Journal of Behavioral and Experimental Finance 43 (2024) 100962

Contents lists available at ScienceDirect

Journal of Behavioral and Experimental Finance

journal homepage: www.elsevier.com/locate/jbef

Full length article

Investigation of the convex time budget experiment by parameter recovery simulation☆

Keigo Inukai [a], Yuta Shimodaira [b,*], Kohei Shiozawa [c]

[a] Department of Economics, Meiji Gakuin University, Japan
[b] Institute of Social and Economic Research, Osaka University, Japan
[c] Department of Economics, Takasaki City University of Economics, Japan

ARTICLE INFO

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

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 discounted 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 range 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 discounted utility model to explore the effect of present bias.

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 prioritize immediate gains over larger future rewards. The discount rate is a key factor in determining the degree to which future payoffs are discounted over time in intertemporal choice problems.

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 decision 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. Various studies have investigated the associations between time inconsistencies in individuals' time preferences and their behavior, such as impulsive discounting among individuals with substance abuse disorders (Kirby et al., 1999), smokers (Bickel et al., 1999), and gamblers (Holt et al., 2003). In addition, several studies have examined consumer financial behavior across domains, such as credit card borrowing (Meier and Sprenger, 2010), creditworthiness (Meier and Sprenger, 2012), and mortgage choices (Atlas et al., 2017).

The convex time budget (CTB) method, developed by Andreoni and Sprenger (2012a, henceforth AS), is one of the main methods used to measure time preferences and has rapidly become a highly influential approach over the past decade.[1] The CTB method attempts to simultaneously elicit the effects of time discounting and the curvature of the utility function with a single instrument. When analyzing preferences from behavioral data collected via the CTB method, researchers do

☆ 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. We gratefully acknowledge the financial support from a grant-in-aid for scientific research (KAKENHI; grant numbers: 17H04780, 18K19954, 19J12097, 19K21701, 19K23195, 20H05631, and 21H05070) from the Japan Society for the Promotion of Science. OnLine English (www.oleng.com.au) edited the manuscript to improve its readability.

* Corresponding author.
E-mail addresses: inukai@eco.meijigakuin.ac.jp (K. Inukai), shimodaira@iser.osaka-u.ac.jp (Y. Shimodaira), shiozawa@tcue.ac.jp (K. Shiozawa).

[1] Another well-known method is the double multiple price list (DMPL), proposed by Andersen et al. (2008), which simultaneously measures the curvature of utility functions and time discounting. While the DMPL method measures an individual's discount rate by temporarily assuming linear utility and then adjusts using the results from measuring risk attitudes, the CTB method uniquely estimates curvature by allowing the choice of interior points on the budget constraint line. For a general overview of time preference elicitation methods, see Cheung (2016) and Cohen et al. (2020).

https://doi.org/10.1016/j.jbef.2024.100962
Received 5 April 2023; Received in revised form 4 June 2024; Accepted 22 July 2024
Available online 26 July 2024

not merely compare the intertemporal allocations across conditions but also estimate the parameters of the quasi-hyperbolic discounted (QHD) utility function (Laibson, 1997; O'Donoghue and Rabin, 1999). 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).

Following AS, one of the main applications of the CTB method is to estimate the present bias parameter, β, of the quasi-hyperbolic discounting model, wherein the parameter β indicates that the discount factor differs by a factor of β depending on whether the earlier period is the present in intertemporal decision-making. Imai et al. (2020) have identified 67 articles that employed the CTB method and presented a meta-analysis of these studies. This meta-analysis showed that, on average, the estimated value of the present bias parameter β ranges from 0.95 to 0.97, indicating that there is very limited evidence of time inconsistency due to present bias.[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 unbiasedness 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 $\beta = 0.97$, which is representative of the value reported in the literature, can we truly claim that there exists a time inconsistency in the individual's behavior?

The present paper studies the properties of the econometric procedures typically applied to CTB data in cases where the true parameter values are known, to examine the unbiasedness and precision with which those true values are able to be recovered. 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 software designed for real data; and finally, comparing the estimated parameters with the ground-truth values to assess the level of accuracy in their recovery.

In particular, we systematically vary the values of the discount factor δ, the present bias parameter β, and utility curvature in the neighborhood of the values that have been reported in the literature, and simulate the choices of agents who act upon such preferences with varying degrees of noise. We apply standard CTB estimation procedures to these data to examine how close the resulting estimates are to the true values used to generate the data.

The results indicate that the CTB design is well-powered to reject the null hypothesis that $\delta = 1$ even when the true value of δ is very close to one. However, the same cannot be said for the key present bias parameter β. For $\beta = 0.97$ with about 5% behavioral noise, the null hypothesis is successfully rejected only around half the time even though it is false. As a result, it is possible to discriminate rather small differences in δ, but not in β.

The CTB estimates of β and δ are strongly negatively correlated, indicating that the estimator struggles to discriminate between the two parameters. This can be explained by the fact that even quite sizeable changes in β are predicted to have much smaller effects upon demand behavior than rather small differences in δ.

Moreover, the estimates of utility curvature are biased in the direction of greater concavity than in the underlying data generating process; this is notable given that CTB studies typically find rather little concavity in the first place.

According to Imai et al.'s (2020) meta-analysis, there may be a tendency for selective reporting of present bias parameter β estimates of less than one, particularly in studies using real effort tasks. In addition, our investigation has revealed imprecision in estimating the present bias parameter β, which can exacerbate the problem of selective reporting of the parameter estimate by reducing the power of a statistical test based 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 QHD utility model — when the true extent of present bias is only modest.

In psychology, the replicability of experimental findings can often be problematic, and in experimental economics, it is a crucial issue 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 the 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 presents 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 unbiasedness and precision of the parameter estimates from the distribution of the estimates. We then demonstrate that combining the CTB method with the quasi-hyperbolic discounting model does not yield high-precision estimates of the present bias parameter. Furthermore, this approach fails to detect time inconsistency when the true value of the present bias parameter β is close to one. In the latter part of Section 3, we discuss the reasons for the low precision in estimating the present bias parameter and explore the effectiveness — or lack thereof — of several potential improvements. 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.[3]

2.1. Behavioral model

We now consider the decision 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 period k days later. It only matters whether the front-end delay, t, is 0 (*i.e.*, present) or not; 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

[2] Cheung et al. (2021), while performing a meta-analysis on the present bias parameter β 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 DMPL method.

[3] The codes for generating synthetic decision data and parameter estimation are available at https://osf.io/j93bx/.

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 discounted (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, $\sigma = (1-\rho)^{-1}$.[4]

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

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

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, on average, have 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 choices 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.

The value of the demand function after adding noise should remain within the interval [0, 1]. Specifically, we draw a random number from the distribution $\mathcal{N}\left(g(\bullet), s\right)$ and accept it as a synthesized decision if it is in $[0, 1]$; otherwise, we repeatedly draw a random number until it does. In other words, the noise associated with the decision follows a truncated distribution. This is because 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.[5]

2.2. Experimental tasks

We have two situations — defined as a combination of the front-end delay t and the delay length k — for the virtual 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 typically ranges from weeks to months and is seldom shorter than one week (Imai et al., 2020). In each situation, we set 21 uniformly spaced prices chosen from $0.6 \le 1+r \le 2$. We fixed income m at 20 for simplicity, as 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 that of AS. The first difference lies in our choice of the price $1+r$ from a range where the interest rate r can be both positive and negative. Few studies employing CTB experiments, including the one by AS, inquire about negative interest rates. However, without addressing negative interest rates, it is impossible to estimate the discount factor for an individual who does not merely discount future payoffs but actually places a premium on them. For such an individual, the discount factor δ will be greater than 1.[6]

Second, we have chosen to set the delay k to a single value, $k = 70$, in this paper. This setting diverges from that of AS, wherein k was varied across three distinct values: 35, 70, and 98. Prior to this study, we created various problem sets by altering several elements based on the AS problem set. We then conducted parameter recovery simulations to compare the recovery performance. As a result, we found that the estimation accuracy was relatively high when limited to a single case of k. Therefore, we will evaluate the estimation accuracy using this setting, which, for the time being, we believe yields the best possible estimation accuracy.[7]

[4] 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 DMPL task — 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 (Andreoni and Sprenger, 2012b; Abdellaoui et al., 2013; Harrison et al., 2013; Andersen et al., 2014; Andreoni and Sprenger, 2015; Cheung, 2015; Miao and Zhong, 2015; Andersen et al., 2018; Cheung, 2020). 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.

[5] The truncated-noise 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 $\mathcal{N}\left(g(\bullet), s\right)$ falls outside of $[0, 1]$. For a discussion on the assumptions of noise and estimation methods, see Section 3.5 and Appendix A.

[6] The impact of excluding problems associated with negative interest rates from the problem set on parameter estimation is discussed in Appendix B. It was found that using the AS problem set, the accuracy of δ estimation significantly declines when the true δ is greater than 1.

[7] Assuming linear utility, individuals shift their consumption from period t to $t+k$ once the price $1+r$ exceeds the threshold; the switching point is determined by $1+r = (\beta^{\mathbf{1}_{t=0}}\delta^k)^{-1}$. The conventional multiple price list method (Coller and Williams, 1999; Harrison et al., 2002) measures the discount factor using this concept, suggesting that varying the price with a fixed delay k is a natural setting for many researchers. Alternatively, keeping $1+r$ constant and slightly adjusting k might equally elicit δ based on the equation $k = -\ln(1+r)/\ln\delta$ (here, the present bias β is ignored). Note that k must be nonnegative because it is impossible to travel into the past and receive a reward after making decisions. For individuals with a true δ exceeding 1, instead of setting k to be negative, the problem set should include cases where $1+r < 1$ for practical δ estimation. Given the concave utility, it is not immediately apparent whether adjustments to price, delay, or employing a hybrid approach, such as AS's design, would be most effective. In addition,

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 values are listed.

