1 reveals that, for the discount factor parameter δ, when the ground-truth value is less than 0.9962, we can reject $\widehat{\delta} = 1$ in over 90% of cases regardless of the amount of added noise. For the case of $\delta > 1$, it may be more challenging to reject null hypotheses compared to the case of $\delta < 1$. To estimate δ accurately to place a premium on future payoffs, it is necessary to collect decision data at negative interest rates. However, in our simulation, we included fewer questions with negative interest rates and therefore, the estimation precision was worse than that of $\delta < 1$.[9]

In contrast, for the present/future bias parameter β, we had more difficulty concluding that the estimates are not equal to 1 compared with the case of the discount factor parameter δ in general. Even when the true β is as small as 0.85, the success rate is below 90% for $s > 0.05$.

3.2 Error Size

To examine the errors of the parameter estimates further, instead of focusing on the estimated uncertainty of the parameter estimates for each individual, we analyze the

[9]The potential impact of excluding problems with negative interest rates from the problem set on parameter estimation is discussed in the latter part of Appendix F.

actual variation of the estimates in a population with the same true parameter value. Here, we assume a population in which the three parameters—δ, β, and $\ln \sigma$—are distributed on a three-dimensional grid of the ground-truth values that we set. Then, we check the distribution of estimates of each parameter in this population.

Figure 2 shows the distribution of the estimated values of δ and β as a box plot (see Appendix A for the $\ln \sigma$ estimates). Each box summarizes 10 replications of all combinations of ground-truth values of β (δ) and $\ln \sigma$, i.e., 700 simulation agents. The two ends of the box represent the first and third quartiles, respectively, and the two ends of the whiskers represent the 5th and 95th percentiles, respectively. On the red line, the error of the estimate is 0. If the box is above or below the red line, then the estimations have less trueness. In most cases, we find that the deviations from the true value fall within the interquartile range of the estimates' distribution.

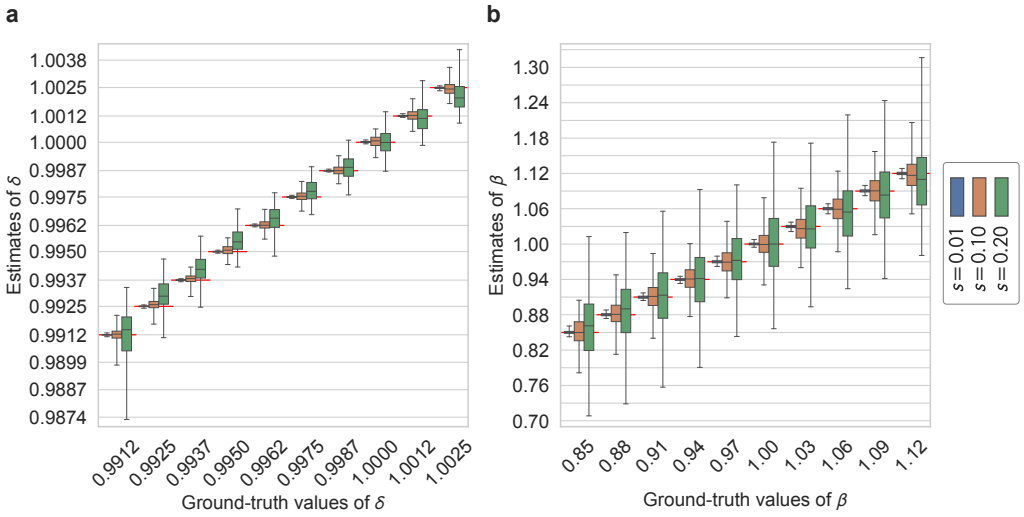


Figure 2: Box plot of the estimates

Notes: Estimates of **a**) δ and **b**) β. Each box summarizes 10 replications of all combinations of ground-truth values of β (δ) and $\ln \sigma$, i.e., 700 simulation agents. The two ends of the box represent the first and third quartiles, respectively, and the two ends of the whiskers represent the 5th and 95th percentiles, respectively. On the red line, the error of the estimate is 0.

In addition to the trueness of estimation, we should understand the resolution of the estimates. If the estimation is obtained using a higher resolution, we can precisely distinguish between any two individuals, even if the actual parameter values are in close proximity to one another. In this context, the resolution, defined as the minimum distance between actual parameter values, can be deemed identifiable by comparing the lengths of the boxes (i.e., interquartile range).

For the discount factor parameter δ, Figure 2 shows that the whiskers of the

estimates for any two adjacent ground-truths do not overlap and can be distinguished from each other for the smallest noise level $s = 0.01$. Even for the most extensive noise $s = 0.20$, the boxes do not overlap, whereas the whiskers do. We conclude that the experimental tasks considered in our simulations have enough resolution that as long as the distance between the true δ values of any two individuals is at least the ground-truth value spacing (1.3×10^{-3}), then we can distinguish between them, even assuming relatively large amounts of noise.

In contrast to the case of δ, Figure 2 reveals that the resolution of the present/future bias parameter β is generally not high. For $s = 0.01$, the whiskers for any two adjacent ground-truths do not overlap in most cases and can be barely distinguished. However, whiskers and boxes often overlap when the noise is more prominent than for $s = 0.01$. For $s = 0.20$, the boxes overlap unless the true values of β are at least 0.1 away from each other. In the case of β, unlike the case of δ, we found that when comparing the magnitude of β for any two individuals using the experimental task we are addressing, the two individuals cannot be distinguished unless their true β values are farther apart than normally assumed.

Relative to the range of the prior distribution of β that we usually assume, the significant variance of the estimates suggests the possibility of errors. It has been argued that focusing only on statistically significant results using low power statistical tests can lead to overestimation of effect sizes (van Zwet and Cator, 2021). A meta-analysis of estimations of the present bias parameter indicated that the reported effect is strong, such that it is suspected to be a publication bias in studies based on real effort tasks (Imai et al., 2020). Our results raise further concerns regarding the overestimation of the present bias effect because greater noise in the estimation produces lower power in the statistical tests.

3.3 Why Is the Present Bias Estimation Resolution Low?

In the CES-QHD utility function, δ and β appeared as the term $D = \beta\delta^k$ for $t = 0$ and as $D = \delta^k$ for $t > 0$. If the available data for parameter estimation is only for the case of $t = 0$, we cannot uniquely identify δ and β. As we indeed have data for both cases, $t = 0$ and $t > 0$, we should be able to identify the parameters mathematically.

Figure 3 shows a scatter plot of the estimated values of δ and β (for $\ln \sigma = 2.67$ and $s = 0.01$; see Appendix B for the scatter plots including all $\ln \sigma$ and s) and a red line that satisfies $\beta\delta^{70} = 1$. Note that both axes use a logarithmic scale centered at 1 and that all points have been offset so that the ground-truth values coincide

with $\delta = \beta = 1$ (indicated by the red cross). What is interesting in Figure 3 is that the points are distributed along the red line. Theoretically, it should be possible to identify δ and β; however, in practice, it is difficult even though the value of D itself can be estimated with reasonable precision.

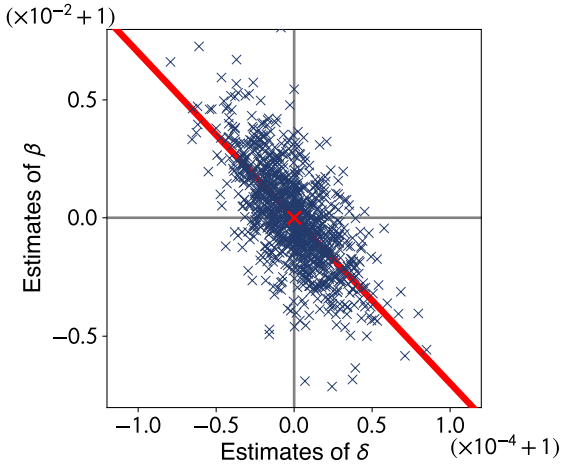


Figure 3: Scatter plot of estimated δ and β

Notes: The figure shows the case $\ln \sigma = 2.67$ and $s = 0.01$. Each point is shifted so that the pair of corresponding ground-truth values coincides with the red marked point $(\delta, \beta) = (1, 1)$. Both axes are on a logarithmic scale centered at 1. On the red line, $\beta\delta^{70} = 1$ is satisfied.

As $\mathrm{d}D/D = \mathrm{d}\beta/\beta + k\,\mathrm{d}\delta/\delta$, a 1% change in β results in a 1% change in D, but a 1% change in δ results in a k% change in D. As the difference between $\delta = 1$ and $\delta = 0.9987$ is 0.13%, the variation in D is 9.1% for $k = 70$. However, $\beta = 0.97$ is 3% smaller than $\beta = 1$ and yields a 3% variation for D, which is three times smaller than that for the δ case.

We can understand the effects of the parameters by depicting the demand curves for several combinations of parameters because the effect of the variation in D on decisions (or the demand function) depends on the curvature parameter $\ln \sigma$ and price $1 + r$. Figure 4 shows the demand curve representing the relationship between the price $1 + r$ and the amount that individuals are willing to allocate to the sooner period for $\ln \sigma = 2.67$. Note that the horizontal axis representing price $1 + r$ uses a logarithmic scale and that the prices are indicated by the vertical lines in the figure. In Figure 4, we can compare the differences in decisions between individuals with $\ln \sigma = 2.67$ and different δ and β. The difference in behavior when only δ decreases from 1 to 0.9987 is the difference between the blue and orange dashed curves. The difference when only β decreases from 1 to 0.97 is the difference between

the blue and green dotted curves. We see that the discount behavior of $\delta = 0.9987$ is more significant than that of $\beta = 0.97$. Given the noise, it is more challenging to test whether the estimated β is less than 1 for an individual whose true β is 0.97 than whether the estimated δ is less than 1 for an individual whose true δ is 0.9987, because the difference in decisions is three times smaller. In the previous subsection, we observed that assuming significant noise $s = 0.20$, the resolution of δ is about 1.3×10^{-3}, which corresponds to the spacing of our ground-truth values, while the resolution of β is 0.1, which corresponds to about three times the spacing of our ground-truth values.

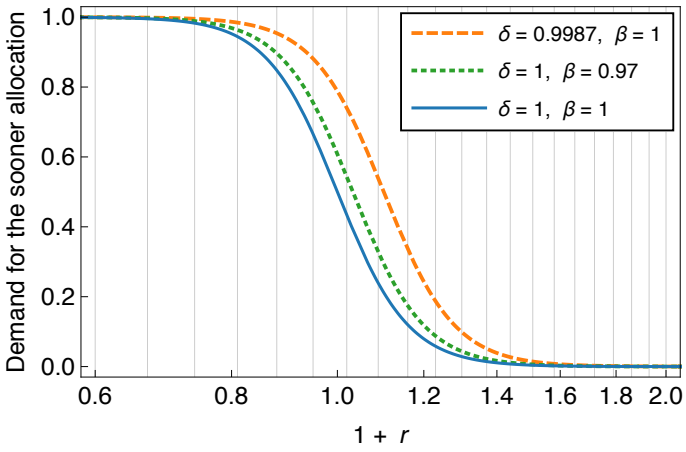


Figure 4: Demand curves

Notes: Each demand curve is for an individual whose curvature parameter is $\ln \sigma = 2.67$. The demand curve represents the relationship between the price $1 + r$ and the amount individuals are willing to allocate to the sooner period. The horizontal axis representing price $1 + r$ uses a logarithmic scale. Note that the individual faces the decision problem of allocating between now ($t = 0$) and $k = 70$ days later with the prices indicated by the vertical lines.

Eventually, the low resolution of β estimation occurs because we try to identify values within a very narrow scope with high precision. As is clear from the comparison of demand curves in Figure 4, when the true difference in the values of β is less than 0.1, it is inherently difficult to identify individuals regardless of the econometric method used because the differences in behavior are small (see Appendix C). In the expanded β scope, it is possible to distinguish between two individuals with given ground-truth values (see Appendix D). Although δ intuitively seems to require a very high resolution because it is a daily discount factor and then values should be accurately estimated to the fourth decimal place, it is possible to estimate it with a sufficient resolution because its scope is broader than that of β. Because k depends on the scale of δ, we must expand the scope of δ if we make k a weekly

discount factor. It should be noted that changing the value of k does not improve the resolution of β's estimation (see Appendix E).

4 Discussion

This paper evaluates the inaccuracy of the estimates for the CES-QHD utility parameters obtained using the CTB experiment (Andreoni and Sprenger, 2012a) by performing parameter recovery simulations (Wilson and Collins, 2019). Figures 1 and 2 demonstrate that the precision of the estimation of the time discount factor δ is sufficient enough to distinguish between $\delta = 0.9987$ and $\delta = 1$. However, the precision of the estimation of β, which represents the present/future bias, is inadequate. It is more challenging to infer that the estimated value of $\beta = 0.97$ is smaller than 1, in comparison with the estimated $\delta = 0.9987$. Our analysis reveals that CTB experiments have attempted to identify small differences in β that were inherently indistinguishable.

Considering the variations in behavior that correspond to the differences in parameter values (as depicted in Figure 4), the true value of β must be less than 0.9 when the estimation is less than 1. In other words, when trying to detect present bias using the problem set used in our simulations, it can only be detected for individuals who discount future payoffs by more than 10%. Given the low resolution of the β estimation, there is a possibility of overestimating or underestimating the effect of behavioral bias by chance, which can make publication bias more problematic.

Although variations in behavior may be subtle and obscured by noise, including more tasks can counteract the impact of noise and enhance the accuracy of estimation. However, it would be impracticable to increase the number of tasks further because of the participants' workload during the experiment.[10] In reality, the CTB experiments conducted subsequent to the original study (Andreoni and Sprenger, 2012a) have generally reduced the number of tasks (Imai et al., 2020).

It may be feasible to enhance the precision of estimation by modifying the design of the problem set, rather than increasing the number of tasks. The variables that can be manipulated in task generation include the sooner date t, the delay period k,

[10] A method for adaptive task generation proposed by Imai and Camerer (2018) could potentially provide a solution to efficiently obtain high-resolution parameter estimation without the need to increase the overall number of tasks.

and the price ratio $1 + r$. In particular, there appears to be scope for improving the generation of price variations.

The problem set used in our simulations includes prices below 1, which implies negative interest rates. This is because our aim is to consider not only individuals who discount future payoffs, but also those who place a premium on such payoffs. If we assume it is possible to disregard atypical individuals who place a premium and exclude them from analyses, we can raise the number of tasks in positive interest rate domains by reducing the number of tasks in negative interest rate domains.

We have attempted to improve the accuracy of the estimation by altering the design of problem sets (see Appendix F). We found that increasing the number of tasks indisputably improves the estimation resolution.[11] However, attempting to modify the delay period k or the price ratio $1 + r$ without increasing the number of tasks did not result in significant improvements in resolution.