Third, we reduced the variation of the front-end delay, t, to two states: present or not present. In AS's experiment, participants made decisions regarding the allocation between the near future and a more distant future for $t = 7$ and $t = 35$ separately. According to the CES-QHD utility function model in Eq. (2), there should be no difference in decisions between $t = 7$ and $t = 35$; however, this variance can affect real behavior. For an actual experimental design, it may be beneficial to specify the variation in t to address behavioral bias, 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 used $\ln\sigma = \ln(1/1-\rho)$ 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, we generated data for 700 synthetic individuals. As mentioned above, there are five levels of noise, denoted by 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. Typically, 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,[8] we also included these values in our set of potential 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 Cobb–Douglas utility curvature) to 5 ($\rho = 0.993$; nearly linear curvature). Recall that the domain of $\ln\sigma$ encompasses all real numbers. Let us assume that $\ln\sigma = 0$ (or $\rho = 0$), which corresponds to the Cobb–Douglas utility function, is at the center of the curvature parameter space. For $\ln\sigma > 0$, the intertemporal allocations become substitutive, and for $\ln\sigma < 0$, they become complementary. As $\ln\sigma \longrightarrow -\infty$ (or $\rho \longrightarrow -\infty$), the utility function approaches 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 approaches a linear function: $U = c_t + \beta^{\mathbf{1}_{t=0}}\delta^k c_{t+k}$, which represents 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 in relatively complementary ways, 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. 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 the CTB method is used (Andersen et al., 2008; Andreoni and Sprenger, 2012a; Andreoni et al., 2015; Cheung, 2020). We also examined the estimation errors for cases where $\ln\sigma < 0$ and have included the results in Appendix C. Our simulations reveal that when the ground-truth value of $\ln\sigma$ is negative, the accuracy of estimating δ and β significantly decreases.

2.4. Estimation methods

As 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 $\tilde{c}_i^j = g(t_j, k_j, 1+r_j, m_j \mid \delta_i, \beta_i, \ln\sigma_i) + \epsilon$. Given the generated data, we estimate the three parameters using a nonlinear least squares method.[9] Following AS, we utilized the "`nl`" command in Stata. Mathematically, the values of $\hat{\delta}$, $\hat{\beta}$, and $\widehat{\ln\sigma}$ minimize the sum of squared residuals:

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

To prevent estimation failures due to nonconvergence of the calculations, we transformed $\ln\sigma$ using a sigmoid function f as $\ln\sigma = f(\theta) = 4\tanh(\theta) + 1.5$, and estimated the latent variable θ.[10]

the impact of the number of conditions on k has been thoroughly examined, as detailed in Section 3.5. Our simulations indicate that no method stands out as unequivocally superior.

[8] 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.

[9] In the AS model, the error term was assumed to follow a censored normal distribution, and a two-limit Tobit model was utilized for estimation. However, the two-limit Tobit model may lead to 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 at the corner ($\tilde{c} = 1$) rather than in a position closer to the theoretical decision. For this reason, we specify an error term with a truncated normal distribution instead of a censored distribution.

[10] For $\ln\sigma > 5.5$, we cannot observe differences in the decision data with practical significance 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 indicates perfect substitutes, and $\ln\sigma$ less than -2.5 indicates perfect complements. This is because changes in $\ln\sigma$ no longer affect behavior $g(\bullet)$ beyond these points, as seen in the demand curves in Appendix C. At these extreme values of $\ln\sigma$, we sometimes face issues with parameter estimations not converging properly. To solve this, we use the S-shaped function $f(\theta)$. The effectiveness of this function was tested. For positive $\ln\sigma$, this transformation had a minimal effect: out of 35,000 agents, failures occurred for 15 agents (0.004%) without the transformation and for 17 agents (0.005%) with it. For negative $\ln\sigma$ (-2.00, -1.22, and -0.44), however, errors occurred for 367 (2%) of the 15,000 agents without the transformation, but we saw no failures when the transformation was applied. For details on how estimates change with and without this transformation, see Appendix C.

For the parameter estimation, we set the convergence criterion to 10^{-5} and the maximum number of iterations at 200. By combining all parameters $(\delta_i, \beta_i, \ln \sigma_i)$, and s, we created a pool of 3,500 synthetic individuals. For each synthetic individual, we regenerated the decision data 10 times. Of the 3,500 synthetic individuals, 3,483 converged all 10 times. The remaining 17 individuals experienced a single failure to converge.

2.5. Performance evaluations

To evaluate the performance of our parameter estimation methods, we employ several quantitative metrics, ensuring comprehensive insight into their unbiasedness and precision. These performance measures are described in detail below.

Rate of rejection of null hypothesis. In Section 3.1, 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. The value 1 indicates the individual does not discount (nor places a premium on) future payoffs. The rate is computed using two-tailed Student's t-tests at a 5% significance level and reflects the estimator's ability to detect deviation from the null hypothesis that $\widehat{\delta} = 1$ and $\widehat{\beta} = 1$.[11] Using this measure, we can discuss how far the true parameter values must differ from 1 to detect the effects of discounting behavior and present bias through parameter estimation based on data obtained from an experiment in a given setting.

Box plot analysis. In Section 3.2, we plot the boxplot representing the actual distribution of the estimates in a population with the same true parameter value to examine the errors of the parameter estimates further. Here, we assume a population in which the three parameters — δ, β, and $\ln \sigma$ — are distributed on a three-dimensional grid according to the ground-truth values we have set. Then, we examine the distribution of estimates of each parameter within this population. Note that the evaluation based on the rejection rate of the null hypothesis was conducted using the standard error of each estimate. By contrast, the boxplot summarizes estimates from various cases.

By displaying the distribution of parameter estimates using box plots, we can simultaneously evaluate how much the estimates deviate from the true values (unbiasedness) and the uncertainty of the estimates (precision). In addition, by displaying box plots for each ground-truth value, we can visually verify how much the true values influence the estimates.

To evaluate the unbiasedness of the estimates, we compare the median of the estimates — represented by the line in the center of the box — with the true value. Estimates occasionally include outliers. Excluding outliers is necessary to calculate the mean of the estimates, but it is challenging to establish a clear criterion for this exclusion. Instead of excluding outliers, we use the median, which is less affected by them. We evaluate bias with positions above or below indicating over- or underestimation.

To evaluate the precision of the estimates, we can use the length of the box in a box plot, namely the interquartile range, or the length of the whiskers, which is the range from the 5th percentile to the 95th percentile.

Now, let us consider the box plots of the estimated values for two ground-truths of δ, δ_0, and δ_1 (where $\delta_0 < \delta_1$). For example, imagine a scenario where the upper end of the whisker of the box plot for δ_0 exactly matches the lower end for δ_1. In this case, when attempting to classify an individual, whose true δ is δ_0, as corresponding to either δ_0 or δ_1, there is a 5% probability of mistakenly classifying it as $\delta = \delta_1$. If the whiskers of the box plots overlap, the error rate will be higher than 5%, and if the upper and lower ends of the boxes coincide, the error rate is 25%.

The importance of precision can be understood by analogy to the resolution of a scale used to measure weight. For example, a (cheap) bathroom scale measures in units of hundredths of a gram, while a pharmaceutical scale can measure up to one ten-thousandth of a gram with precision. Measuring instruments do not always need to have high resolution; the precision of a pharmaceutical scale is unnecessary for measuring body weight, and conversely, a bathroom scale is completely unsuitable for measuring the weight of the medication prescribed to a patient. Most scales have their resolution clearly indicated. Researchers must understand the resolution of the measuring instrument they intend to use in advance.

In our context, the resolution is defined as the minimum distance between true parameter values. If the estimation is obtained using a higher resolution, it becomes possible to precisely distinguish between any two individuals, even if the actual parameter values are very close. Researchers can assess identifiability by comparing the lengths of the boxes or whiskers. When evaluating the precision of parameter estimation, whether to base the discussion on the length of the box or the length of the whiskers in a box plot depends on the specific requirements of the research.

Supplemental analyses. To supplement the analysis using the rejection rate of the null hypothesis and box plots of the estimates, several analyses are conducted.

To investigate the relationships between the estimated values of the three parameters in the utility function, Section 3.3 presents scatter plots of the estimated curvature parameter $\ln \sigma$ versus the signed estimation errors for δ and β. Section 3.4 shows scatter plots of the estimated values of δ versus β.

Furthermore, in Section 3.5, we attempt to modify the experimental settings or estimation methods to improve the estimated performance. To compare the results of each modification attempt with the main simulation results, we use the median of the absolute estimation errors for each of the three parameters (and the 95% confidence interval of the median obtained through bootstrap simulations) as the evaluation measure. In the Appendix, we present the rejection rates of the null hypothesis and the box plots of the estimated values for each simulation.

3. Results

This section is structured into five subsections, each addressing distinct aspects of the performance of parameter estimation.

First, the detectability of time discounting parameters, δ and β, is evaluated by the rate of rejection of the null hypothesis, revealing that δ can be reliably distinguished from 1 for $\delta < 1$, whereas β estimates struggle to differentiate from 1 under noisy conditions.

Second, we evaluated the unbiasedness and precision of the estimates by displaying box plots of the estimated values for δ and β. The box plots show that δ and β estimates tend to shift toward 1. Generally, δ exhibits higher precision and distinguishability between adjacent values than β.

Third, analysis of the curvature parameter $\ln \sigma$ indicates a significant bias in its estimates, notably when substantial noise is involved, impacting the precision of δ and β estimates, but not leading to systematic over- or underestimation.

Fourth, the intrinsic difficulty in uniquely identifying β due to minimal theoretical behavioral differences between closely spaced true values is dissected, highlighting the challenge in accurate present-bias estimation.

Finally, various strategies to enhance parameter estimation accuracy are assessed, finding that expanded problem sets generally improve precision, but merely altering delay periods or price ratios does not yield significant accuracy gains.

[11] 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 D). Therefore, we chose the jackknife method to avoid underestimating the precision, *i.e.*, to be conservative about what we are trying to conclude.

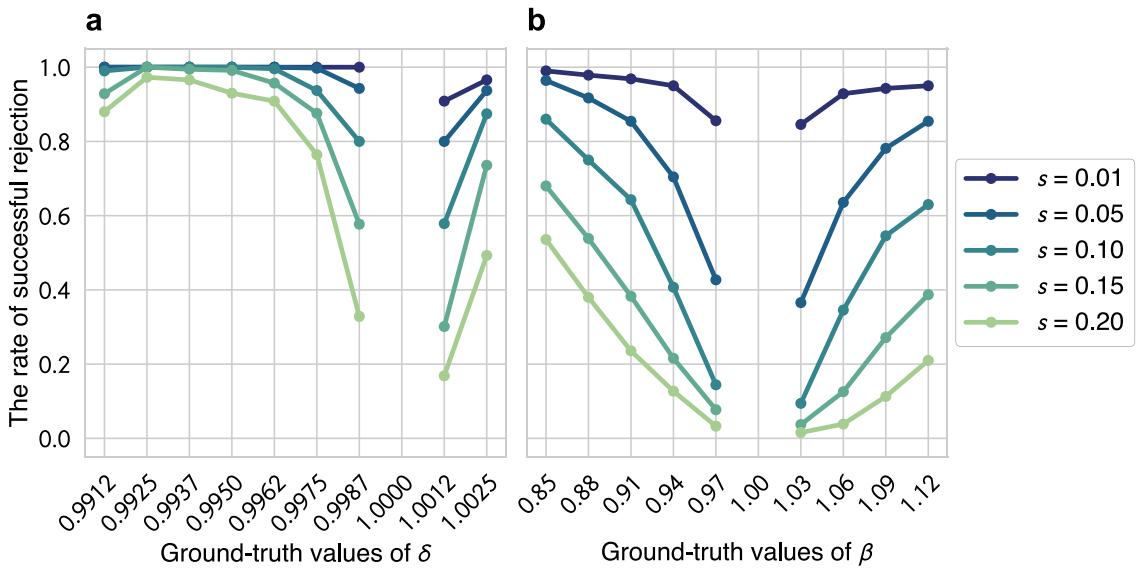


Fig. 1. Rate of successful rejection of the null hypothesis.
Notes: Tests on (**a**) $\hat{\delta} = 1$ and (**b**) $\hat{\beta} = 1$. Each point summarizes 10 replications of all combinations of the ground-truth values of β (or δ) and $\ln \sigma$, *i.e.*, 700 simulation agents.

3.1. Detectability of time discounting

Fig. 1 shows the rate of successfully rejected null hypotheses, such that $\hat{\delta} = 1$ and $\hat{\beta} = 1$. Each point summarizes the results of 10 replications of all combinations of the ground-truth values of β (or δ) and $\ln \sigma$, *i.e.*, 700 simulation agents.

An inspection of Fig. 1 reveals that for the discount factor parameter δ, when the ground-truth value is less than 0.9962, we can reject the null hypothesis $\hat{\delta} = 1$ 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 with the case of $\delta < 1$. As previously discussed, estimating an individual's δ when the true value exceeds one necessitates collecting data related to negative interest rates. Although tasks concerning negative interest rates were indeed included in our problem set, their quantity was smaller compared with tasks related to positive interest rates. The low success rate in cases where the true δ exceeds one may be attributed to this scarcity of tasks.

By contrast, regarding the present/future bias parameter β, we encountered more difficulty in concluding that the estimates are not equal to 1, compared with the case of the discount factor parameter δ, in general. When $s = 0.05$, to reject the null hypothesis $\hat{\beta} = 1$ with a success rate of over 90%, the true β must be less than 0.91 or greater than 1.12. For $s > 0.05$, even when the true β is as small as 0.85, the success rate falls below 90%.

Let us recall the meta-analysis by Imai et al. (2020), which found that the average estimated values of individual β ranged from 0.95 to 0.97. Our results show that when attempting to determine the presence of behavioral bias from parameter estimation for individuals with a true β value of 0.97, unless dealing with individuals who are not perturbed by noise (*i.e.*, with $s = 0.01$, meaning individuals who can almost accurately respond to the value of the demand function), the correct detection rate falls below 50%. This suggests that detecting time inconsistency from parameter estimates obtained through CTB experiments for individuals whose true value of β is 0.97 is challenging.

3.2. Unbiasedness and precision

Fig. 2 shows the distribution of the estimated values of δ and β as a box plot (refer to the subsequent subsection for the $\ln \sigma$ estimates). Each box summarizes the outcomes of 10 replications across all combinations of ground-truth values of β (or δ) and $\ln \sigma$, *i.e.*, 700 simulation agents.

When examining the median values of the estimates, it is evident that both δ and β tend to shift closer to 1 than do the ground-truth values. This observation suggests that the estimates for δ and β often underestimate the effect of discounts (or premiums). In most cases, we find that deviations from the true value fall within the interquartile range of the estimates' distribution. However, when $s = 0.20$, there are some instances where the true values lie outside the interquartile range of the δ estimates.

Next, we evaluate the precision of the estimation. For the discount factor parameter δ, Fig. 2 shows that the whiskers of the estimates for any two adjacent ground-truths do not overlap and can be distinguished from each other at the smallest noise level $s = 0.01$. Even with the most substantial noise, $s = 0.20$, the boxes do not overlap, although the whiskers do. We conclude that the experimental tasks considered in our simulations have enough precision 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}), we can distinguish between them, even assuming relatively large amounts of noise.

In contrast to the case of δ, Fig. 2 reveals that the precision 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 barely be 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 solely on statistically significant results using low-power statistical tests can lead to an 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, suggesting a potential for 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.

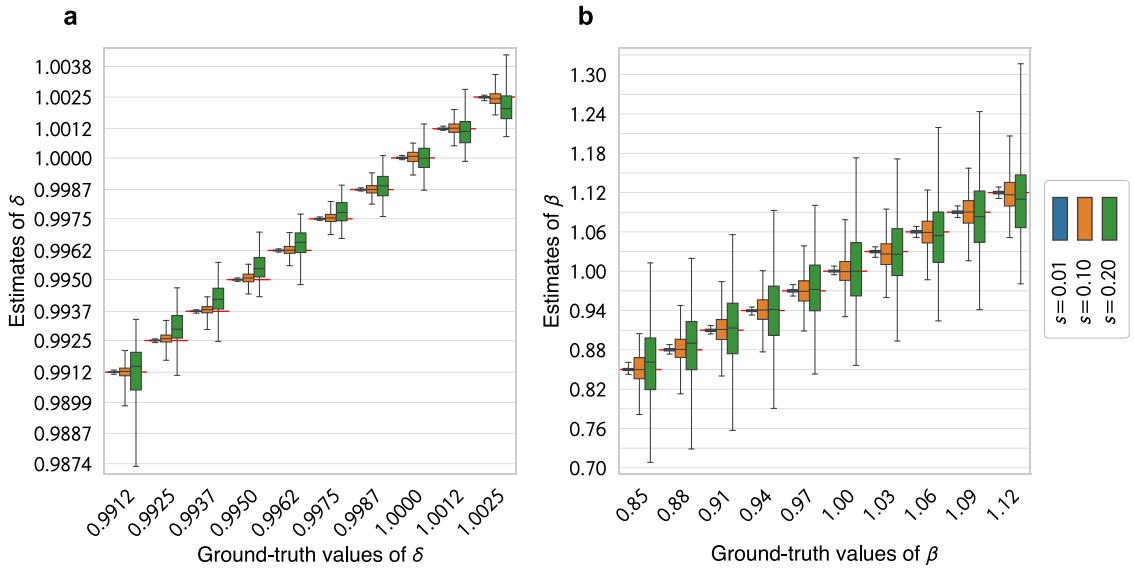


Fig. 2. Box plot of the estimates.
Notes: Estimates of (**a**) δ and (**b**) β. Each box summarizes 10 replications of all combinations of ground-truth values of β (or δ) and $\ln\sigma$, *i.e.*, 700 simulation agents. The line in the center of the box represents the median. The two ends of the box represent the first and third quartiles, respectively, while the two ends of the whiskers represent the 5th and 95th percentiles, respectively. On the red line, the error of the estimate is 0.