The difficulty in estimating β primarily stems from the mathematical structure of the CES-QHD utility model combined with the CTB experimental tasks. Note that it is entirely possible to discern differences in behavior by actual humans in CTB experiments, which can be detected as outcomes of present bias—these behaviors may not be captured by the CES-QHD utility model.[12] We believe that researchers who persist in using the CTB method will necessitate a significant overhaul of behavior modeling. We additionally recommend the use of parameter recovery simulations.

[11] If it is acceptable for us to disregard the individual heterogeneity and all data of a single population can be merged for parameter estimation, then a significant amount of data is available for accurate estimation.

[12] For example, the analyses summarized in Table II of Augenblick et al. (2015) and Table 2 of Cheung et al. (2022) attempt to assess the magnitude of the present bias effect without employing the parameters of the CES-QHD utility function.

References

Abdellaoui, M., H. Bleichrodt, O. L'Haridon, and C. Paraschiv (2013): "Is there one unifying concept of utility? An experimental comparison of utility under risk and utility over time," *Management Science*, 59, 2153–2169.

Andersen, S., G. W. Harrison, M. I. Lau, and E. E. Rutström (2008): "Eliciting Risk and Time Preferences," *Econometrica*, 76, 583–618.

——— (2014): "Discounting behavior: A reconsideration," *European Economic Review*, 71, 15–33.

Andreoni, J., M. A. Kuhn, and C. Sprenger (2015): "Measuring time preferences: A comparison of experimental methods," *Journal of Economic Behavior & Organization*, 116, 451–464.

Andreoni, J. and C. Sprenger (2012a): "Estimating Time Preferences from Convex Budgets," *American Economic Review*, 102, 3333–3356.

——— (2012b): "Risk Preferences Are Not Time Preferences," *American Economic Review*, 102, 3357–3376.

——— (2015): "Risk preferences are not time preferences: Reply," *American Economic Review*, 105, 2287–2293.

Augenblick, N., M. Niederle, and C. Sprenger (2015): "Working over Time: Dynamic Inconsistency in Real Effort Tasks," *The Quarterly Journal of Economics*, 130, 1067–1115.

Blumenstock, J., M. Callen, and T. Ghani (2018): "Why Do Defaults Affect Behavior? Experimental Evidence from Afghanistan," *American Economic Review*, 108, 2868–2901.

Camerer, C. F., A. Dreber, E. Forsell, T. H. Ho, J. Huber, M. Johannesson, M. Kirchler, J. Almenberg, A. Altmejd, T. Chan, E. Heikensten, F. Holzmeister, T. Imai, S. Isaksson, G. Nave, T. Pfeiffer, M. Razen, and H. Wu (2016): "Evaluating replicability of laboratory experiments in economics," *Science*, 351, 1433–1436.

Carvalho, L. S., S. Meier, and S. W. Wang (2016): "Poverty and Economic Decision-Making: Evidence from Changes in Financial Resources at Payday," *American Economic Review*, 106, 260–84.

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

——— (2020): "Eliciting utility curvature in time preference," *Experimental Economics*, 23, 493–525.

Cheung, S. L., A. Tymula, and X. Wang (2021): "Quasi-hyperbolic Present Bias: A Meta-analysis," Life Course Centre Working Paper No. 2021-15, Life Course Centre.

——— (2022): "Present bias for monetary and dietary rewards," *Experimental Economics*, 25, 1202–1233.

Dantas, A. M., A. T. Sack, E. Bruggen, P. Jiao, and T. Schuhmann (2022): "The effects of probiotics on risk and time preferences," *Scientific Reports*, 12, 12152.

Epper, T. and F. D. Helga (2015): "Comment on "Risk Preferences Are Not Time Preferences": Balancing on a Budget Line," *American Economic Review*, 105, 2261–2271.

Harrison, G. W., M. I. Lau, and E. E. Rutström (2013): "Identifying time preferences with experiments: Comment," `https://cear.gsu.edu/wp-2013-09-identifying-time-preferences-with-experiments-comment/`. Accessed November 9, 2022.

Holt, C. A. and S. K. Laury (2002): "Risk Aversion and Incentive Effects," *American Economic Review*, 92, 1644–1655.

Imai, T. and C. F. Camerer (2018): "Estimating Time Preferences from Budget Set Choices Using Optimal Adaptive Design," `https://www.taisukeimai.com/api/resources/adaptive_ctb.pdf`. Accessed November 30, 2022.

Imai, T., T. A. Rutter, and C. F. Camerer (2020): "Meta-Analysis of Present-Bias Estimation using Convex Time Budgets," *The Economic Journal*, 131, 1788–1814.

Inukai, K., Y. Shimodaira, and K. Shiozawa (2022): "Revisiting CES utility functions for distributional preferences: Do people face the equality–efficiency

trade-off?" ISER Discussion Paper No. 1195, Institute of Social and Economic Research, Osaka University.

Laibson, D. (1997): "Golden Eggs and Hyperbolic Discounting," *The Quarterly Journal of Economics*, 112, 443–478.

Miao, B. and S. Zhong (2015): "Comment on "Risk Preferences Are Not Time Preferences": Separating Risk and Time Preference," *American Economic Review*, 105, 2272–2286.

O'Donoghue, T. and M. Rabin (1999): "Doing It Now or Later," *American Economic Review*, 89, 103–124.

——— (2015): "Present Bias: Lessons Learned and To Be Learned," *American Economic Review*, 105, 273–279.

Takeuchi, K. (2012): "Time Discounting: The Concavity of Time Discount Function: An Experimental Study," *Journal of Behavioral Economics and Finance*, 5, 2–9.

Thöni, C. (2015): "A note on CES functions," *Journal of Behavioral and Experimental Economics*, 59, 85–87.

van Zwet, E. W. and E. A. Cator (2021): "The significance filter, the winner's curse and the need to shrink," *Statistica Neerlandica*, 75, 437–452.

Wilson, R. C. and A. G. Collins (2019): "Ten simple rules for the computational modeling of behavioral data," *eLife*, 8, e49547.

A Box Plots of the $\ln\sigma$ Estimates

Figure A.1 illustrates the distribution of the estimated curvature parameter $\ln\sigma$ through box plots, which were not included in the main text. Each box summarizes 10 replications of all combinations of ground-truth values of δ and β, representing a total of 1,000 simulation agents. The first and third quartiles are depicted at the two ends of the box, while the 5th and 95th percentiles are represented by the two ends of the whiskers. The error of the estimate is 0 for the red line.

To ensure convergence of the $\ln\sigma$ estimates, we used the sigmoid function $\ln\sigma = f(\theta) = 4\tanh(\theta) + 1.5$ to transform $\ln\sigma$, after which we searched for the latent variable θ. This manipulation allowed the estimated $\ln\sigma$ to range between -2.5 and 5.5. The black horizontal dashed lines in the figure represent the boundaries of $\ln\sigma$.

Unlike the parameters δ and β, the $\ln\sigma$ estimates are subject to heavy bias and tend to underestimate, particularly when the added noise is substantial. At $s = 0.20$, $\ln\sigma$ appears to saturate at approximately 2.2.

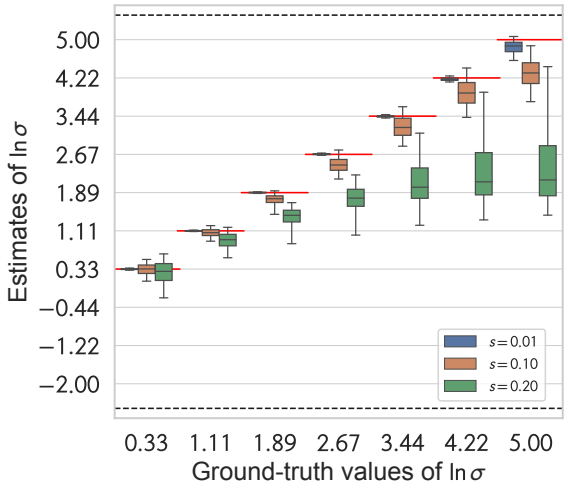


Figure A.1: Box plots of the $\ln\sigma$ estimates

B Scatter Plots of Estimated δ–β

Figure B.1 depicts scatter plots presenting the estimated δ and β for each pair of specific ground-truth δ and β. Each δ–β plot comprises estimates for all combinations of ground-truth $\ln \sigma$ and s. The color intensity of the dots represents the magnitude of the added noise, with darker dots representing smaller noise. The horizontal and vertical axes signify the δ and β estimates, respectively, while the grid displays the ground-truth values. Note that although the ground-truth values are linearly equally spaced, both axes adopt a logarithmic scale, resulting in unevenly spaced grid lines. Figure B.1 is a matrix, with columns denoting ground-truth δ and rows representing ground-truth β. Each pair of ground-truth values is indicated by a red cross.

To better understand the relationship between δ and β, we have drawn a red curve, which follows a straight line on log–log graphs, that satisfies $\beta\delta^{70} = \text{const.}$ and passes through the red crosses. The scatter plot points are either located near the red line or in a more vertical distribution than the red line.

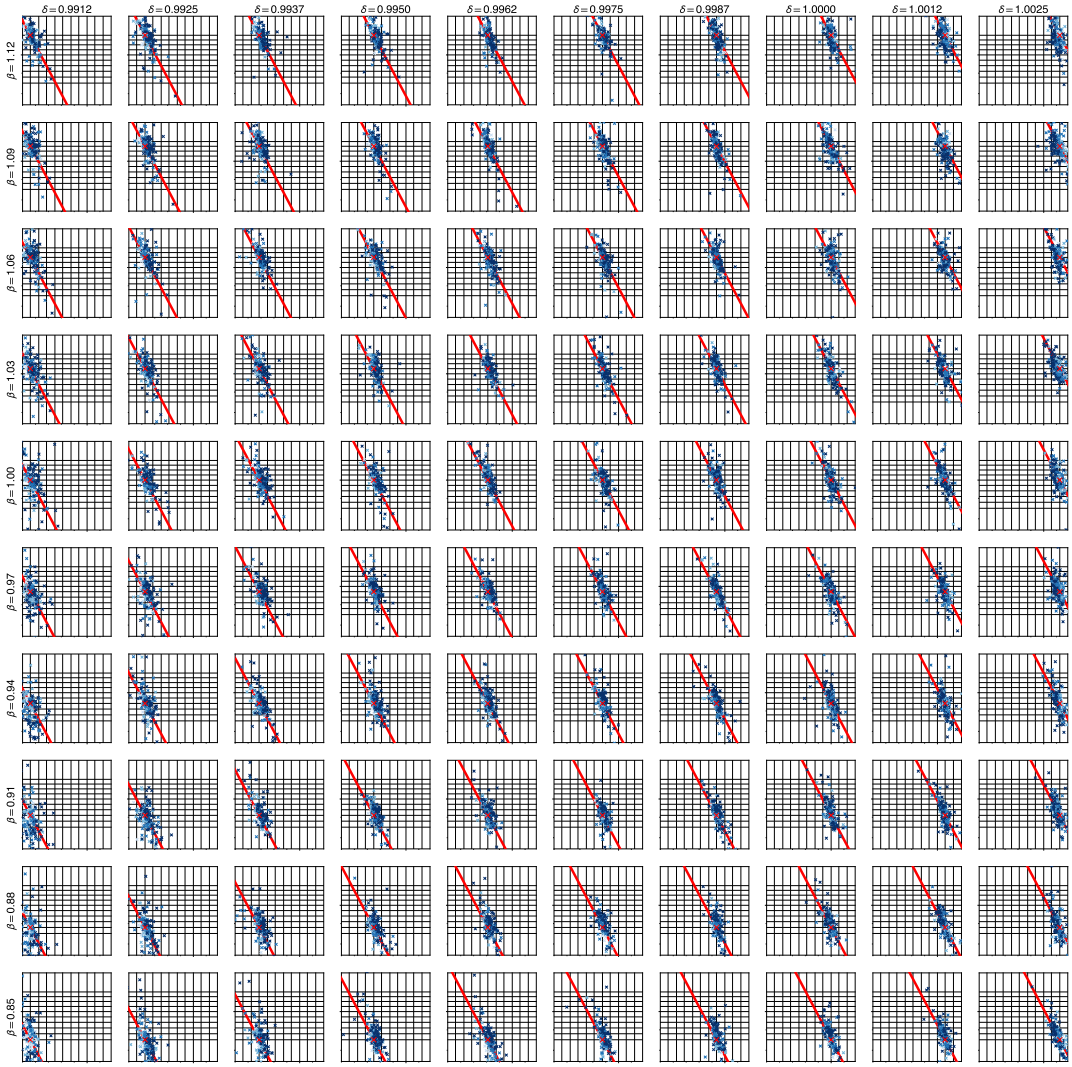


Figure B.1: Scatter plots of estimated δ–β

C Detecting the Present Bias by Using a Paired Two-sample t-test

Here, we identify the existence of present bias by comparing the decision data values directly rather than by estimating the β parameter. We conducted a two-sided paired t-test to compare the two series of decisions, with one set for $t = 0$ and another for $t > 0$, across 21 different prices. Figure C.1 displays the proportion of cases in which the null hypothesis is rejected, in which there was no difference in decisions at the 5% significance level. Compared with the success rates observed using the β estimates in the main analysis (see the panel **b** of Figure 1), detecting present bias by directly comparing decisions is more challenging.

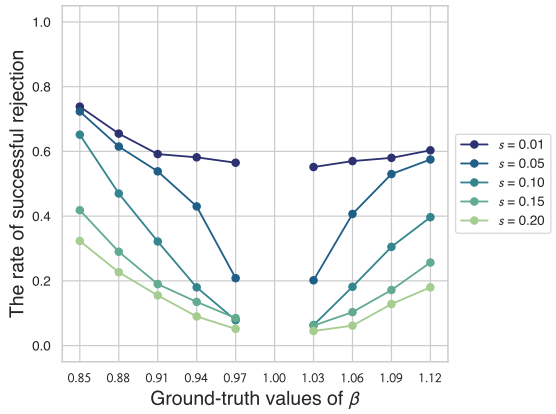


Figure C.1: Rate of successful rejection of the null hypothesis in which there was no difference in decisions between $t = 0$ and $t > 0$

D Re-running the Simulation with the Expanded Ground-truth β Spacing

In the main analysis, we concluded that the low precision of the β estimation was due to the narrow range of the ground-truth values, which made identification difficult. Herein, we present the outcomes of re-executing the simulation with the ground-truth β spacing expanded by roughly three times. To this end, we used 10 evenly spaced values from the range $0.50 \leq \beta \leq 1.40$ for the ground-truth β and the same ground-truth values for δ and $\ln \sigma$ as in the main analysis.

Figure D.1 shows box plots illustrating the dispersion of the estimated δ, β, and $\ln \sigma$. The distributions of the δ and $\ln \sigma$ estimates remain largely unchanged from those in the primary analysis, whereas the box length in the distribution of the β estimates indicates that they are precise enough to distinguish between neighbors even when the noise size is $s = 0.20$.