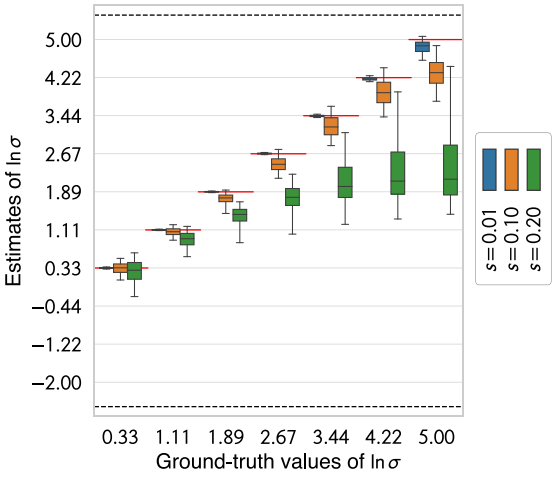


Fig. 3. Box plots of the $\ln\sigma$ estimates.
Notes: Each box summarizes 10 replications of all combinations of ground-truth values for δ and β, representing a total of 1,000 simulation agents. The two ends of the box represent the first and third quartiles, respectively, while the two ends of the whiskers represent the 5th and 95th percentiles, respectively. On the red line, the error of the estimate is 0. To converge the estimation, we used the sigmoid function $\ln\sigma = f(\theta) = 4\tanh(\theta) + 1.5$, which allowed the estimated $\ln\sigma$ to range between -2.5 and 5.5. The black horizontal dashed lines in the figure denote the boundaries of $\ln\sigma$.

3.3. Curvature parameter

Fig. 3 illustrates the distribution of the estimated curvature parameter, $\ln\sigma$, through box plots. 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 $s = 0.20$, $\widehat{\ln\sigma}$ appears to saturate at approximately 2.2.

Next, we examine how the estimated curvature relates to the errors in estimating the parameters δ and β. Fig. 4 presents a scatter plot that illustrates the estimation errors of δ and β against the estimated values of the curvature parameter $\ln\sigma$, revealing several noteworthy observations.

First, it is evident that when the estimated value of $\ln\sigma$ is small, indicating a stronger concavity, both δ and β exhibit more significant estimation errors. This observation implies that the degree of concavity significantly affects the precision of estimating these parameters. Second, regardless of the estimated values of $\ln\sigma$, the distribution of errors for both parameters is symmetrical around zero. This symmetry suggests that the estimation of δ and β, though affected by the curvature parameter in terms of precision, does not suffer from a systematic bias toward either overestimation or underestimation.

Furthermore, attention is drawn to the region indicating linearity, where larger values of $\widehat{\ln\sigma}$ are observed. Observations for the lower noise level, $s = 0.05$, reveal that as the ground-truth values of $\ln\sigma$ increase, the horizontal width of the distribution also broadens. This broadening indicates a decline in the precision of $\widehat{\ln\sigma}$. For the high noise level, $s = 0.20$, the precision of $\widehat{\ln\sigma}$ diminishes to an extent where distinguishing between groups of ground-truth values becomes challenging. However, it is imperative to note that despite the diminished estimation precision of $\ln\sigma$, it does not necessarily translate to more significant estimation errors for $\hat{\delta}$ and $\hat{\beta}$.

The development of methods such as CTB and DMPL emerged from the recognition of certain biases inherent in experimental settings. Traditionally, several experiments e.g., (Coller and Williams, 1999; Harrison et al., 2002) have assumed linear utility and measured the discount factor δ. However, it was noted that this approach often leads to either an underestimation of δ or, on the flip side, an overestimation of the discount rate (for an overview, see Frederick et al., 2002).

This issue arises when attempting to fit the behaviors of individuals, who actually have concave utilities, into a linear utility model. Due to what is known as Jensen's inequality, a bias in the discount factor invariably occurs mathematically. Nonetheless, our simulation results indicate that if the estimated $\ln\sigma$ is somewhat large, the bias in estimating the discount factor can be considered minor, even if the true curvature is incorrectly estimated to be more linear than it actually is.[12]

[12] Our findings align with those of Cheung (2020), who observed that changing discount rates based on the utility function's concavity has a minimal impact. Cheung (2020) measured curvature using a method distinct from the CTB approach. The results showed that while people's preferences tend to lean toward concavity, they are, to a degree, nearly linear.

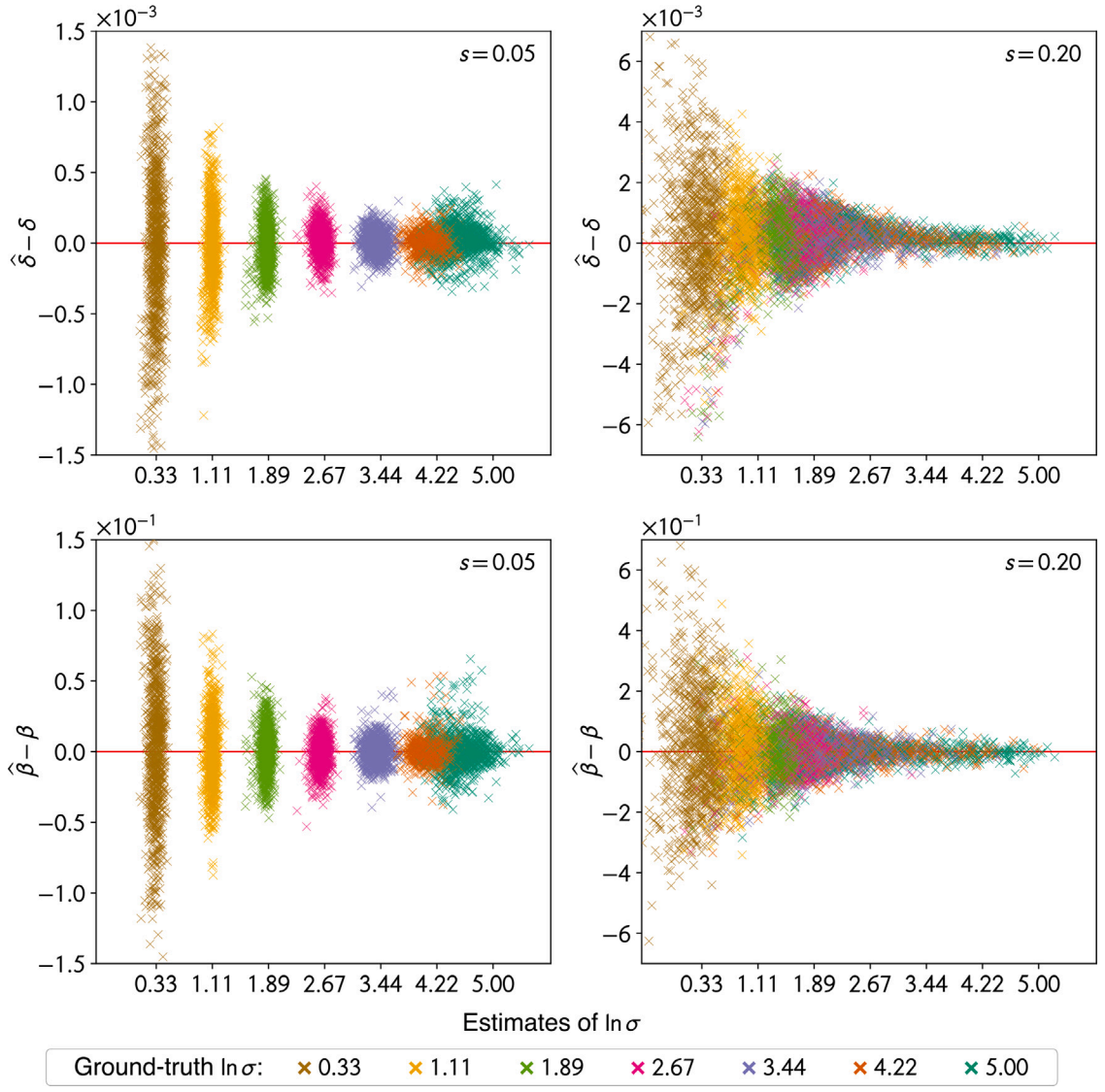


Fig. 4. Scatter plots of estimates $\widehat{\ln\sigma}$ vs. estimation errors of $\hat{\delta}$ and $\hat{\beta}$.
Notes: The vertical axis measures the signed error, which is the difference between the estimated and ground-truth values. The colors of the markers differ according to the ground-truth values of $\ln\sigma$. The top and bottom panels are related to $\hat{\delta}$ and $\hat{\beta}$, respectively. The panels on the left and the right correspond to the cases of $s = 0.05$ and $s = 0.20$, respectively.

3.4. Why is the present bias estimation precision low?

In the CES-QHD utility function, δ and β appear as the term in $D = \beta\delta^k$ for $t = 0$ and as $D = \delta^k$ for $t = 1$. If the available data for parameter estimation is only for the case of $t = 0$, we cannot uniquely identify δ and β. However, as we indeed have data for both cases, $t = 0$ and $t = 1$, we should be able to identify the parameters mathematically.

Fig. 5 shows a scatter plot of the estimated values of δ and β (for $\ln\sigma = 2.67$ and $s = 0.01$; see Appendix E for the scatter plots including all values of $\ln\sigma$ and s), along with a red line that satisfies the equation $\beta\delta^{70} = 1$. Note that both axes use a logarithmic scale centered at 1, and 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 Fig. 5 is the distribution of points along the red line. Theoretically, identifying δ and β should be possible; however, in practice, it is difficult, even though the value of D itself can be estimated with reasonable accuracy.