Figure D.2 displays the successful null hypothesis rejection rates for both $\widehat{\delta} = 1$ and $\widehat{\beta} = 1$. Compared with panel **b** of Figure 1, the lines in Figure D.2 are shifted upward overall, giving the impression of an improved success rate. It must be noted, however, that we have only expanded the range of the ground-truth β that we consider, and nothing has been done to improve the precision of the estimation. Ultimately, the primary takeaway is that the conventional CTB experiment can only measure β with a rough resolution; however, this method has been used to examine marginal differences within a narrow range.

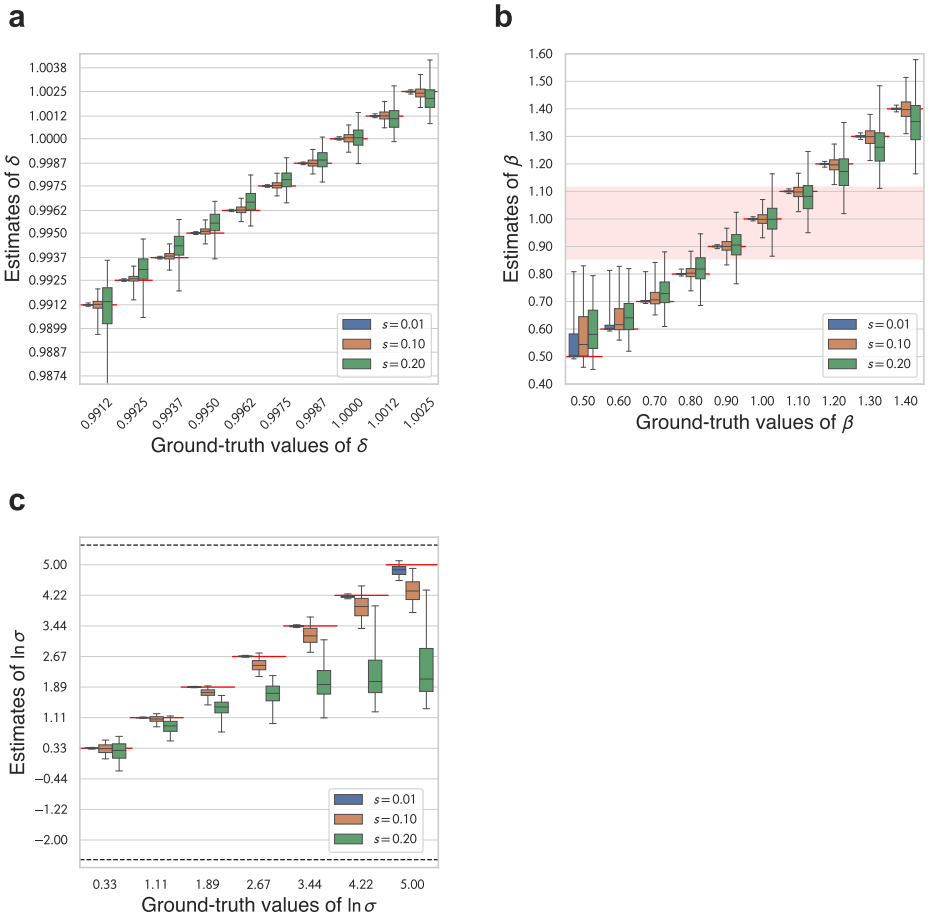


Figure D.1: Box plots of estimates for the expanded ground-truth β range

Notes: Estimates of **a**) δ, **b**) β, and **c**) $\ln\sigma$. Shaded area in panel **b** represents the range of unexpanded ground-truth β: $0.85 \leq \beta \leq 1.12$.

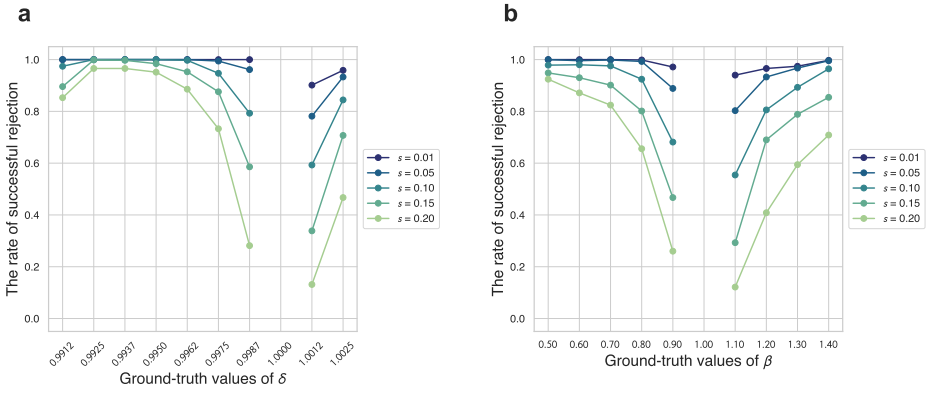


Figure D.2: Rate of successful rejection for the expanded ground-truth β range

Notes: Tests on **a**) $\widehat{\delta} = 1$ and **b**) $\widehat{\beta} = 1$.

E Re-running the Simulation with $k = 1$

The magnitude of k, representing the duration of the intertemporal period, is reliant on the scale of δ and the value of k does not inherently enhance parameter estimation accuracy. Suppose we let $k = 1$ for one day. Figures E.1 and E.2 present the results of the recovery simulation with data produced under the same ground-truth values as in the primary analysis. The precision of δ estimation is significantly low. Hence, in an experimental arrangement with only one day between periods, it is difficult to estimate δ within the given ground-truth range. When compared with panel **b** of Figure 1, we observe that the precision of β estimation is nearly the same for both $k = 70$ and $k = 1$.

Let us consider a scenario where k is set to one "period", where one period corresponds to 70 days and δ represents a discount factor for one period instead of a daily discount factor. To be more precise, we set k to 1 and generate data by raising the ground-truth value of δ used in the main analysis to the power of 70. The results are shown in Figures E.3 and E.4. Comparing these figures to Figure 1, we observe that the resolution of β remains unchanged, and there is only a slight difference in the resolution of δ over the transformed range.

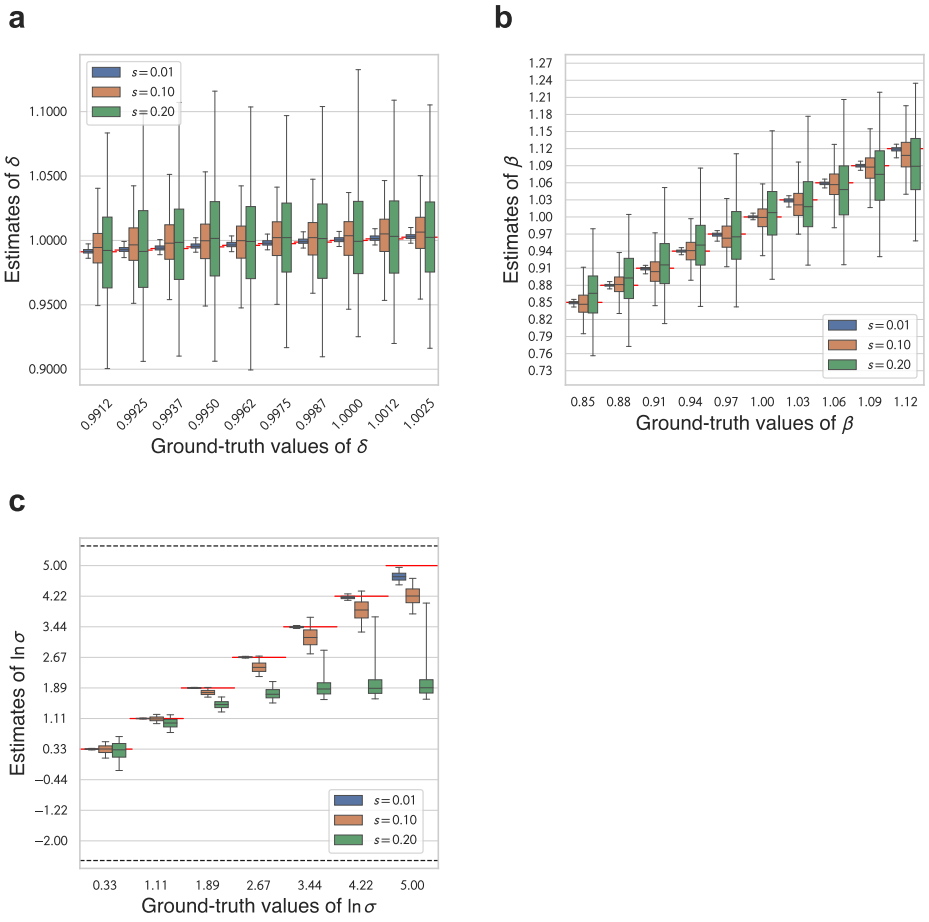


Figure E.1: Box plots of estimates for $k = 1$ without expanding ground-truth δ range

Notes: Estimates of **a**) δ, **b**) β, and **c**) $\ln \sigma$.

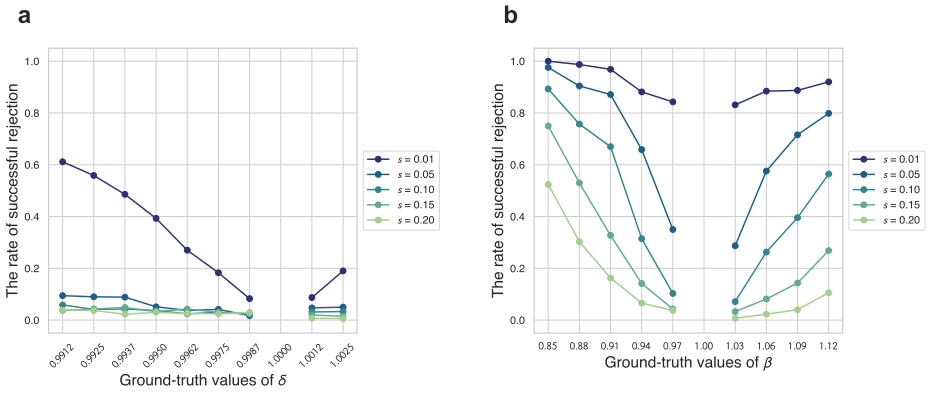


Figure E.2: Rate of successful rejection for $k = 1$ without expanding ground-truth δ range

Notes: Tests on **a**) $\widehat{\delta} = 1$ and **b**) $\widehat{\beta} = 1$.

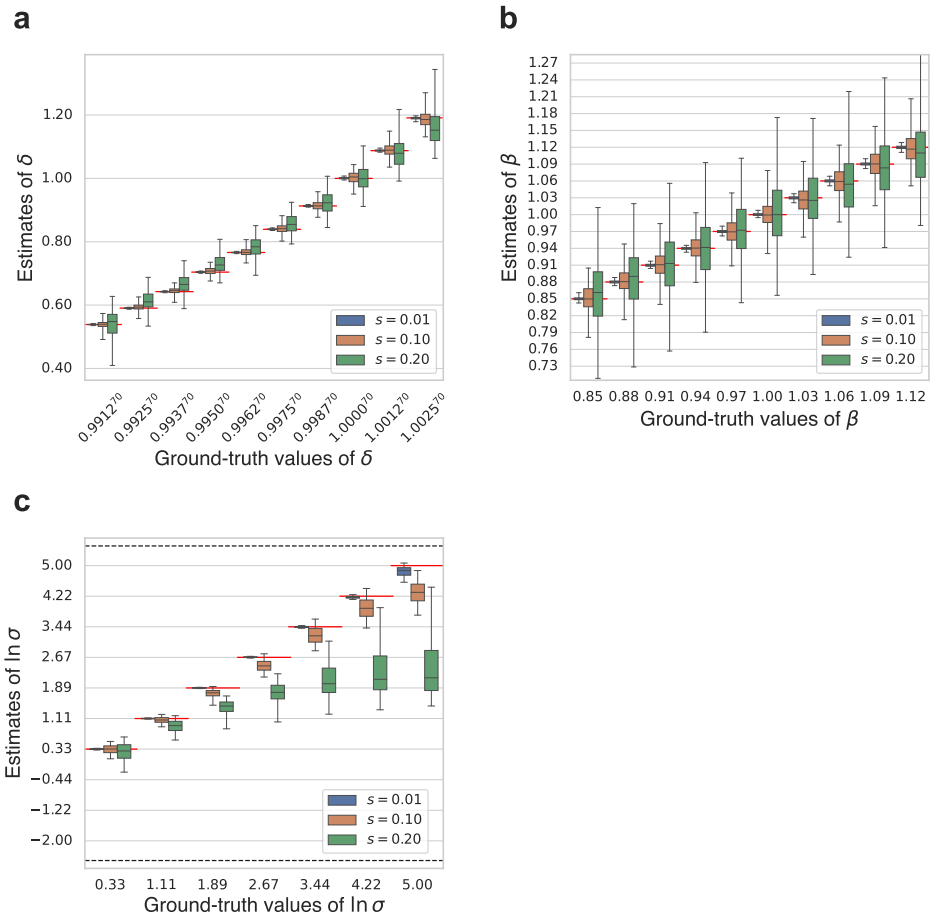


Figure E.3: Box plots of estimates for $k = 1$ with expanding ground-truth δ range

Notes: Estimates of **a**) δ, **b**) β, and **c**) $\ln \sigma$.

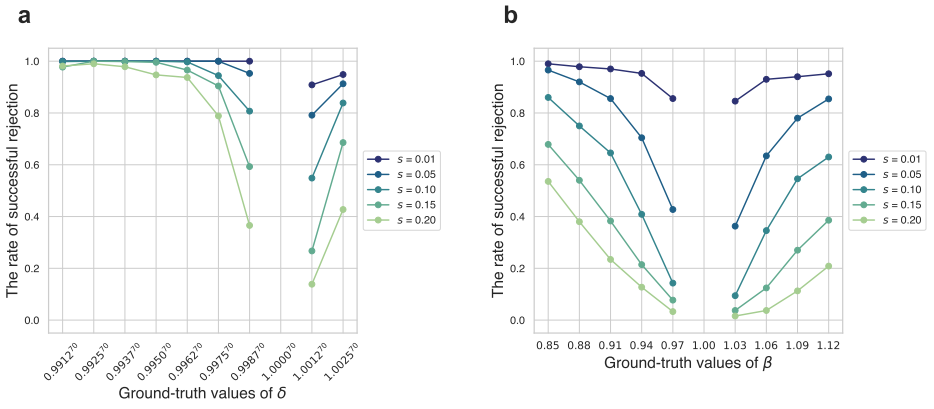


Figure E.4: Rate of successful rejection for $k = 1$ with expanding ground-truth δ range

Notes: Tests on **a**) $\widehat{\delta} = 1$ and **b**) $\widehat{\beta} = 1$.

F Design of Problem Set

In this section, we examine the effect of the problem set design on the accuracy of estimation. The main analysis uses a problem set called PS0, where m is fixed at 20, k is fixed at 70, and 21 uniformly spaced prices were drawn from the range of $0.6 \leq 1 + r \leq 2$ for $t = 0$ and $t = 1$, respectively, for a total of 42 problems.

To begin with, we investigate the impact of merely increasing the number of problems. For PS1, we draw 42 prices, twice the number of PS0, from the same price range. For PS2, we draw 210 prices, 10 times the number of PS0, from the same price range.

We then consider the strategy of altering the number of types for each variable without increasing the total number of problems. For PS3, we set the number of prices $1+r$ to seven for each combination of t and k, instead of increasing the number of k to three (35, 70, and 98). PS4 differs from PS3 in that the values of k are altered to 35, 175, and 350. For PS5, we set the number of prices $1 + r$ to three (0.9, 1.2, and 1.5) for each combination of t and k, instead of increasing the number of k to seven (14, 28, 42, 56, 70, 84, and 98). For PS6, we fixed k to 70 and draw 14 prices from the range $0.6 \leq 1 + r \leq 2$, once at $t = 0$ and twice at $t = 1$. This corresponds to setting two different t situations for $t > 0$ (i.e., two decision makings for $k = 70$: one between 7 and 77 days later and the other between 35 and 105 days later). For PS7, we neglected the existence of individuals who place a premium on future payoffs and did not consider negative interest rates. We fixed k to 70 and drew 21 prices from the range $1.05 \leq 1 + r \leq 2$ for $t = 0$ and $t = 1$, respectively.

We conducted simulations using the problem sets PS1 to PS7 and the original problem set used by AS. AS's problem set is summarized in Table F.1.

Table F.1: Problem set of AS

t	k	$(1 + r, m)$				
0	35	(1.05, 20)	(1.11, 20)	(1.25, 20)	(1.25, 25)	(1.43, 20)
0	70	(1.05, 20)	(1.11, 20)	(1.25, 20)	(1.25, 25)	(1.43, 20)
0	98	(1.05, 20)	(1.25, 20)	(1.25, 25)	(1.54, 20)	(2.00, 20)
7	35	(1.05, 20)	(1.11, 20)	(1.25, 20)	(1.25, 25)	(1.43, 20)
7	35	(1.05, 20)	(1.11, 20)	(1.25, 20)	(1.25, 25)	(1.43, 20)
7	70	(1.00, 20)	(1.05, 20)	(1.11, 20)	(1.25, 20)	(1.43, 20)
35	70	(1.05, 20)	(1.11, 20)	(1.25, 20)	(1.25, 25)	(1.43, 20)
35	98	(1.05, 20)	(1.25, 20)	(1.25, 25)	(1.54, 20)	(2.00, 20)
35	98	(1.05, 20)	(1.25, 20)	(1.25, 25)	(1.54, 20)	(2.00, 20)

For PS1 and PS2, we computed the standard error of the estimates using the inverse of the negative Hessian. However, for PS3–PS7 and AS, we computed them using the jackknife method, as in the main analysis.

Figure F.1 displays the box plots of the estimates, and Figure F.2 illustrates the rate of successful rejection of the null hypothesis for the problem sets PS1–PS7 and AS. We also present the data of PS0 as a dashed line in Figure F.2.

Figures F.3 and F.4 illustrate the absolute magnitude of the error and the standard error of the estimate by problem set, respectively. In both figures, we have depicted the mean and median along with bootstrap 95% confidence intervals. To depict the mean value, we have used a logarithmic scale as some values are excessively large.

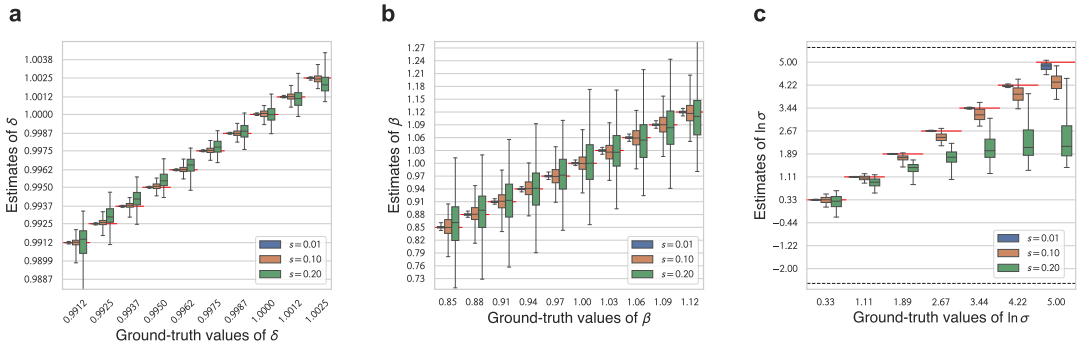
Comparing the estimation precision of PS0 with that of PS1 and PS2, it becomes apparent that the latter two exhibit an improvement in precision. This enhancement is attributed to an increase in the number of problems, as precision is seen to be positively correlated with the number of tasks. Despite PS2 comprising a total of 420 tasks, however, it remains challenging to reject the hypothesis that an individual's β estimate with a true value of 0.97 is not equal to 1 when the noise size is $s = 0.20$. With regard to PS3 to PS6, no particular problem set appears to be definitively superior to PS0. While PS4 displayed improved precision in estimating δ for ground-truth values near 1, precision worsened for ground-truth values smaller than 0.9950, and overall estimation precision for β also decreased.

For PS7 and AS, which do not involve prices with negative interest rates, the precision of the δ estimates is inferior for individuals whose ground-truth δ is greater than 1. It is acknowledged that significant errors in parameter estimation are inevitable, especially for individuals with preferences close to linear. To simplify, we assume here that an individual's preferences are represented by a completely linear utility. They allocate all tokens to a later period if the offered price exceeds a certain threshold, known as the switching point. Specifically, if $1 + r > (\beta^{\mathbf{1}_{t=0}}\delta^k)^{-1}$, they allocate all tokens to the later period; otherwise, to the sooner period. If the minimum offering price exceeds their switching point, $(\beta^{\mathbf{1}_{t=0}}\delta^k)^{-1}$, then that individual will always allocate all endowments to the later period in any problem, and thus it is impossible to estimate their switching point from the observed data. To extract the switching point for an individual who highly values future profits (i.e., whose discount factor δ is greater than 1), we need to examine whether they are willing to allocate to the later periods despite the nominal decrease in allocated payoffs, where the interest rate is negative.

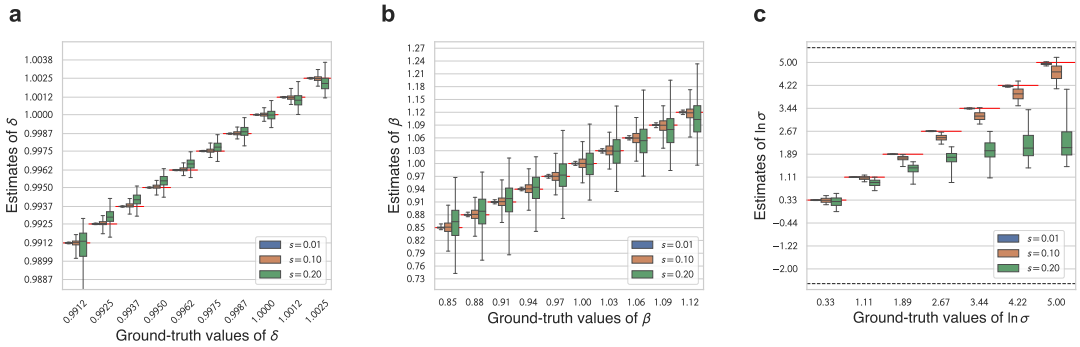
The panels located in the upper left and middle left of Figure F.5 illustrate

the regions for PS7 and AS, respectively, in which the values of δ and β satisfy the requirement that the switching point is either less than the minimum price (orange) or greater than the maximum price (blue). Note that the area is shaded for each value of k, as AS comprises three different types of k. The grid of black dots depicts the simulated ground-truth values. In the upper center and middle center panels of Figure F.5, heatmaps are presented for PS7 and AS, respectively, which exhibit the medians of the Euclidean distance between the ground-truth (δ, β) and the estimates $(\widehat{\delta}, \widehat{\beta})$ within each cell. For pairs that are darker in color, the estimation error is more significant. The northeast area of the heatmaps indicates a worsening in the accuracy of the estimation for both PS7 and AS. When comparing the left panel to the right panel, it becomes apparent that the cells are darker for parameter combinations that correspond to the shaded regions, which further reduces the estimation's accuracy. For individuals whose preferences are not relatively linear (ground-truth $\ln \sigma = 0.33, 1.11, 1.89$), there is no noticeable pattern of the estimation error being significant for a specific combination of parameters (see the upper right and middle right panels of Figure F.5).

PS0: $\{t \mid 0, 1\} \times \{k \mid 70\} \times \{1 + r \mid 0.60, 0.67, \ldots, 2.00\}$ $(\# = 42)$



PS1: $\{t \mid 0, 1\} \times \{k \mid 70\} \times \{1 + r \mid 0.60, 0.63, \ldots, 2.00\}$ $(\# = 84)$



PS2: $\{t \mid 0, 1\} \times \{k \mid 70\} \times \{1 + r \mid 0.600, 0.607, \ldots, 2.000\}$ $(\# = 420)$

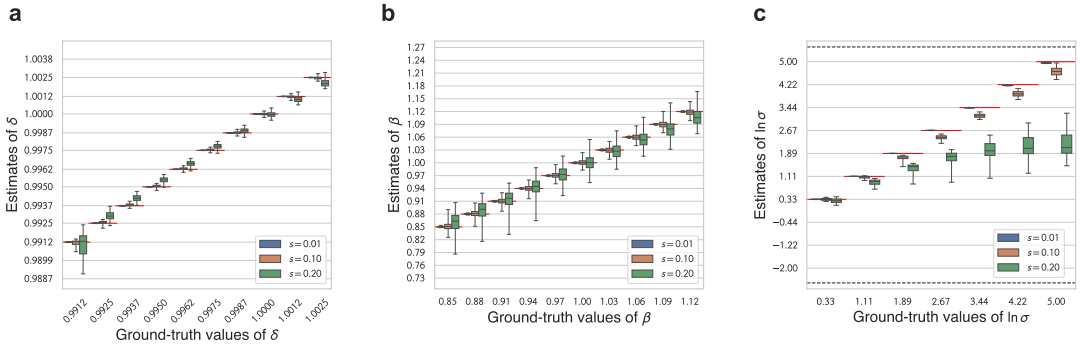
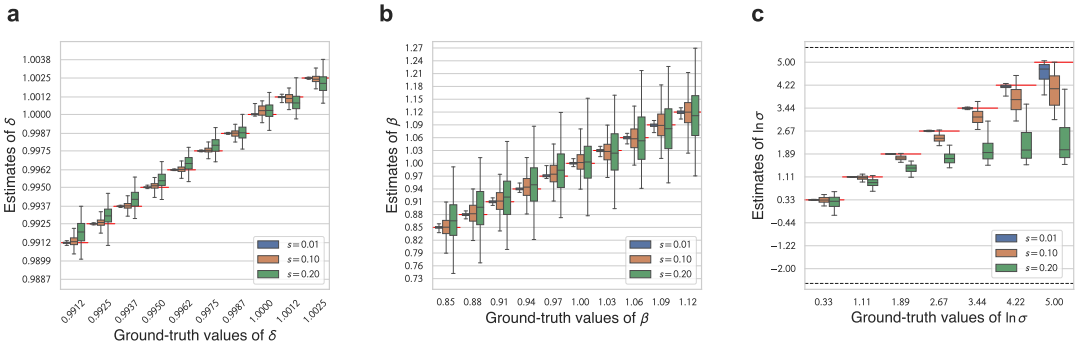


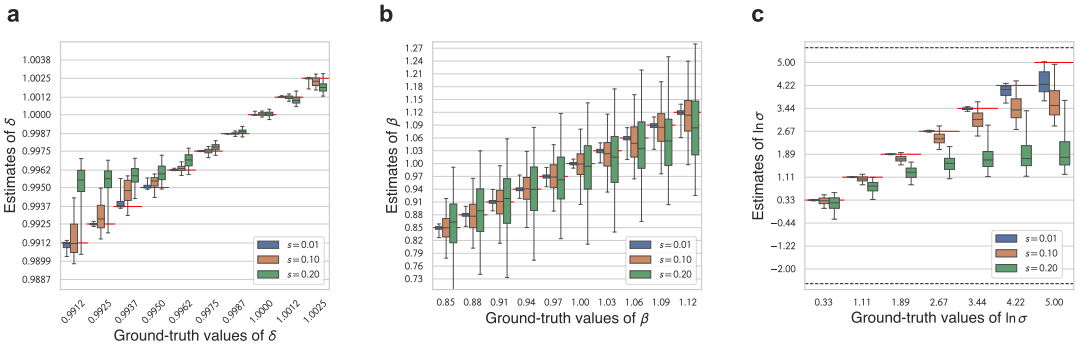
Figure F.1: Box plots of estimates for each problem set

Notes: Estimates of **a**) δ, **b**) β, and **c**) $\ln \sigma$. For PS0, panels **a** and **b** are reshown as Figure 2 and panel **c** is reshown as Figure A.1.