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, whereas a 1% change in δ results in a k% change in D. Given that 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 $\beta = 1$, yielding a 3% variation in D, which is three times smaller than the variation seen with δ.

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 $1 + r$.

Fig. 6 shows the demand curve representing the relationship between the price $1 + r$ and the amount that individuals are willing to allocate to the earlier period for $\ln\sigma = 2.67$. Note that the horizontal axis, representing price $1+r$, uses a logarithmic scale and the prices are indicated on the vertical lines in the figure.

In Fig. 6, 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 illustrated by the difference between the blue and orange dashed curves. The difference when only β decreases from 1 to 0.97 is represented by the difference between the blue and green dotted curves. We observe that the discount behavior for $\delta = 0.9987$ is more significant than that for $\beta = 0.97$.

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 to ascertain

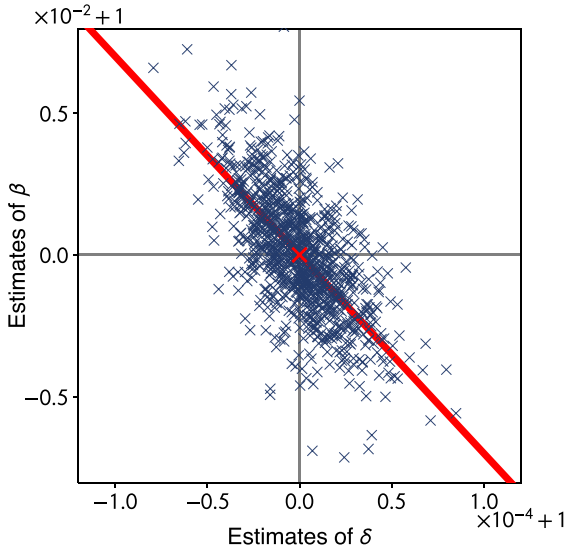


Fig. 5. Scatter plot of estimated δ and β.
Notes: The figure shows the case $\ln\sigma = 2.67$ and $s = 0.01$. Each point is shifted so that the pair of corresponding ground-truth values aligns with the red point $(\delta, \beta) = (1, 1)$. Both axes are presented on a logarithmic scale, centered at 1. Along the red line, the equation $\beta\delta^{70} = 1$ is satisfied.

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 precision of δ estimation is sufficient to distinguish between individuals with a true difference of about 1.3×10^{-3}, which corresponds to the spacing of our ground-truth values, while the precision of β estimation is just enough to distinguish a difference of about 0.1, which is approximately three times the spacing of our ground-truth values.

Eventually, the low precision of β estimation occurs because we attempt to identify values within a very narrow range with high precision. As is clear from the comparison of demand curves in Fig. 6, when the true difference in the values of β is less than 0.1, identifying individuals becomes inherently difficult regardless of the econometric method used, because the differences in behavior are minimal. In the expanded β range, distinguishing between two individuals with given ground-truth values becomes possible (see Appendix F).

Although δ intuitively seems to require very high precision because it is a daily discount factor, and values should thus be precisely estimated to the fourth decimal place, estimating it with sufficient precision is feasible due to its broader range compared with that of β. As k depends on the scale of δ, we must expand the range of δ if we are to make k a weekly discount factor. It should be noted that changing the value of k does not improve the precision of β's estimation (see Appendix G).

3.5. Some attempts to improve parameter estimation

In the previous subsections, we discussed how accurately we can estimate the utility parameters by understanding the relationship between the shape of the utility function and our expected range of these estimates. However, the accuracy of estimating parameters is also influenced by the methods employed for calculation and the setup of the experiments. Next, we examine how improving these computation methods and experiment designs can enhance the accuracy of our estimates. Figs. 7 and 9 display the median absolute errors of the parameter estimates for each described scenario, offering a comparison with the main simulation, labeled as "BASE".

We begin by exploring a situation in which the true values of the parameters δ and $\ln\sigma$ are already known. How well can we then estimate the parameter β? To investigate this, we again utilized the synthesized decision data, which was employed in the main simulation. Here, we use the true values of δ and $\ln\sigma$ to focus solely on estimating β. This simulation is labeled as M1 in Fig. 7 (see Appendix H for details). The results show an improvement in the overall accuracy of the β estimates. For instance, when the actual value of β is 0.97, our ability to correctly reject the null hypothesis that $\widehat{\beta} = 1$ increases from 43% to 59% when the noise level is $s = 0.05$. However, at a higher noise level of $s = 0.20$, the success rate remains at 13%.

Estimating δ and $\ln\sigma$ using only partial data, where a front-end delay exists (*i.e.*, $t = 1$), did not, surprisingly, make the estimates of δ less accurate compared with the accuracy when using all the data to estimate all three parameters. This result may suggest one reason to use the CTB method: if the goal is not to detect the effects of present bias, but solely to estimate the discount factor and the curvature of utility simultaneously, it is possible to achieve sufficient accuracy for δ with only 21 tasks.

Furthermore, we experimented with keeping δ and $\ln\sigma$ at their estimated values while estimating β over the entire dataset. This simulation is labeled as M2 in Fig. 7 (see Appendix I for details). We observed a slight deterioration in the median error magnitude in the aggregated data. However, an improvement was noted in the performance of detecting the present bias through testing the null hypothesis that $\widehat{\beta} = 1$. If the true value of β is 0.97, the percentage of successfully rejecting the null hypothesis increases from 43% to 57% when the noise level is $s = 0.05$. Although this improvement is modest, it is important to note that dividing the estimation process into two stages actually leads to better outcomes. This finding is significant, suggesting that a two-stage approach could provide more reliable results.

We also consider a method that exclusively uses true curvature parameters, assuming these can be accurately identified — setting aside the debate on their compatibility with risk attitudes — to estimate δ and β. This simulation is labeled as M3 in Fig. 7 (see Appendix J for details). This leads to an important question: does using true curvature parameters solely for estimating δ and β make the estimations more accurate? Surprisingly, the accuracy in estimating δ and β may worsen. This unexpected outcome often occurs in situations where the value of $\ln\sigma$ is high, suggestive of a linear utility. In such cases, the values of the demand function tend to disregard minor differences in estimating parameters and take on extreme values of 0 or 1, consequently making accurate parameter estimation challenging.

One possible solution to address this problem is to cap the $\ln\sigma$ value at 2.5, regardless of its actual value. Adopting this solution improves the success rate of detecting the discounting, $\widehat{\delta} < 1$. However, while the estimation precision improves, a noticeable estimation bias emerges. Moreover, adopting this strategy has mixed results on accurately identifying the present bias, $\widehat{\beta} < 1$: success rates drop when the true value is 0.97, yet improve for a true value of 0.85. It is important to be careful when estimating δ and β by setting the curvature parameter equal to a fixed value, as this can harm the accuracy of estimations.

Instead of using the entire dataset simultaneously, we divided it based on the front-end delay, t. This approach allows us to identify the discount factors in two distinct scenarios: $\tilde{\delta}_{t=1} = 1 \times \delta$ for $t = 1$ and $\tilde{\delta}_{t=0} = \beta^{1/70}\delta$ for $t = 0$. By calculating the ratio of these discount factors across t, we gained insight into what we refer to as the present bias parameter: $\beta = (\tilde{\delta}_{t=0}/\tilde{\delta}_{t=1})^{70}$. This simulation is labeled as M4 in Fig. 7 (see Appendix K for details). However, our simulation revealed that this method decreases the accuracy of the β estimation.

We then sought a more straightforward method to observe present bias by directly examining decisions, thus avoiding the need to estimate the parameter β. To this end, we employed the two-sided paired t-test to assess decisions across 21 varying prices at the two front-end

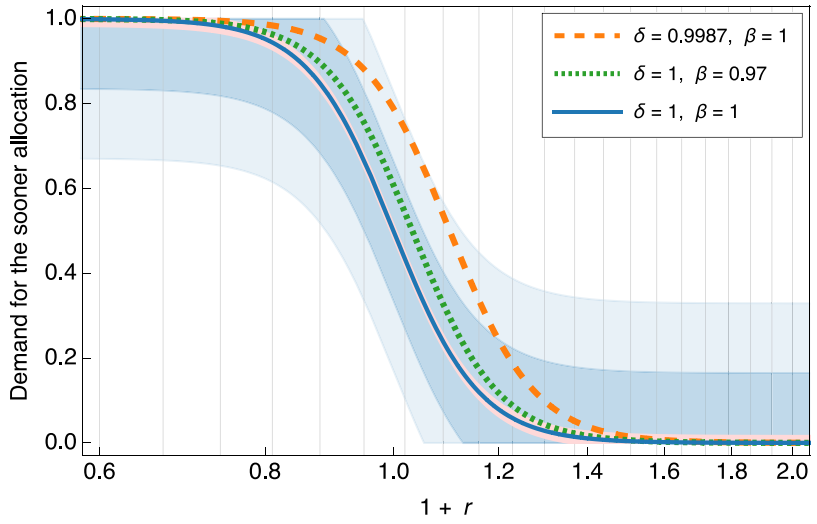
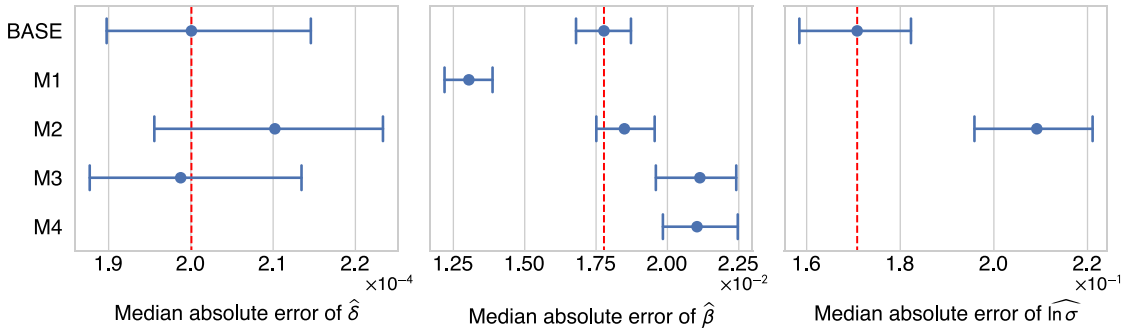