PS3: $\{t \mid 0, 1\} \times \{k \mid 35, 70, 98\} \times \{1 + r \mid 0.60, 0.83, \ldots, 2.00\}$ $(\# = 42)$



PS4: $\{t \mid 0, 1\} \times \{k \mid 35, 175, 350\} \times \{1 + r \mid 0.60, 0.83, \ldots, 2.00\}$ $(\# = 42)$



PS5: $\{t \mid 0, 1\} \times \{k \mid 14, 28, 42, 56, 70, 84, 98\} \times \{1 + r \mid 0.9, 1.2, 1.5\}$ $(\# = 42)$

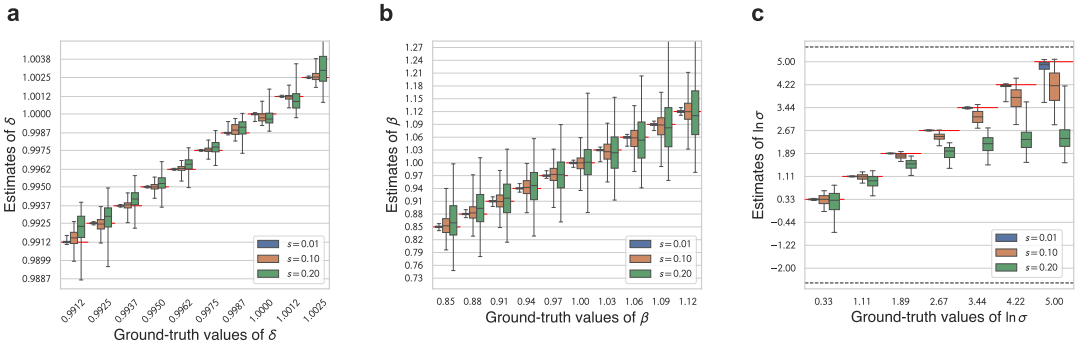
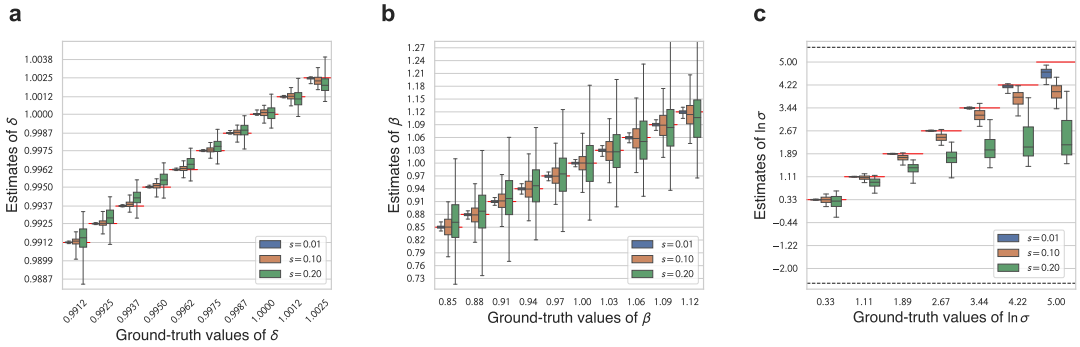
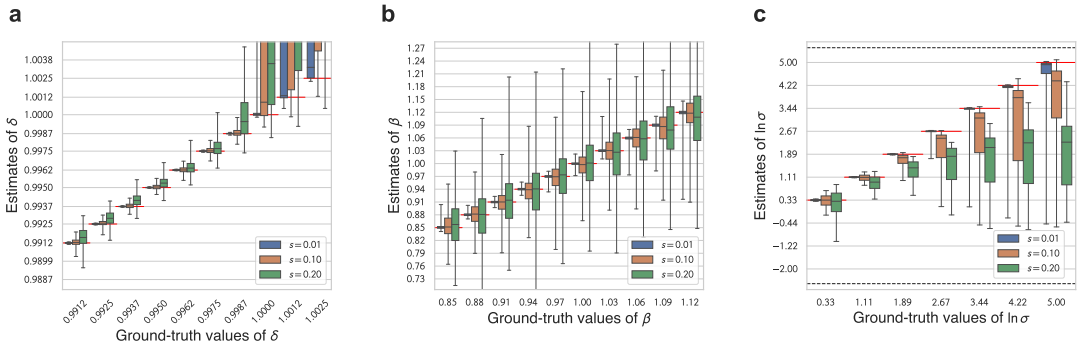


Figure F.1: Box plots of estimates for each problem set (cont'd)

PS6: $\{t \mid 0, 1, 1\} \times \{k \mid 70\} \times \{1 + r \mid 0.60, 0.71, \ldots, 2.00\}$ $(\# = 42)$



PS7: $\{t \mid 0, 1\} \times \{k \mid 70\} \times \{1 + r \mid 1.05, 1.10, \ldots, 2.00\}$ $(\# = 42)$



AS (summarized in Table F.1, $\# = 45$)

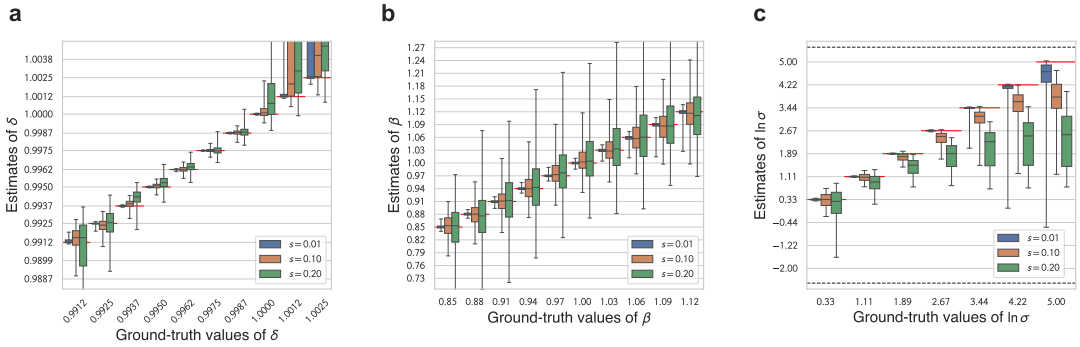
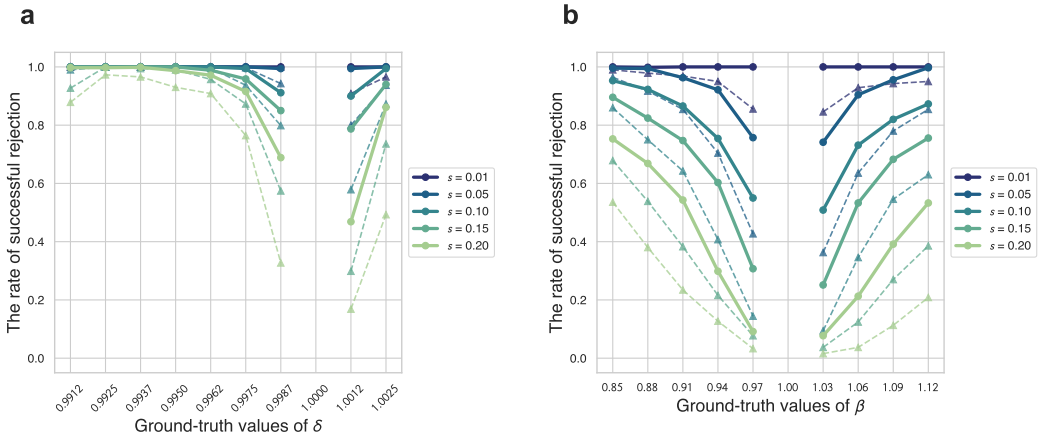


Figure F.1: Box plots of estimates for each problem set (cont'd)

PS1: $\{t \mid 0, 1\} \times \{k \mid 70\} \times \{1 + r \mid 0.60, 0.63, \ldots, 2.00\}$ $(\# = 84)$



PS2: $\{t \mid 0, 1\} \times \{k \mid 70\} \times \{1 + r \mid 0.600, 0.607, \ldots, 2.000\}$ $(\# = 420)$

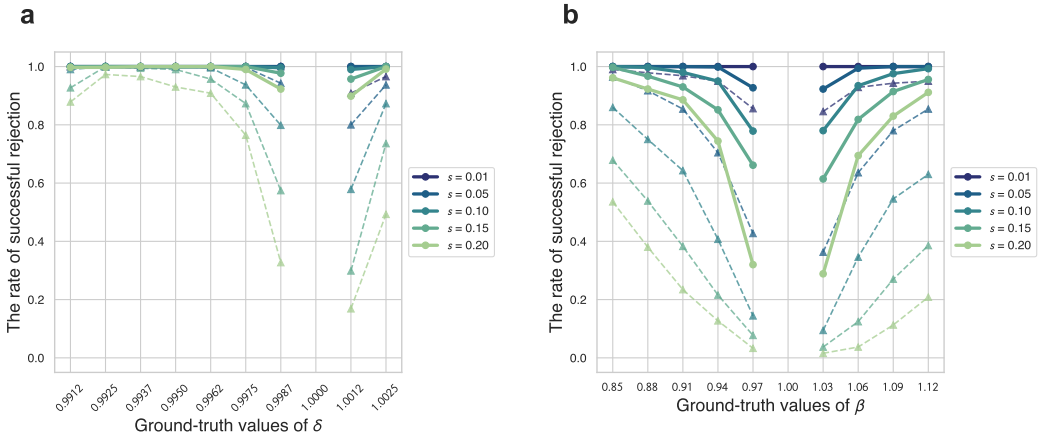
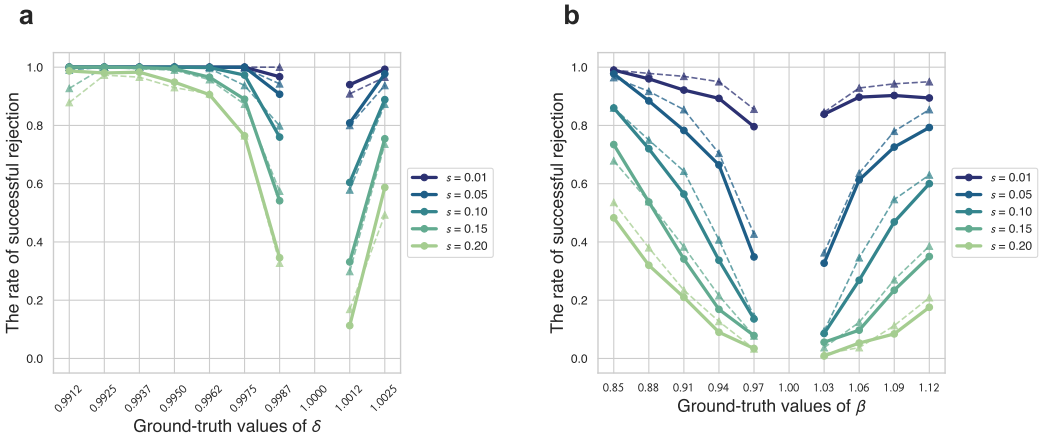


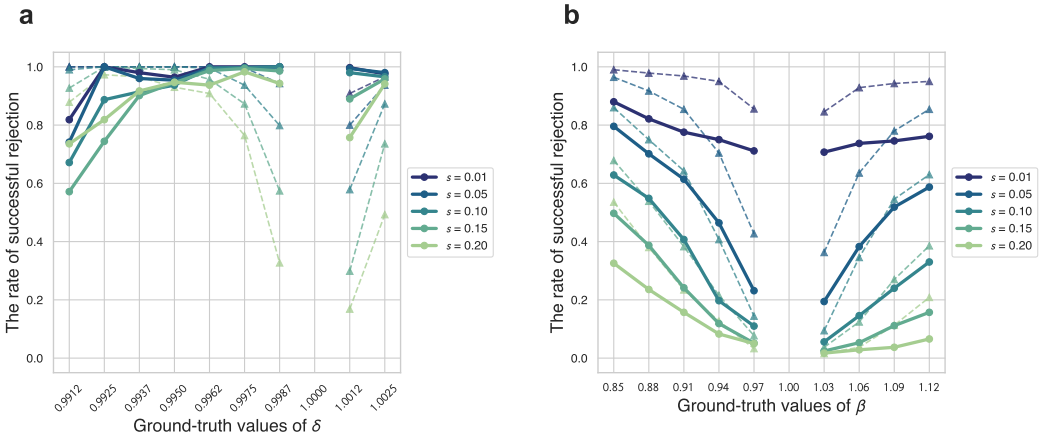
Figure F.2: Rate of successful rejection for each problem set

Notes: Tests on **a**) $\widehat{\delta} = 1$ and **b**) $\widehat{\beta} = 1$. Dashed lines are for PS0.

PS3: $\{t \mid 0, 1\} \times \{k \mid 35, 70, 98\} \times \{1 + r \mid 0.60, 0.83, \ldots, 2.00\}$ $(\# = 42)$



PS4: $\{t \mid 0, 1\} \times \{k \mid 35, 175, 350\} \times \{1 + r \mid 0.60, 0.83, \ldots, 2.00\}$ $(\# = 42)$



PS5: $\{t \mid 0, 1\} \times \{k \mid 14, 28, 42, 56, 70, 84, 98\} \times \{1 + r \mid 0.9, 1.2, 1.5\}$ $(\# = 42)$

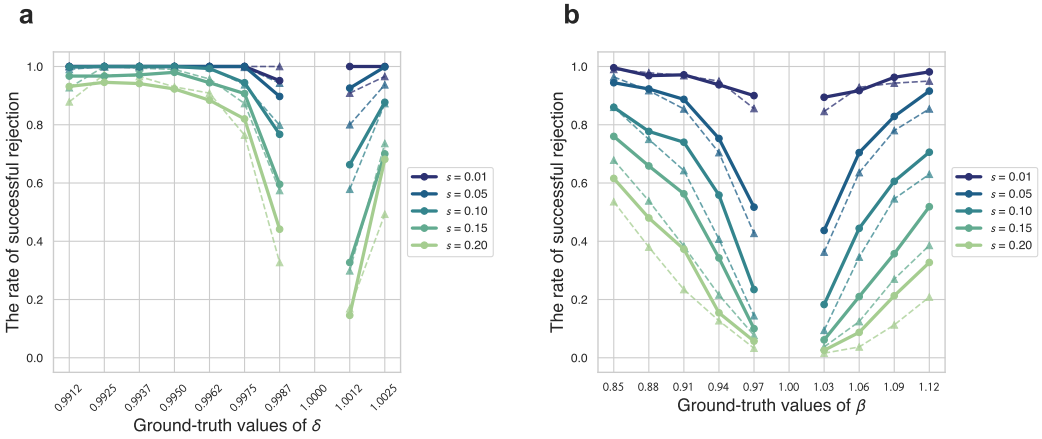
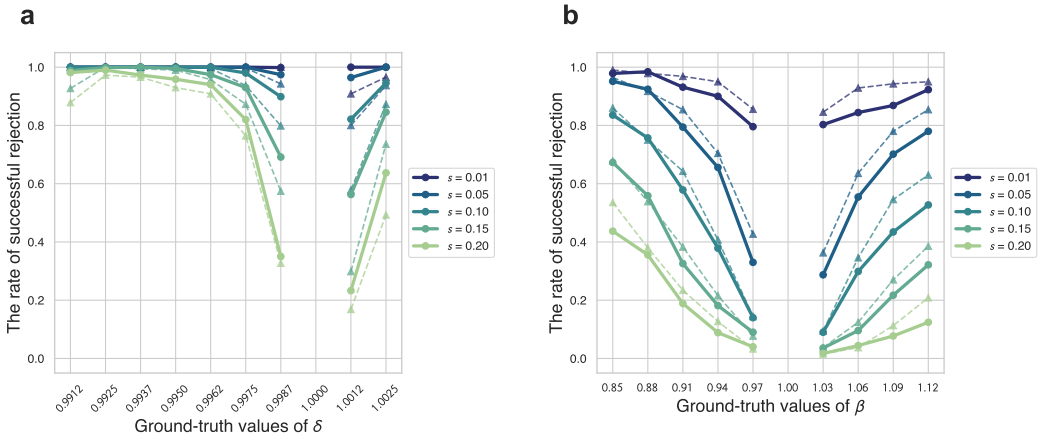
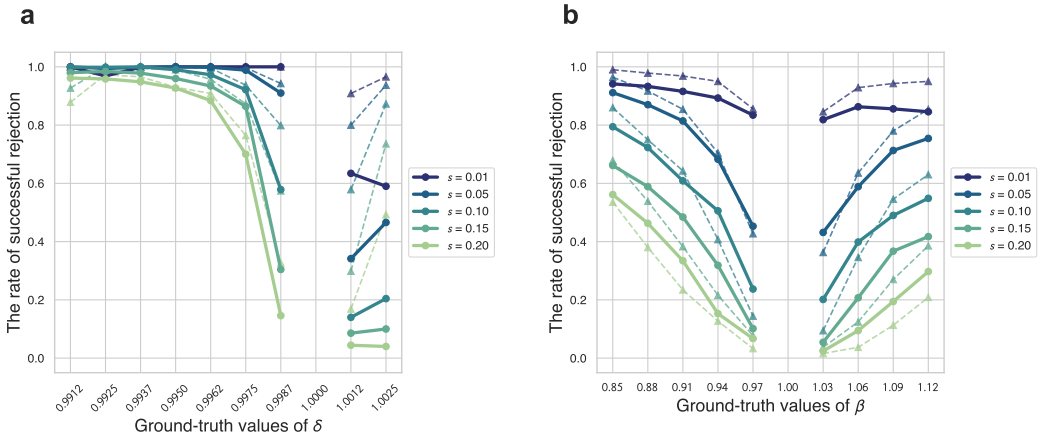