Fig. 6. Demand curves.
Notes: Each demand curve corresponds to an individual whose curvature parameter is $\ln\sigma = 2.67$. The demand curve represents the relationship between the price $1+r$ and the amount individuals are willing to allocate to the earlier period. The horizontal axis, representing price $1+r$, employs a logarithmic scale. The curve, depicted as a solid blue line, is accompanied by three types of error bands. The shaded area represents the range from the 5th to the 95th percentile of noise. The colors signify different bandwidths: red (indicating a very narrow bandwidth) corresponds to $s = 0.01$, dark blue to $s = 0.10$, and light blue to $s = 0.20$. The individual faces the decision problem of allocating endowments based on the prices, which are indicated by the vertical lines, between the present ($t = 0$) and a future period ($k = 70$ days later).



BASE	The main simulation.
M1	The simulation that estimates only the parameter β while fixing the parameters δ and $\ln\sigma$ at their ground-truth values.
M2	The simulation of the two-stage estimation that estimates the parameters δ and $\ln\sigma$ using partial data, where $t = 1$. Then, fixing these estimated values, it estimates only β using the full dataset.
M3	The simulation that estimates the parameters δ and β while fixing the parameter $\ln\sigma$ at its ground-truth value.
M4	The simulation that estimates the biased discount factor $\tilde{\delta}_{t=0} = \beta^{1/70}\delta$ and the curvature $\ln\sigma$ using partial data, where $t = 0$. Then, it estimates β using the obtained $\tilde{\delta}_{t=0}$ estimate and the δ estimate from M2.

Fig. 7. Median absolute errors of the estimates for each method.*Notes:*Error bars represent the bootstrap 95% confidence intervals.

delays, $t = 0$ and $t = 1$. Our results indicate that this direct comparison of decisions proved more challenging for identifying present bias than estimating the parameter β (see Appendix L for details).

Harrison et al. (2013) pointed out that in CTB experiments, individuals frequently choose the corner points on the budget line. They suggested that a multinomial logit model could better capture this behavior than AS's approach of applying the least squares method to fit demand functions. This insight prompts further investigation into how well multinomial logit models can estimate.

The proposed method divides the budget line into 101 equal segments, ranging from 0 to 100, thus transforming a continuous choice into 101 specific options. To find out how likely each choice is, we calculate their probabilities. The chance of picking the ith option is shown by the formula $\exp(U_i)/\sum_{j=0}^{100}\exp(U_j)$, where U_i represents the CES-QHD utility value of choosing the ith option. We utilized these probabilities for the maximum likelihood estimation.

When comparing the accuracy of the estimates from the multinomial logit models with those obtained by directly fitting demand

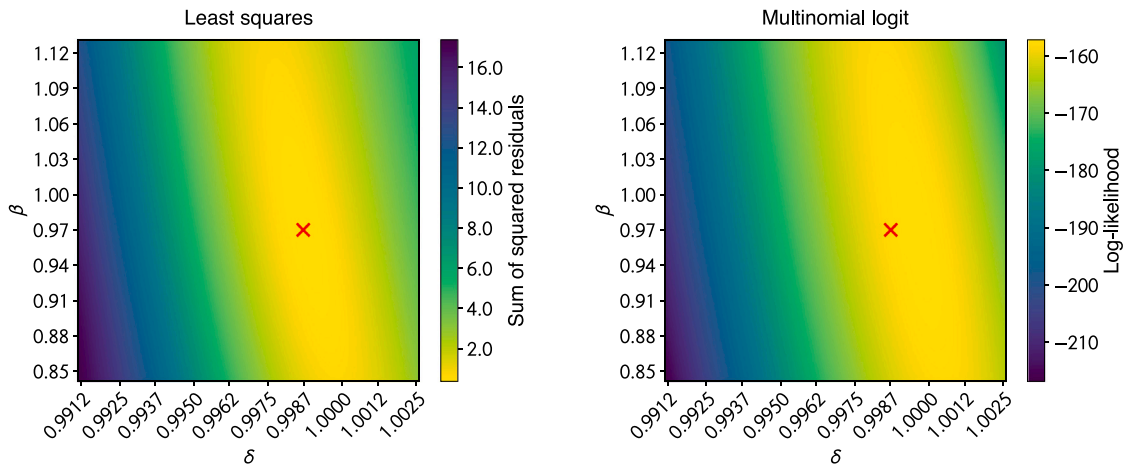


Fig. 8. Goodness of fit with respect to δ and β.
Notes: The goodness of fit is visualized using heatmaps with respect to δ and β, fixing the value of $\ln\sigma$ to its ground-truth value. For illustrative purposes, we generated the figures for one of the synthetic decision datasets with the truncated noise, where the agent's ground-truth values are $\delta = 0.9987$, $\beta = 0.97$, $\ln\sigma = 2.67$, and $s = 0.10$. The red point indicates the pair of ground-truth values. **Left**: The sum of squared residuals in the least squares method. **Right**: The log-likelihood function in the multinomial logit model.

functions, the multinomial logit model estimates generally exhibit less unbiasedness and precision (see Appendix A for details). This difference in accuracy requires careful interpretation because the two approaches employ distinct computational methods.[13]

Exploring these results further, the heatmaps in Fig. 8 visualize the model's goodness of fit with respect to the parameters δ and β for one set of synthesized data. In the left panel of the figure, we observe the squared residuals for the least squares method; in the right panel, the figure displays the log-likelihood function for the multinomial logit model. These visuals aid in comparing the two methods of estimating parameters.

A significant observation is that, for both methods, gradients in the goodness of fit with respect to β were flatter than those with respect to δ. This implies that accurately finding the optimal point for β is more difficult. Whether using the least squares approach or the multinomial logit model approach, estimating β accurately is more challenging than estimating δ.

Let us now explore how we can better estimate variables in the design of tasks for CTB experiments.

In our main simulation, we use a set of problems, labeled "BASE" in Fig. 9: we fix m at 20 and k at 70, and choose 21 prices that range between 0.6 and 2 for two front-end delays, $t = 0$ and $t = 1$. This results in 42 distinct problems. We have the flexibility to modify several aspects of the experimental tasks, such as the start period t, the delay k, and the prices $1+r$. By adjusting these elements of the experimental design, we aim to enhance the accuracy of our parameter estimations (see Appendix M for details).

First, we aimed to determine whether merely adding more problems would prove beneficial. Therefore, in PS1, we selected 42 prices, which is double that of the BASE, all within the same range. For PS2, we aimed even higher and chose 210 prices, which is ten times larger than the BASE, again within the same range. [14]

We discovered that having a greater number of problems unequivocally improves our estimates. Comparing the estimation precision of PS1 and PS2 with that of BASE, it is clear that PS1 and PS2 show an improvement in precision. This enhancement is attributed to the increased number of problems, as precision appears to be positively correlated with the number of tasks. Despite PS2 comprising a total of 420 tasks, it remains challenging to reject the null hypothesis that an individual's β estimate, with a true value of 0.97, is equal to 1 when the noise size is $s = 0.20$.

When the number of prices remains at 21, increasing the variety of delay k to two cases (PS3: $k = 35, 70$) and to 10 cases (PS4: $k = 35, 42, 49, \ldots, 98$) also improves estimation precision. PS1 and PS3, as well as PS2 and PS4, have the same number of tasks, and their estimation precision is approximately the same.

We then consider the strategy of altering the number of types for each variable without increasing the total number of problems, which remains at 42.

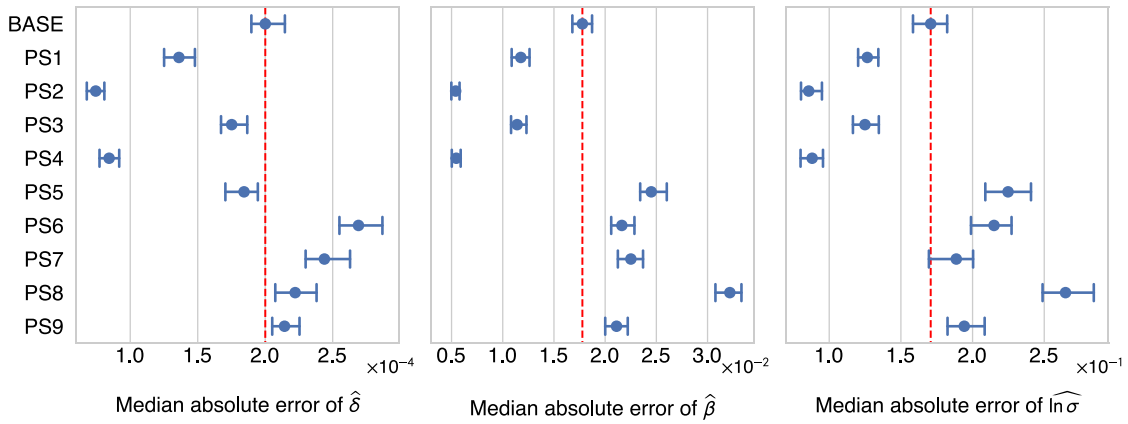
In BASE, while k was fixed to one case, $1 + r$ was set to 21 different cases. Conversely, in PS5, we explore a method in which the price is fixed to one case, whereas k is set to 21 different cases. As already discussed, k must always be a positive value. Furthermore, if the price is set at a value that is only greater than 1, estimating δ becomes challenging for individuals with a true δ value greater than 1. Therefore, instead of setting k to a negative value for some problems, the interest rate was made negative. For $1 + r = 1.25$, 15 cases of k (23, 24, 25, 27, 29, 31, 34, 38, 43, 50, 60, 75, 105, 181, 789) were set; and for $1 + r = 0.75$, six cases of k (39, 50, 67, 96, 158, 393) were set, totaling 21 combinations of $1 + r$ and k. As there are two types for each of the 21 tasks, $t = 0$ and $t = 1$, the total number of problems equals 42. Here, the values of k were chosen to ensure that the equation, $k = 70\ln(1.25)/\ln(1 + r)$ for any prices $1 + r > 1$ in BASE, and $k = 70\ln(0.75)/\ln(1+r)$ for any prices $1+r < 1$ in BASE, is satisfied.[15]

For PS6, 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 values to seven (35, 45.5, 56, 66.5, 77, 87.5, 98). In PS7, we increased the number of prices, $1 + r$, to seven for each combination of t and k, as opposed to setting the quantity of k equal to three values (35, 70, and 98). PS8 differs from PS7 in that the values of k are altered to 35, 175, and 350. For PS9, we fixed k at 70 and drew 14 prices from the range $0.6 \leq 1 + r \leq 2$, once at $t = 0$ and twice at $t = 1$. PS9 corresponds to setting two different t situations for when $t > 0$—*i.e.*, two

[13] The fitting of demand functions was conducted using Stata's "`nl`" command, which relies on the Gauss–Newton algorithm. Conversely, the maximum likelihood estimation for the multinomial logit model utilized R's "`optim`" function, based on the Nelder–Mead algorithm.

[14] For PS1, PS2, and the subsequently mentioned PS3 and PS4, because evaluating the standard error of the parameter estimates using the jackknife method would take a long time, we opted for an alternative approach by computing using the inverse of the negative Hessian.

[15] For individuals with linear utility, where the true β is 1, the switch in choices from the early period to the late period, due to price increases, occurs at a price of $1 + r = \delta^{-k}$. Using this equation, we determined the value of k corresponding to each problem in BASE.



	t	k	$1+r$	#
BASE	$\{0,1\}$	$\times\{70\}$	$\times\{\underbrace{0.60, 0.67, \ldots, 2.00}_{21}\}$	42
PS1	$\{0,1\}$	$\times\{70\}$	$\times\{\underbrace{0.60, 0.63, \ldots, 2.00}_{42}\}$	84
PS2	$\{0,1\}$	$\times\{70\}$	$\times\{\underbrace{0.600, 0.607, \ldots, 2.000}_{210}\}$	420
PS3	$\{0,1\}$	$\times\{35, 70\}$	$\times\{\underbrace{0.60, 0.67, \ldots, 2.00}_{21}\}$	84
PS4	$\{0,1\}$	$\times\{\underbrace{35, 42, 49, \ldots, 98}_{10}\}$	$\times\{\underbrace{0.60, 0.67, \ldots, 2.00}_{21}\}$	420
PS5	$\{0,1\}$	$\times\{\underbrace{23, \ldots, 789}_{15}\}$	$\times\{1.25\}$	42
	$\{0,1\}$	$\times\{\underbrace{39, \ldots, 393}_{6}\}$	$\times\{0.75\}$	
PS6	$\{0,1\}$	$\times\{\underbrace{35, 45.5, 56, \ldots, 98}_{7}\}$	$\times\{0.9, 1.2, 1.5\}$	42
PS7	$\{0,1\}$	$\times\{35, 70, 98\}$	$\times\{\underbrace{0.60, 0.83, \ldots, 2.00}_{7}\}$	42
PS8	$\{0,1\}$	$\times\{35, 175, 350\}$	$\times\{\underbrace{0.60, 0.83, \ldots, 2.00}_{7}\}$	42
PS9	$\{0,1,1\}$	$\times\{70\}$	$\times\{\underbrace{0.60, 0.71, \ldots, 2.00}_{14}\}$	42

Fig. 9. Median absolute errors of the estimates for each problem set.*Notes:* Error bars represent the bootstrap 95% confidence intervals.

cases for $k = 70$: one occurring between 7 and 77 days later, and the other between 35 and 105 days later.

Upon inspecting Fig. 9, it becomes apparent that, in terms of median absolute error size, PS5 through PS9 exhibit more degraded performance than BASE. Although one might intuitively think that having multiple k conditions, as is the case with the AS problem set, would reduce estimation errors more than having only one k value, in reality, no clear patterns regarding the relationship between problem set composition and estimation performance were found. 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 accuracy.

4. Discussion

This paper evaluates the inaccuracy of the estimates of the CES-QHD utility parameters obtained using the CTB experiment (Andreoni and Sprenger, 2012a) by performing parameter recovery simulations (Wilson and Collins, 2019). Figs. 1 and 2 demonstrate that the precision of the estimation of the time discount factor δ is sufficient 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, compared 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 demand behavior predicted by a β value of 0.97 compared with a β value of 1 (as depicted in Fig. 6), the differences are sufficiently small and are barely detectable in the presence of noise. Any true value of β closer to 1 than 0.9 cannot be reliably identified. Unfortunately, however, the estimates of β reported in the literature, in fact, fall within that range. Given the low precision of the β estimation, there is a possibility of both overestimating and 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, incorporating more tasks can counteract the impact of noise and

enhance the accuracy of estimation. However, it would be impractical to increase the number of tasks further due to the workload imposed on participants during the experiment. 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).

Here, although not exhaustive, let us discuss some indicative considerations regarding parameter estimation accuracy using the multiple price list (MPL) method (Coller and Williams, 1999; Harrison et al., 2002; Andersen et al., 2008) compared with the CTB method. To summarize the conclusion first, the problem we identified in the CTB method — the low precision in estimating the present bias parameter β in the CES-QHD utility function — is not immediately resolved by adopting the MPL method.

The tasks in the MPL experiment are nested within the tasks in the CTB experiment. In the CTB experiment, participants can choose any point on the budget line, whereas in the MPL experiment, participants are forced to choose between two corner points on the budget line. Therefore, for individuals with concave utility who selected interior points in the CTB experiment, their decision data degenerate into binary data in MPL experiments, resulting in a loss of information. To the extent that the goal is to estimate the parameters of the CES-QHD utility function, the decision data obtained through the CTB experiment is at least as rich in terms of information as that obtained through the MPL experiment.

Andersen et al. (2008) proposed the DMPL method to correct the bias in discount rate estimation that results from assuming linear utility. However, the DMPL method — correcting the discount rate using measures of risk attitudes — remains controversial (Andreoni and Sprenger, 2012b; Abdellaoui et al., 2013; Harrison et al., 2013; Andersen et al., 2014; Andreoni and Sprenger, 2015; Cheung, 2015; Miao and Zhong, 2015; Andersen et al., 2018; Cheung, 2020). Even if we accept that corrections based on risk attitude measurements are valid, it should be noted that this concerns the issue of parameter estimation unbiasedness and is not directly related to the precision of the estimation.

When considering the actual implementation of the MPL experiment, it is often the case that forcing respondents to indicate their unique switching point prevents inconsistencies e.g., (Andersen et al., 2006; Tanaka et al., 2010).[16] In such cases, the parameters characterizing individuals are determined from the consistent switching points with a one-to-one correspondence; thus, it can only be said that the amount of noise is at most equal to the step in the price list. However, it should be noted that assuming the estimation error is limited to approximately the price increments mistakenly overlooks the estimation errors due to probabilistic noise in individual decision-making. Concerning probabilistic noise, it is necessary to iteratively conduct several MPL tasks and observe the variations in responses.

Meanwhile, there are benefits to using MPL experiments in terms of improving estimation accuracy. Once it has been determined that the switching point is between two prices from the list, it is possible to adaptively present a new list of prices to more precisely identify the switching point between those two prices. In the case of CTB experiments, however, parameter estimation can only be performed once all decision data are collected, making it difficult to generate tasks adaptive to respondents' answers.[17]

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.[18] We believe that researchers persisting in the use of the CTB experiment will need to significantly overhaul behavior modeling. We additionally recommend the use of parameter recovery simulations.

CRediT authorship contribution statement

Keigo Inukai: Writing – review & editing, Methodology, Funding acquisition, Conceptualization. **Yuta Shimodaira:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Kohei Shiozawa:** Writing – review & editing, Methodology, Conceptualization, Funding acquisition.