Figure F.2: Rate of successful rejection for each problem set (cont'd)

PS6: $\{t \mid 0, 1, 1\} \times \{k \mid 70\} \times \{1 + r \mid 0.60, 0.71, \ldots, 2.00\}$ $(\# = 42)$



PS7: $\{t \mid 0, 1\} \times \{k \mid 70\} \times \{1 + r \mid 1.05, 1.10, \ldots, 2.00\}$ $(\# = 42)$



AS (summarized in Table F.1, $\# = 45$)

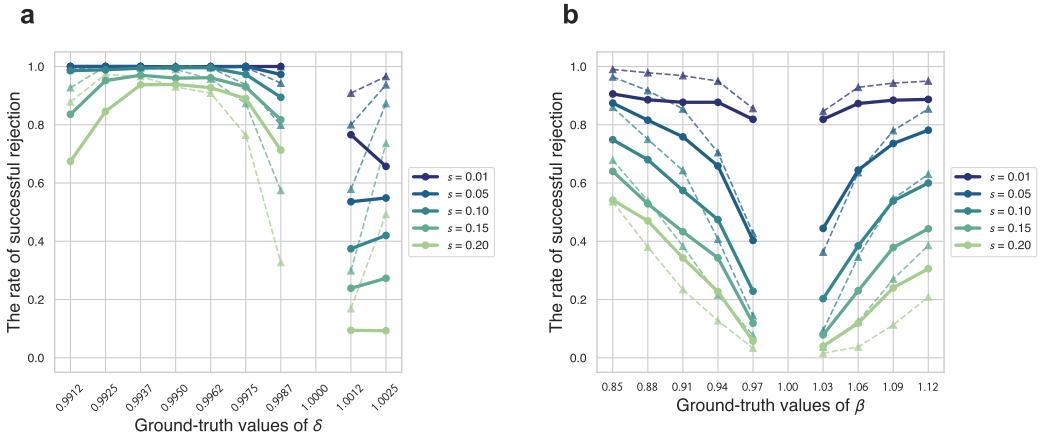


Figure F.2: Rate of successful rejection for each problem set (cont'd)

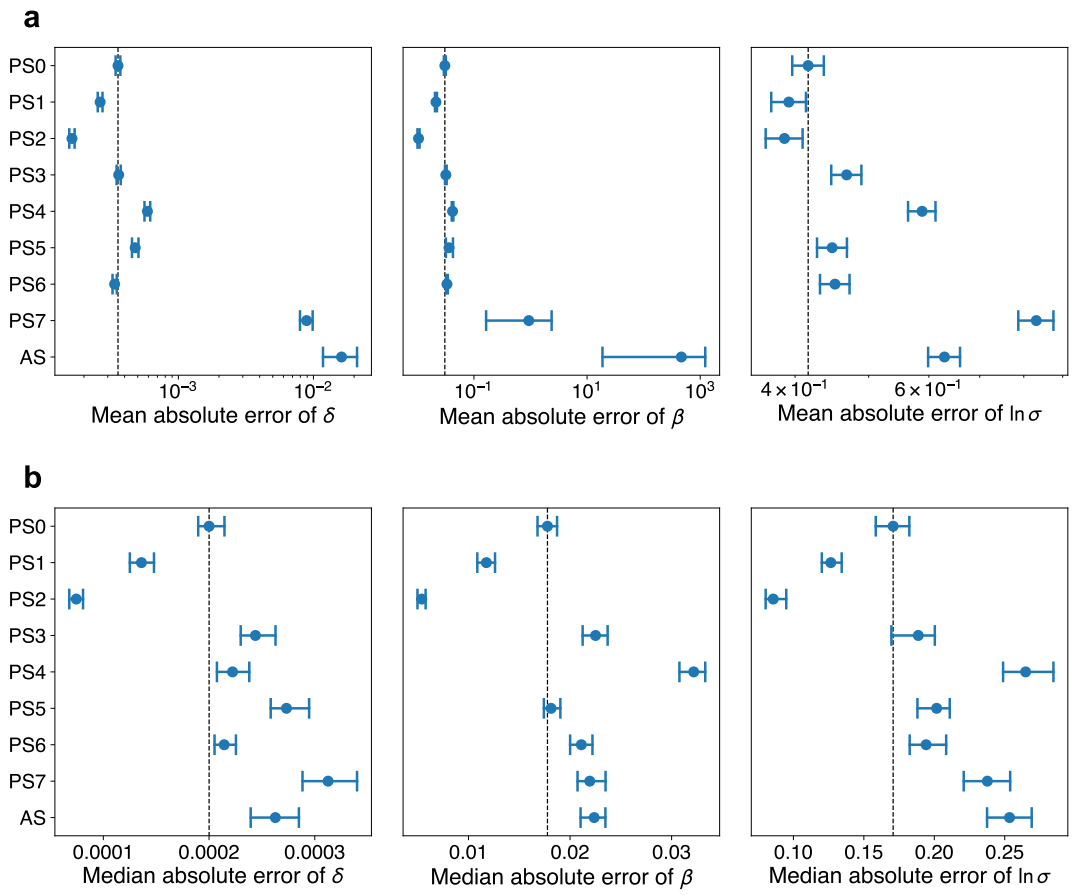


Figure F.3: Absolute errors of the estimates for each problem set

Notes: Represented by **a**) mean and **b**) median. Error bars represent the bootstrap 95% confidence intervals.

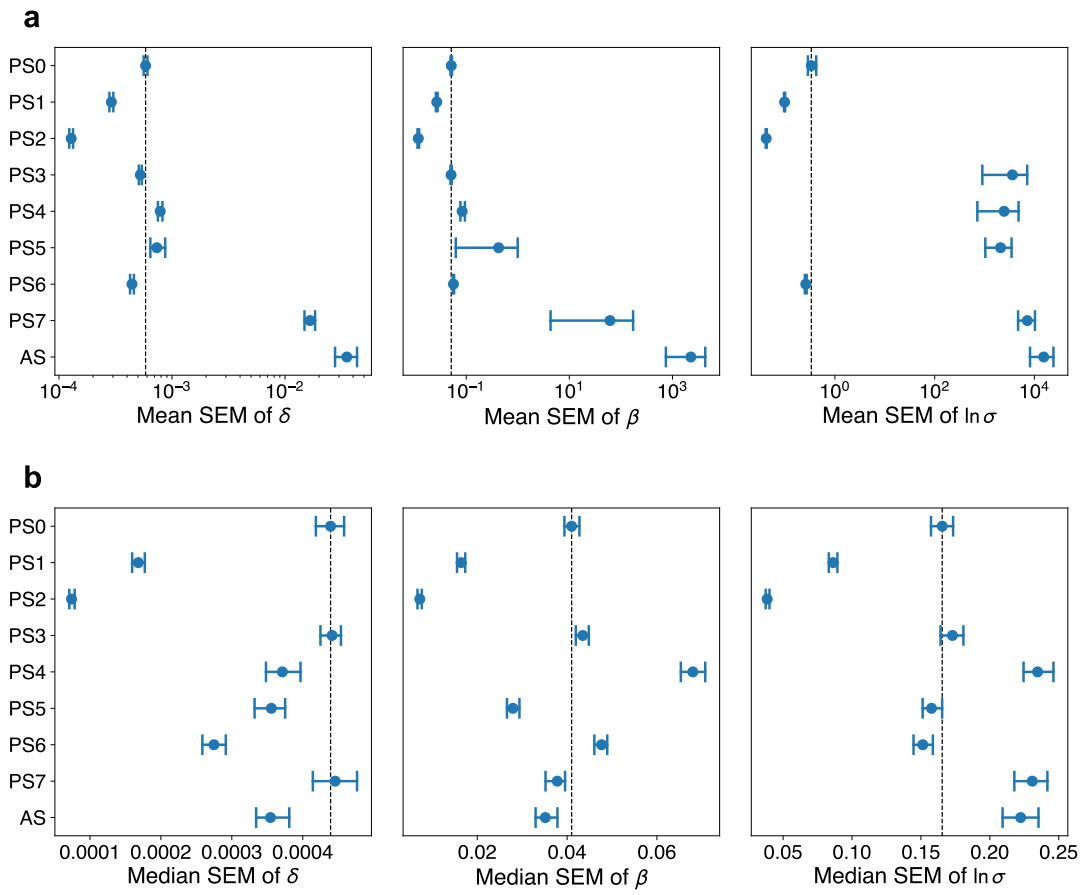


Figure F.4: Standard errors of the estimates for each problem set

Notes: Represented by **a**) mean and **b**) median. Error bars represent the bootstrap 95% confidence intervals.

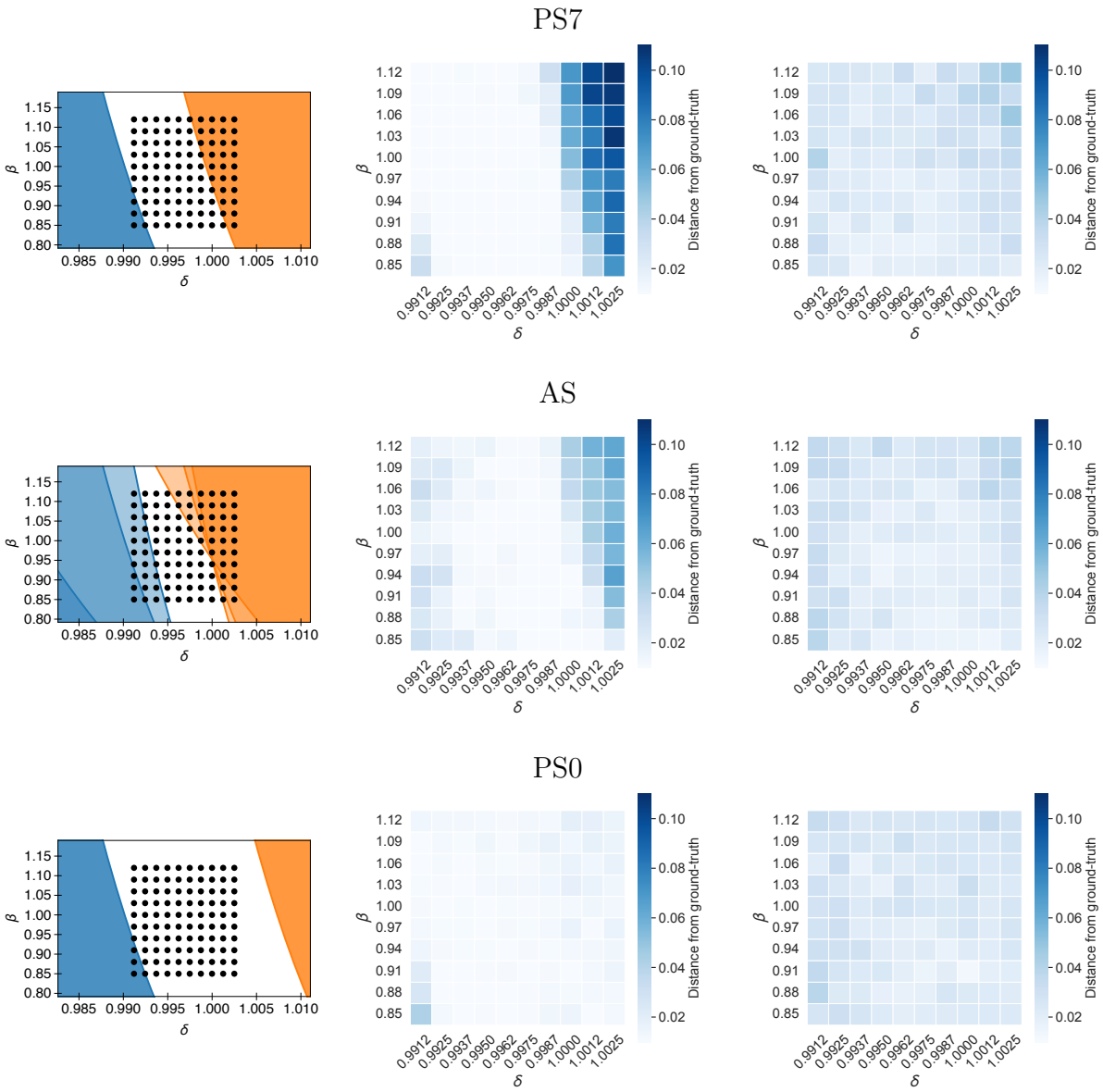


Figure F.5: Range of parameters for which estimation accuracy is inevitably low

Notes: **Left**: Region plots where the switching point is outside the range of prices of the problem set for individuals for whom the utility function is linear. **Center**: Heatmaps representing the median Euclidean distance between the ground-truth and the estimated values within each cell. **Right**: Heatmaps of synthetic individuals with relatively nonlinear utility ($\ln \sigma = 0.33, 1.11, 1.89$).

G Negative Ground-truth $\ln \sigma$

In the main analysis, the ground-truth values of the curvature parameter $\ln \sigma$ were restricted to positive values. However, in this section, we examine the accuracy and precision of parameter estimation for negative values of $\ln \sigma$, representing lower elasticities. Specifically, we used ground-truth values of $\ln \sigma$ at -2.00, -1.22, and -0.44 (-6.39, -2.39, and -0.56 in ρ), along with the same ground-truth values for δ and β as in the main analysis.

The lower panel of Figure G.1 presents box plots of the estimates obtained for negative ground-truth values of $\ln \sigma$. In panels **a** and **b**, each plot provides a summary of 10 replications of all combinations of ground-truth values of β (δ) and $\ln \sigma$, respectively, thus involving 300 simulation agents. In panel **c**, each plot provides a summary of 10 replications of all combinations of ground-truth values of δ and β, thus involving 1,000 simulation agents.

Upon comparing the upper and lower rows of panels **a** and **b**, it appears that the medians of the estimates of δ and β do not show any significant difference. However, for negative ground-truth values of $\ln \sigma$, we observe that the distribution range of estimates (length of boxes and whiskers) is considerably larger and hence, the estimation precision is comparatively low. The lower row panel **c** displays the estimates of $\ln \sigma$, where the boxes with ground-truth values of -2.00 and -1.22 overlap even at $s = 0.10$. Therefore, distinguishing between them for negative values of $\ln \sigma$ can be very challenging.

Figure G.2 depicts the success rate of null hypothesis rejection when $\ln \sigma$ values are negative. Each data point is obtained by conducting 10 replications of all combinations of ground-truth values of β (δ) and $\ln \sigma$, totaling 300 simulation agents. In comparison to the scenario where $\ln \sigma$ is positive (as illustrated in Figure 1), the rate of successful null hypothesis rejection is typically lower.

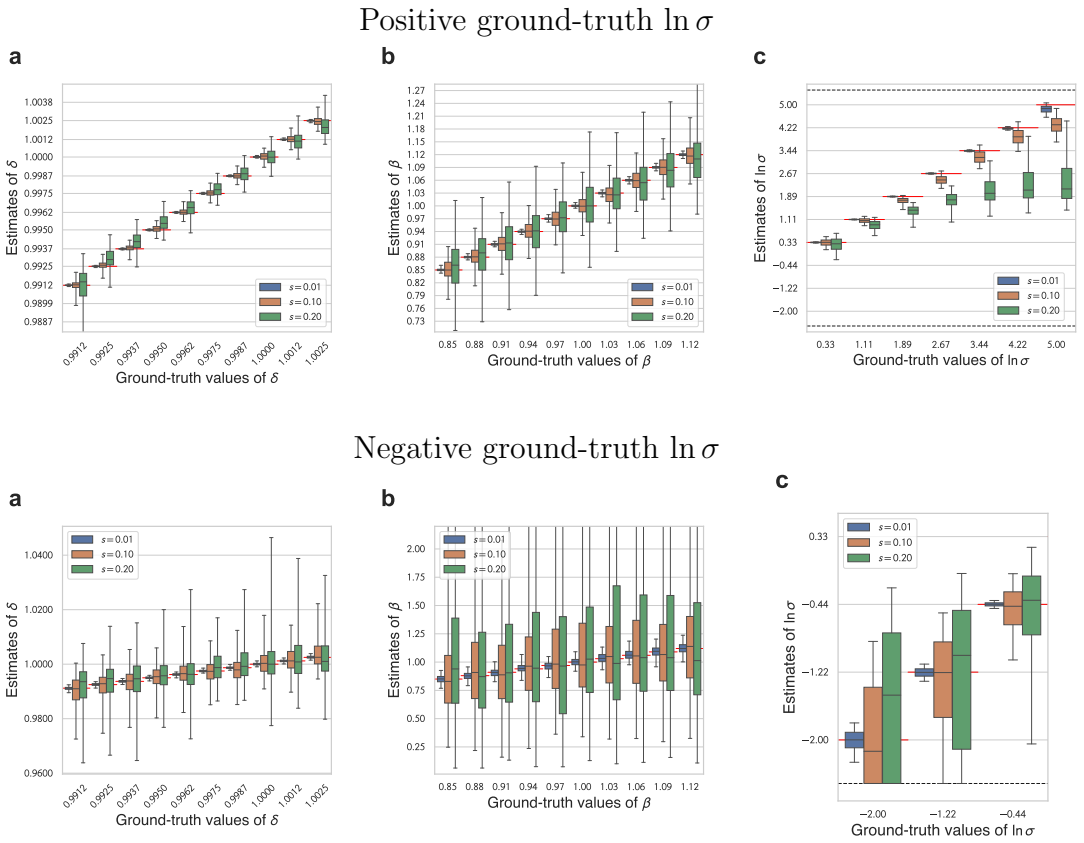


Figure G.1: Box plots of estimates for positive and negative ground-truth $\ln \sigma$

Notes: Estimates of **a**) δ, **b**) β, and **c**) $\ln \sigma$. For positive ground-truth $\ln \sigma$, panels **a** and **b** are reshown as Figure 2 and panel **c** is reshown as Figure A.1.

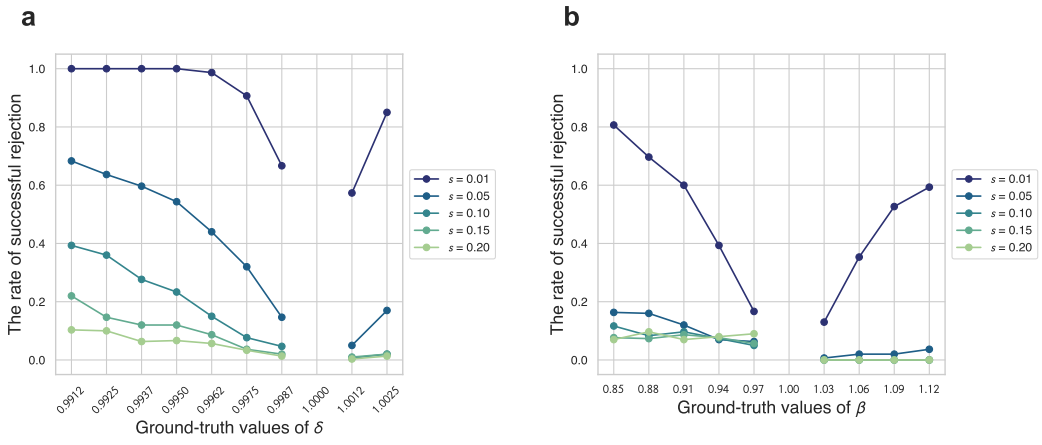


Figure G.2: Rate of successful rejection for negative ground-truth $\ln \sigma$

Notes: Tests on **a**) $\widehat{\delta} = 1$ and **b**) $\widehat{\beta} = 1$.

H Censored Noise

In the main analysis, the synthesized decision data were subjected to a truncated distribution of added noise. In this section, we investigate the estimation using censored-noised decision data.

Figure H.1's lower row displays box plots of the estimates for the censored-noised data. In general, there is little disparity in the accuracy and precision of the estimates between noise types. However, the estimation of $\ln \sigma$ at $s = 0.20$ exhibits a notable difference: the corrected-noised data do not saturate, as opposed to the truncated-noised data.

Figure H.2 displays the successful rejection rates for both censored noise (represented by solid lines) and truncated noise (represented by dashed lines). Although there is generally little difference in the success rates between the two noise types, censored noise appears to have a higher success rate when the noise level is large.

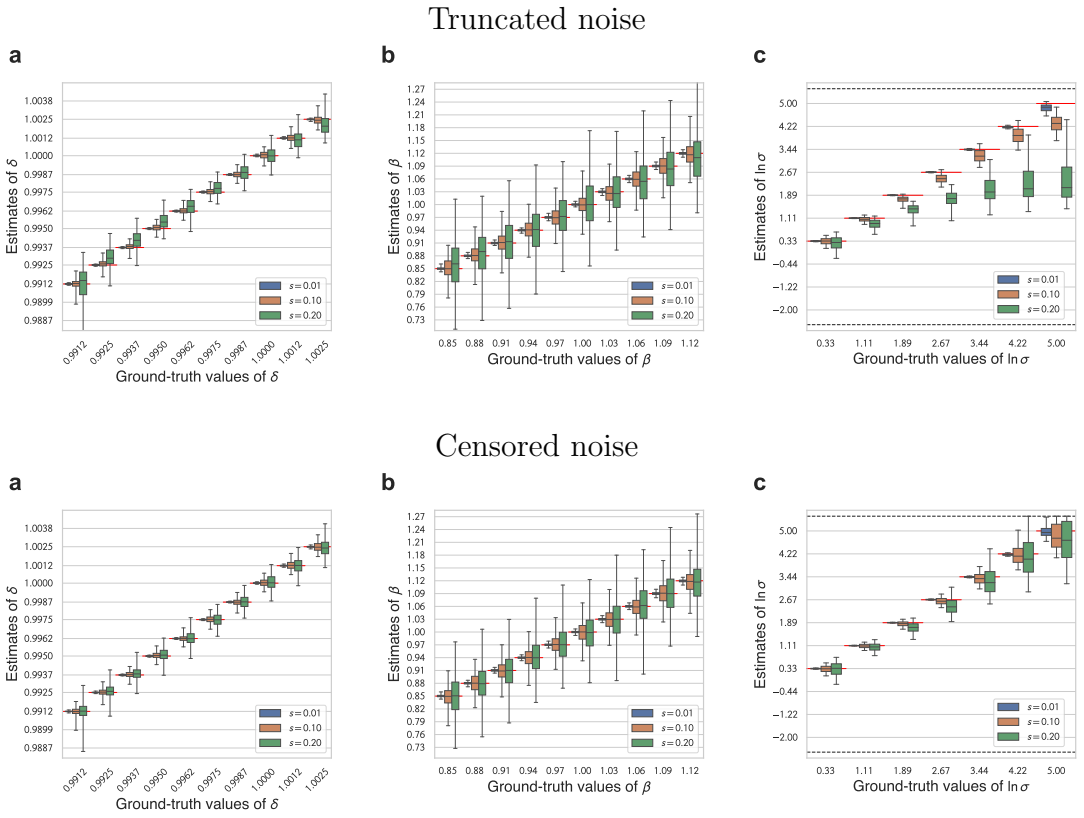


Figure H.1: Box plots of estimates for the truncated and censored noise

Notes: Estimates of **a**) δ, **b**) β, and **c**) $\ln\sigma$. For the truncated noise, panels **a** and **b** are reshown as Figure 2 and panel **c** is reshown as Figure A.1.

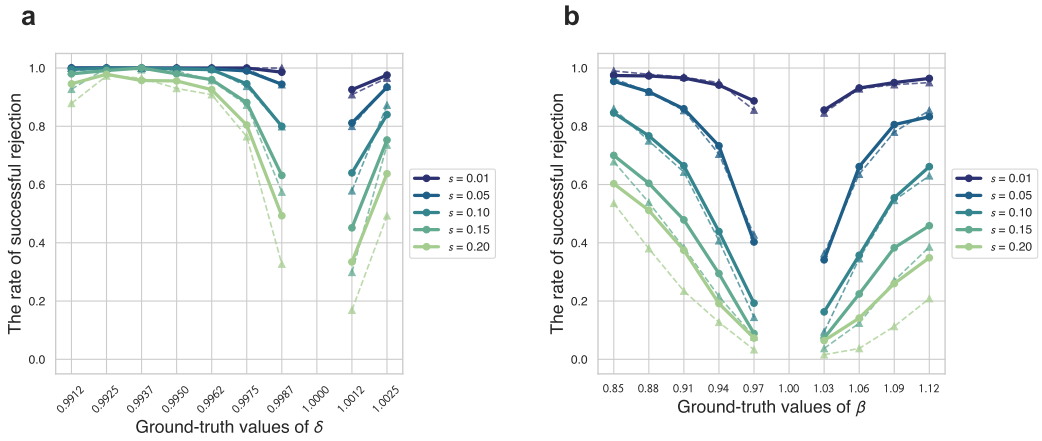


Figure H.2: Rate of successful rejection for the censored and truncated noise

Notes: Tests on **a**) $\widehat{\delta} = 1$ and **b**) $\widehat{\beta} = 1$. Solid lines are for the censored noise and dashed lines are for the truncated noise.

I Evaluation of Standard Errors by the Bootstrap Method

In the main analysis, we used the jackknife method to assess the standard errors of the estimates. In this section, we examine the success rate of the null hypothesis when standard errors are evaluated by the bootstrap method.

The upper row of Figure I.1 depicts the rate of successful rejection of the null hypothesis for the bootstrap method (represented by solid lines) and the jackknife method (represented by dashed lines). The difference in the success rate for δ is minimal across the evaluation methods. However, for β, the success rate tends to be lower with standard errors computed by the bootstrap method in comparison to the jackknife method.

A significant contrast is discernible when the problem set lacks a question concerning negative interest rates. The success rate is illustrated in the lower row of Figure I.1, employing problem set PS7 in Appendix F. In the case of δ's success rate, we note that for $s = 0.01$, there are instances where the success rate falls below 80%.

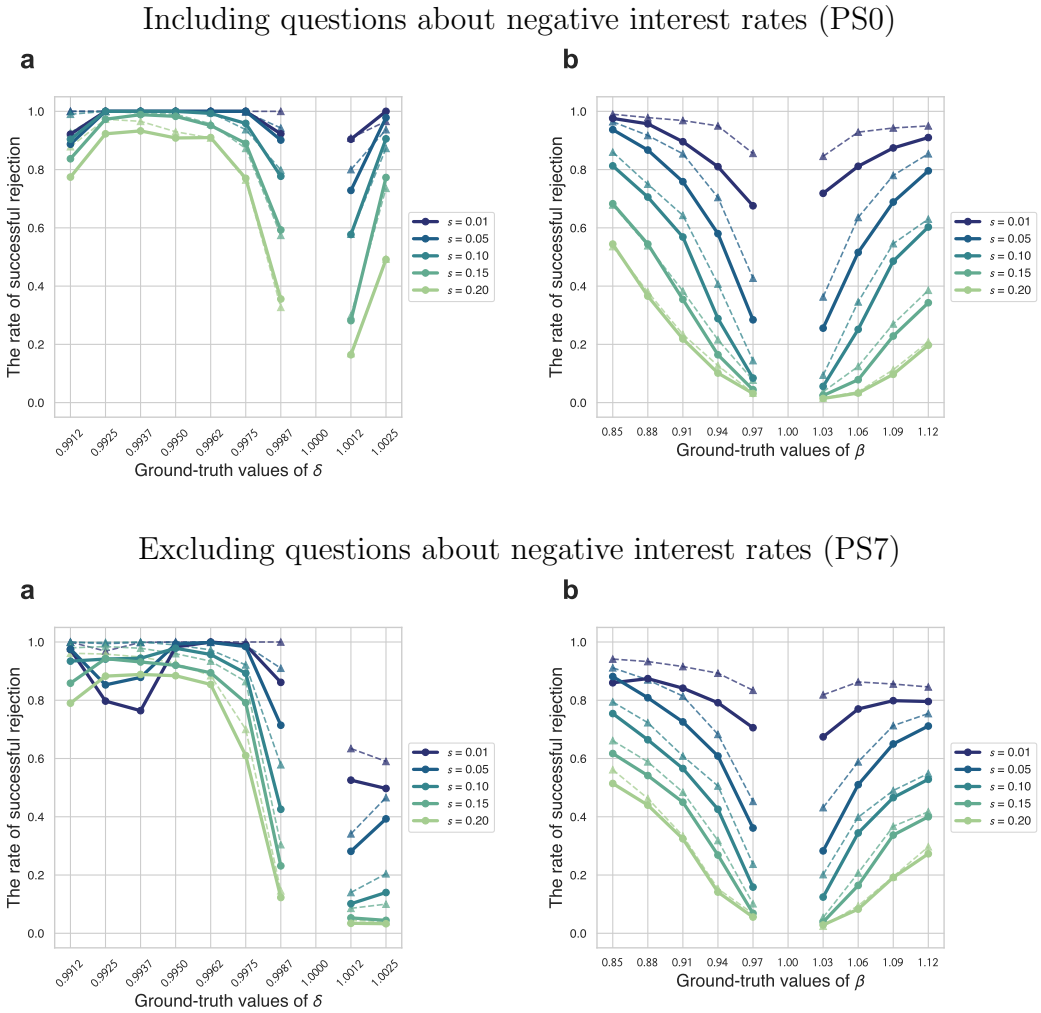


Figure I.1: Rate of successful rejection for the bootstrap and jackknife method

Notes: Tests on **a**) $\widehat{\delta} = 1$ and **b**) $\widehat{\beta} = 1$. Solid lines are for the bootstrap method and dashed lines are for the jackknife method.