Declaration of competing interest

None

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jbef.2024.100962.

References

Abdellaoui, M., Bleichrodt, H., L'Haridon, O., Paraschiv, C., 2013. Is there one unifying concept of utility? An experimental comparison of utility under risk and utility over time. Manage. Sci. 59, 2153–2169. http://dx.doi.org/10.1287/mnsc.1120.1690.

Andersen, S., Harrison, G.W., Lau, M.I., Rutström, E.E., 2006. Elicitation using multiple price list formats. Exp. Econ. 9, 383–405. http://dx.doi.org/10.1007/s10683-006-7055-6.

Andersen, S., Harrison, G.W., Lau, M.I., Rutström, E.E., 2008. Eliciting risk and time preferences. Econometrica 76, 583–618. http://dx.doi.org/10.1111/j.1468-0262.2008.00848.x.

Andersen, S., Harrison, G.W., Lau, M.I., Rutström, E.E., 2014. Discounting behavior: A reconsideration. Eur. Econ. Rev. 71, 15–33. http://dx.doi.org/10.1016/j.euroecorev.2014.06.009.

Andersen, S., Harrison, G.W., Lau, M.I., Rutström, E.E., 2018. Multiattribute utility theory, intertemporal utility, and correlation aversion. Internat. Econom. Rev. 59, 537–555. http://dx.doi.org/10.1111/iere.12279.

Andreoni, J., Kuhn, M.A., Sprenger, C., 2015. Measuring time preferences: A comparison of experimental methods. J. Econ. Behav. Organ. 116, 451–464. http://dx.doi.org/10.1016/j.jebo.2015.05.018.

Andreoni, J., Sprenger, C., 2012a. Estimating time preferences from convex budgets. Amer. Econ. Rev. 102, 3333–3356. http://dx.doi.org/10.1257/aer.102.7.3333.

Andreoni, J., Sprenger, C., 2012b. Risk preferences are not time preferences. Amer. Econ. Rev. 102, 3357–3376. http://dx.doi.org/10.1257/aer.102.7.3357.

Andreoni, J., Sprenger, C., 2015. Risk preferences are not time preferences: Reply. Amer. Econ. Rev. 105, 2287–2293. http://dx.doi.org/10.1257/aer.20150311.

Atlas, S.A., Johnson, E.J., Payne, J.W., 2017. Time preferences and mortgage choice. J. Mar. Res. 54, 415–429. http://dx.doi.org/10.1509/jmr.14.0481.

Augenblick, N., Niederle, M., Sprenger, C., 2015. Working over time: Dynamic inconsistency in real effort tasks. Q. J. Econ. 130, 1067–1115. http://dx.doi.org/10.1093/qje/qjv020.

Bickel, W.K., Odum, A.L., Madden, G.J., 1999. Impulsivity and cigarette smoking: delay discounting in current, never, and ex-smokers. Psychopharmacology 146, 447–454. http://dx.doi.org/10.1007/PL00005490.

Blumenstock, J., Callen, M., Ghani, T., 2018. Why do defaults affect behavior? Experimental evidence from Afghanistan. Amer. Econ. Rev. 108, 2868–2901. http://dx.doi.org/10.1257/aer.20171676.

Camerer, C.F., Dreber, A., Forsell, E., Ho, T.H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., Wu, H., 2016. Evaluating replicability of laboratory experiments in economics. Science 351, 1433–1436. http://dx.doi.org/10.1126/science.aaf0918.

[16] Regarding multiple switching in MPL experiments, Yu et al. (2021) provided a detailed investigation.

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

[18] 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.

Carvalho, L.S., Meier, S., Wang, S.W., 2016. Poverty and economic decision-making: Evidence from changes in financial resources at payday. Amer. Econ. Rev. 106, 260–284. http://dx.doi.org/10.1257/aer.20140481.

Cheung, S.L., 2015. Comment on risk preferences are not time preferences: On the elicitation of time preference under conditions of risk. Amer. Econ. Rev. 105, 2242–2260. http://dx.doi.org/10.1257/aer.20120946.

Cheung, S.L., 2016. Recent developments in the experimental elicitation of time preference. J. Behav. Exp. Finance 11, 1–8. http://dx.doi.org/10.1016/j.jbef.2016.04.001.

Cheung, S.L., 2020. Eliciting utility curvature in time preference. Exp. Econ. 23, 493–525. http://dx.doi.org/10.1007/s10683-019-09621-2.

Cheung, S.L., Tymula, A., Wang, X., 2021. Quasi-hyperbolic present bias: A meta-analysis. In: Life Course Centre Working Paper No. 2021-15. Life Course Centre, URL https://ssrn.com/abstract=3909663.

Cheung, S.L., Tymula, A., Wang, X., 2022. Present bias for monetary and dietary rewards. Exp. Econ. 25, 1202–1233. http://dx.doi.org/10.1007/s10683-022-09749-8.

Cohen, J., Ericson, K.M., Laibson, D., White, J.M., 2020. Measuring time preferences. J. Econ. Lit. 58, 299–347. http://dx.doi.org/10.1257/JEL.20191074.

Coller, M., Williams, M.B., 1999. Eliciting individual discount rates. Exp. Econ. 2, 107–127. http://dx.doi.org/10.1007/bf01673482.

Dantas, A.M., Sack, A.T., Bruggen, E., Jiao, P., Schuhmann, T., 2022. The effects of probiotics on risk and time preferences. Sci. Rep. 12, 12152. http://dx.doi.org/10.1038/s41598-022-16251-x.

Frederick, S., Loewenstein, G., O'donoghue, T., 2002. Time discounting and time preference: A critical review. J. Econ. Lit. 40, 351–401. http://dx.doi.org/10.1257/jel.40.2.351.

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

Harrison, G.W., Lau, M.I., Williams, M.B., 2002. Estimating individual discount rates in Denmark: A field experiment. Amer. Econ. Rev. 92, 1606–1617. http://dx.doi.org/10.1257/000282802762024674.

Holt, D.D., Green, L., Myerson, J., 2003. Is discounting impulsive?: Evidence from temporal and probability discounting in gambling and non-gambling college students. Behav. Process. 64, 355–367. http://dx.doi.org/10.1016/S0376-6357(03)00141-4.

Holt, C.A., Laury, S.K., 2002. Risk aversion and incentive effects. Amer. Econ. Rev. 92, 1644–1655. http://dx.doi.org/10.1257/000282802762024700.

Imai, T., Camerer, C.F., 2018. Estimating time preferences from budget set choices using optimal adaptive design. https://www.taisukeimai.com/api/resources/adaptive_ctb.pdf. (Accessed 30 November 2022).

Imai, T., Rutter, T.A., Camerer, C.F., 2020. Meta-analysis of present-bias estimation using convex time budgets. Econ. J. 131, 1788–1814. http://dx.doi.org/10.1093/ej/ueaa115.

Inukai, K., Shimodaira, Y., Shiozawa, K., 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, URL https://econpapers.repec.org/paper/dprwpaper/1195.htm.

Kirby, K.N., Petry, N.M., Bickel, W.K., 1999. Heroin addicts have higher discount rates for delayed rewards than non-drug-using controls. J. Exp. Psychol. General 128, 78–87. http://dx.doi.org/10.1037/0096-3445.128.1.78.

Laibson, D., 1997. Golden eggs and hyperbolic discounting. Q. J. Econ. 112, 443–478. http://dx.doi.org/10.1162/003355397555253.

Meier, S., Sprenger, C., 2010. Present-biased preferences and credit card borrowing. Am. Econ. J. Appl. Econ. 2, 193–210. http://dx.doi.org/10.1257/app.2.1.193.

Meier, S., Sprenger, C., 2012. Time discounting predicts creditworthiness. Psychol. Sci. 23, 56–58. http://dx.doi.org/10.1177/0956797611425931.

Miao, B., Zhong, S., 2015. Comment on risk preferences are not time preferences: Separating risk and time preference. Amer. Econ. Rev. 105, 2272–2286. http://dx.doi.org/10.1257/aer.20131183.

O'Donoghue, T., Rabin, M., 1999. Doing it now or later. Amer. Econ. Rev. 89, 103–124. http://dx.doi.org/10.1257/aer.89.1.103.

O'Donoghue, T., Rabin, M., 2015. Present bias: Lessons learned and to be learned. Amer. Econ. Rev. 105, 273–279. http://dx.doi.org/10.1257/aer.p20151085.

Tanaka, T., Camerer, C.F., Nguyen, Q., 2010. Risk and time preferences: Linking experimental and household survey data from Vietnam. Amer. Econ. Rev. 100, 557–571. http://dx.doi.org/10.1257/aer.100.1.557.

Thöni, C., 2015. A note on CES functions. J. Behav. Exp. Econ. 59, 85–87. http://dx.doi.org/10.1016/j.socec.2015.10.001.

van Zwet, E.W., Cator, E.A., 2021. The significance filter, the winner's curse and the need to shrink. Stat. Neerl. 75, 437–452. http://dx.doi.org/10.1111/stan.12241.

Wilson, R.C., Collins, A.G., 2019. Ten simple rules for the computational modeling of behavioral data. eLife 8, http://dx.doi.org/10.7554/eLife.49547.

Yu, C.W., Zhang, Y.J., Zuo, S.X., 2021. Multiple switching and data quality in the multiple price list. Rev. Econ. Stat. 103, 136–150. http://dx.doi.org/10.1162/rest_a_00895.