J Demand Curve for Individuals Who Have an Extreme $\ln \sigma$

Figure J.1 illustrates the demand curves of individuals with extremely large or small values of the curvature parameter $\ln \sigma$. Both individuals with $\ln \sigma = 6$ and 7 allocate all of their resources to the sooner period when prices are lower than the switching point $(\beta^{\mathbf{1}_{t=0}} \delta^k)^{-1}$, and all to the later period otherwise. Their behavior is similar, with no discernible difference. For $\ln \sigma = 5$, there is a slight variation in behavior compared with $\ln \sigma = 6$, as they allocate a positive amount to the sooner period when the price is $1 + r = 1.44$, which is higher than the switching point. However, no other differences in behavior are observed. For $\ln \sigma = -2$, -3, and -4, there are some differences in the demand curves, but they are negligible.

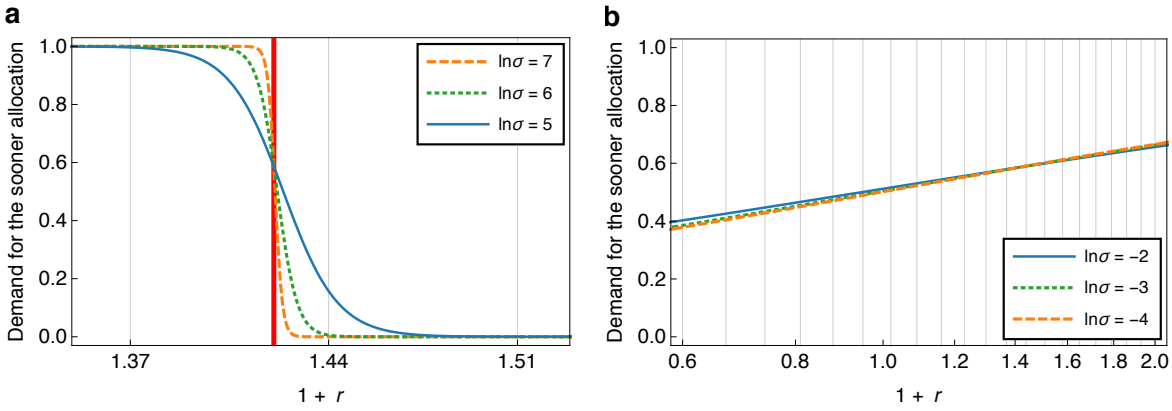


Figure J.1: Demand curves for extreme values of $\ln \sigma$

Notes: Each demand curve corresponds to an individual with different values of the curvature parameter: **a**) $\ln \sigma \geq 5$, and **b**) $\ln \sigma \leq -2$. Here, $\delta = 0.9950$ and $\beta = 1$. The horizontal axis, which represents the price $1 + r$, is presented on a logarithmic scale. Note that the individual faces the decision of allocating resources between the present time $(t = 0)$ and 70 days later $(k = 70)$, at prices indicated by the vertical lines in the figure. For panel **a**, only the neighborhood of the switching point, $(\beta^{\mathbf{1}_{t=0}} \delta^k)^{-1}$ (indicated as a red line), is displayed.

K Transformation of $\ln \sigma$

In the main analysis, we applied a sigmoid function $\ln \sigma = f(\theta) = 4 \tanh(\theta) + 1.5$ to transform $\ln \sigma$ and searched for the latent variable θ to avoid estimation failures. In this section, we examine the estimation results without the transformation $f(\theta)$.

The lower row of Figure K.1 presents box plots of the estimates without the transformation $f(\theta)$. We observe minimal differences in the estimation results with and without the transformation. Figure K.2 displays box plots for negative true $\ln \sigma$, as discussed in Appendix G. For negative ground-truth $\ln \sigma$, the whiskers are more extended than with the transformation, indicating that some estimates may be considered outliers.

Figure K.3 depicts the success rate of rejecting the null hypothesis with (represented by dashed lines) and without (represented by solid lines) the transformation $f(\theta)$. The difference in the success rate between the two cases is negligible.

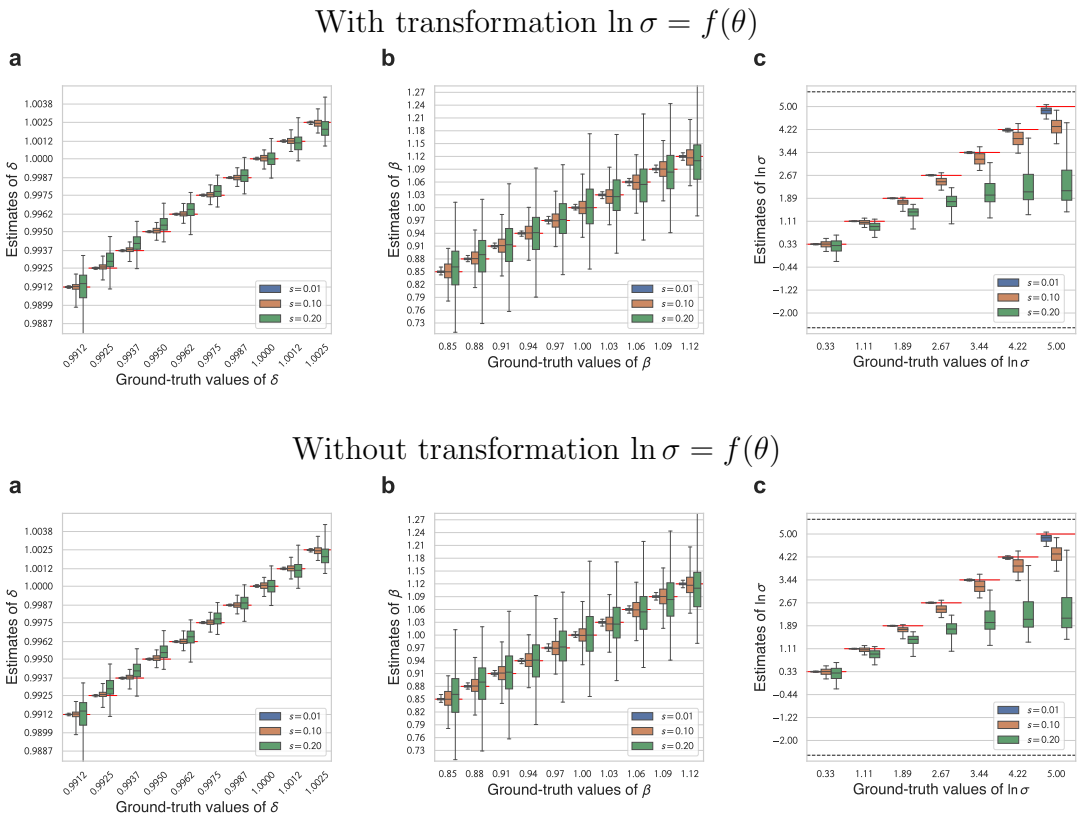


Figure K.1: Box plots of estimates with and without the transformation $f(\theta)$ for positive ground-truth $\ln\sigma$

Notes: Estimates of **a**) δ, **b**) β, and **c**) $\ln\sigma$. For the with-transformation case, panels **a** and **b** are reshown as Figure 2 and panel **c** is reshown as Figure A.1.

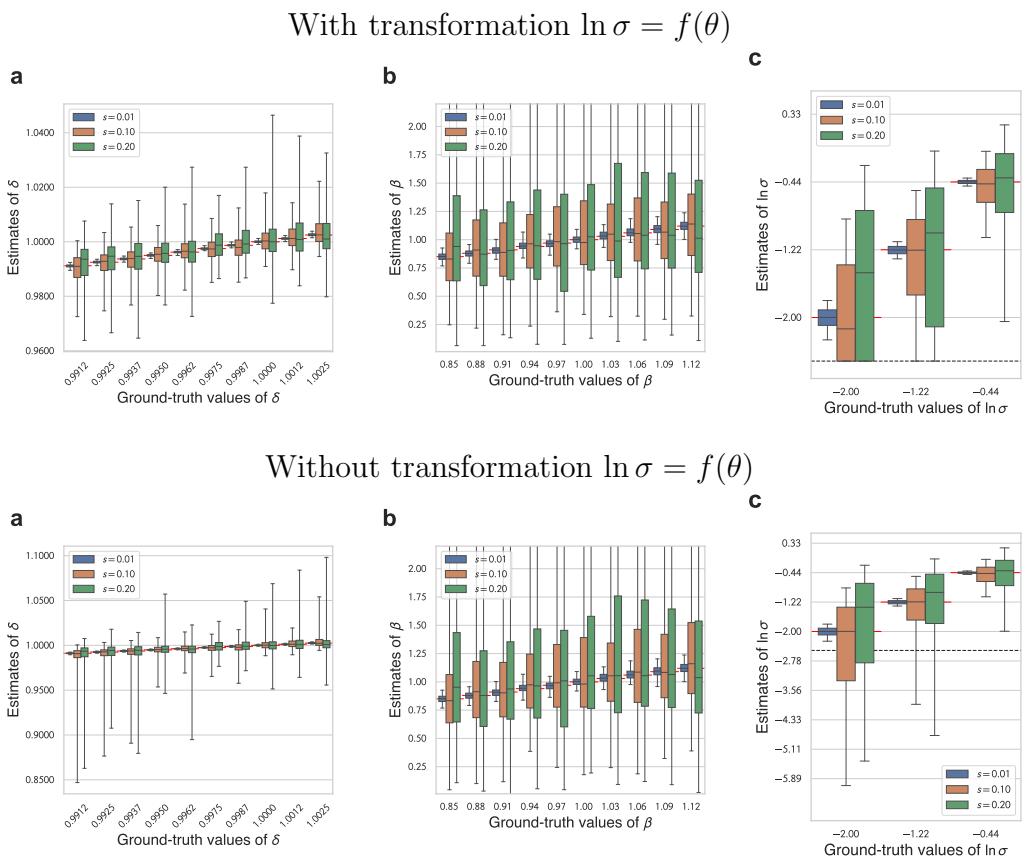


Figure K.2: Box plots of estimates with and without the transformation $f(\theta)$ for negative ground-truth $\ln\sigma$

Notes: Estimates of **a**) δ, **b**) β, and **c**) $\ln\sigma$. For the with-transformation case, panel **a**, **b**, and **c** are reshown as Figure G.1.

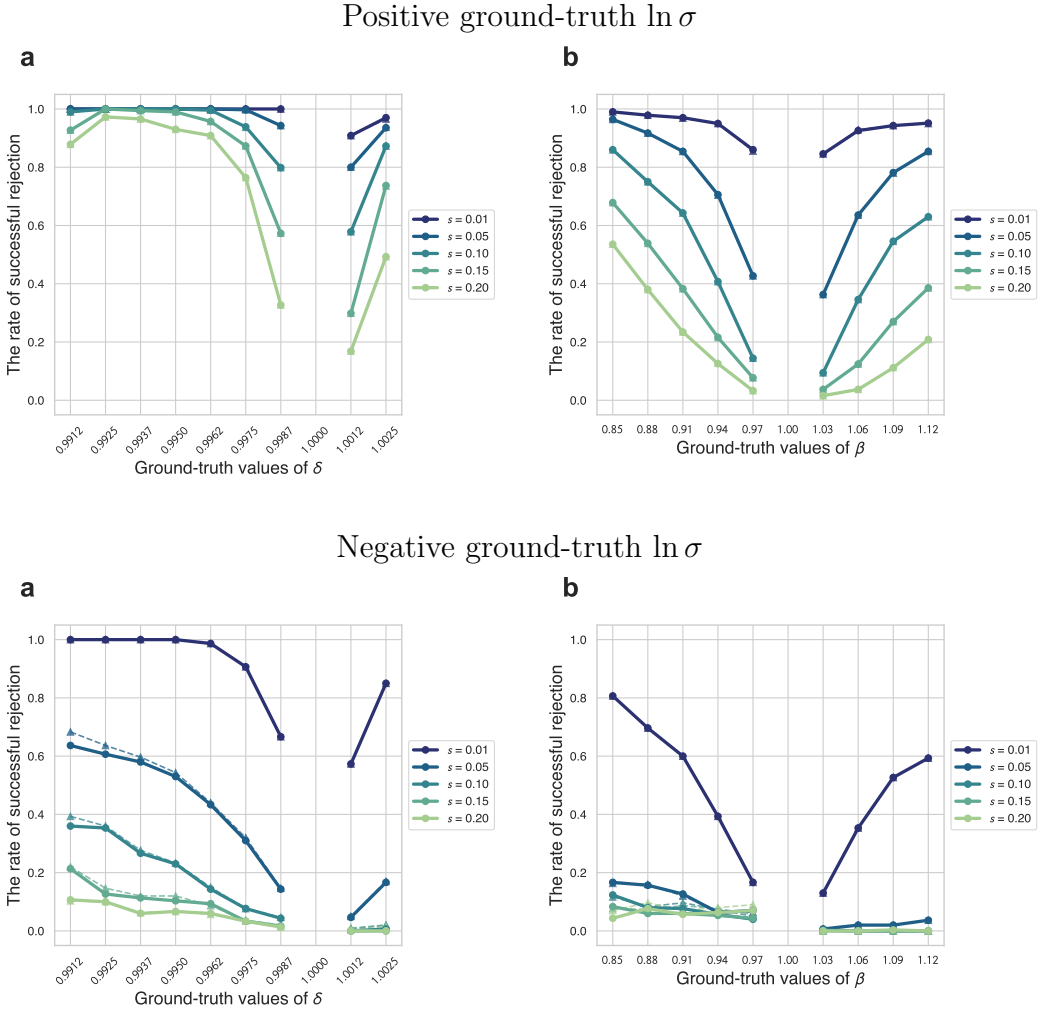


Figure K.3: Rate of successful rejection without and with the transformation

Notes: Tests were conducted on the estimated parameters of **a**) $\widehat{\delta} = 1$ and **b**) $\widehat{\beta} = 1$. Solid lines represent the results without the transformation, whereas dashed lines represent the results with the transformation.