Empirical Test on MEL Task

Zijian Zark Wang

2023-04-17

1 The Model

1.1 Attention Reallocation Process

Suppose a decision maker is to evaluate a sequence of rewards, which depends on the state of world s. The time length of this sequence is T. Let t denote each time point, $t \in [0, T]$. The decision maker draws a random sample of the time points within $[0, T]$, and aggregate the utilities that could be obtained at each sampled time point, to construct a value representation of s. Her objective is to find a sampling strategy $f(t, s)$, which denotes the probability of selecting a time point t into the sample under a certain state s. I denote the prior probability of s occurring by $p(s)$ and the prior probability of selecting t into the sample by $p(t)$. The decision maker wants to maximize her total utility through f; thus, time points with larger rewards should be sampled more frequently (that is, she wants to selectively pay more attention on these time points). However, processing reward information and shifting attention trigger a cognitive cost. Let $u(.)$ denote the instantaneous utility function. Following the rational inattention literature (Matějka and McKay, 2015; Caplin et al., 2022; Maćkowiak et al., 2023), I frame the decision maker's optimization problem as

$$\max_{f} \int u(t,s)f(t,s)dtds - \lambda I(t;s)$$
$$s.t. \int f(t,s)dt = p(s), \forall s$$

where $\lambda I(t;s)$ is the attention cost function, λ is a fixed parameter and $I(t;s)$ denotes Shannon mutual information.

$$I(t;s) = \int f(t,s) \log \left(\frac{f(t,s)}{p(t)p(s)} \right) dtds$$

The solution needs to satisfy $f(t|s) \propto p(t)e^{u(t,s)/\lambda}$, where $\int f(t|s)dt = 1$.

Considering discrete time $t \in \{0, 1, ..., T\}$, we can replace $p(t)$ and $f(t|s)$ by the decision weight (or discounting factor) assigned to time point t. Let w_t^0 denote the initial weight assigned to t and w_t denote the decision weight after optimization, i.e. $w_t^0 \equiv p(t)$, $w_t \equiv f(t|s)$, then $\{w_t^0\}_{t=0}^T$ and $\{w_t\}_{t=0}^T$ can represent the initial and adjusted attention allocations. I assume the decision maker is initially time-consistent, that is, $w_t^0 = \delta^t$, where δ is the exponential discounting parameter, $\delta \in (0, 1]$.

1.2 SS or LL

In a "Money Earlier or Later" (MEL) task, the decision maker needs to make a choice between different options. For each option, she is informed the amount of reward she would receive in a specific time point t, which I denote by x_t. The reward is pre-determined, thus the state s is fixed. The decision maker simply wants to choose the option that has the maximum $\sum_t w_t u(x_t)$, where $w_t \propto \delta^t e^{u(x_t)/\lambda}$ and $\sum_{t=0}^T w_t = 1$. Suppose the reward only arrives at time point T, that is, $x_t = 0$ for all $t < T$. The optimal weight for T is

$$w_T = \frac{1}{1 + G(T) \cdot e^{-v(x_T)/\lambda}}$$

where

$$G(T) = \begin{cases} T & , \delta = 1 \\ \dfrac{1}{1-\delta}(\delta^{-T} - 1) & , 0 < \delta < 1 \end{cases}$$

When $\delta = 1$, I say the initial attention is uniformly allocated.

2 Data

To test the capacity of attention-adjusted discounting factor in explaining the findings of MEL experiments, I select two open datasets. The first dataset is from Ericson et al. (2015), containing 23,131 observations from 939 participants; the second is from Chávez et al. (2017), containing 34,515 choices from 1,284 participants. Hereafter, I term the first dataset as *Ericson* data, and the second dataset as *Chávez* data. Each dataset has been used in more than one previous study.[1] Readers interested in the empirical method and the results of this paper can easily compare them with those of other papers. For both datasets, the corresponding experiments require the participants to answer a series of choice questions. In each question, participants need to make a choice between an small sooner reward (denoted by SS) and a large later reward (denoted by LL). I denote the reward magnitude and delay by x_s and t_s for option SS, and by x_l and t_l for option LL, where $x_l > x_s > 0$, $t_l > t_s$.

Notably, there are significant differences in experimental procedure underlying the datasets. In *Ericson* data, the participants are recruited via Mechanical Turk. Rewards are framed in US dollars, ranging from 0.03 USD to 10,100 USD. 33% of the observations have a x_s no larger than 5 USD, 30% of the x_s are between 5 USD and 100 USD, 22% of the x_s are between 100 USD and 1,000 USD, while 15% are between 1,000 USD and 10,000 USD. Delays are framed in weeks. Delays for SS range from 0 to 2 weeks, and the maximum delay for LL is 5 weeks. In *Chávez* data, the participants are Mexican high school and first-year university students, and attending the experiment is a class requirement. Rewards are framed in Mexican pesos, ranging from 11 MXD (approx. 0.6 USD) to 85 (approx. 4.7 MXD). Delays are framed in days. All sooner rewards are delivered "today" ($t_s = 0$), and the maximum delay for LL is 186 days. Given that the two datasets employ different frames and ranges for rewards and delays, a same model fitted on these datasets may have different parameters and goodness of fit.

I mainly focus on out-of-sample model performance. For each dataset, I randomly draw the responses from 20% of the participants as the test sample, and set the rest as the train

[1]For example, *Ericson* data is used by Wulff and van den Bos (2018) for discussing the proper ways to compare different choice models. *Chávez* data is used by Gershman and Bhui (2020) for testing their proposed attention-based choice model.

sample. To mitigate the overfitting issue, I implement a 10-fold cross-validation procedure on the train sample.

3 Empirical Strategy

I test three types of intertemporal choice model: discounted utility model, trade-off model, and heuristic model.

The discounted utility model assumes that the decision maker tends to choose the option with greater discounted utility. Let the discounted utility for option j ($j \in \{l, s\}$) be $v_j = d(t_j)u(x_j)$, where $d(.)$ is the discounting function. Suppose the decision maker's perceived discounted utility for each option, denoted by is $\tilde{v_l}$ and $\tilde{v_s}$, is noisy. I set $\tilde{v_l} = v_l + \eta_l$, $\tilde{v_s} = v_s + \eta_s$. When η_l and η_s are independent noises and both follow $Gumble(0, \rho)$, where the scale parameter $\rho \in (0, \infty)$, the probability that the decision maker chooses LL is

$$P\{\tilde{v_s} \leq \tilde{v_l}\} = \frac{1}{1 + \exp\{-\frac{1}{\rho}(v_l - v_s)\}}$$

The trade-off model (Scholten and Read, 2010; Scholten et al., 2014) assumes that when thinking of whether to choose LL, the decision maker tends to make a comparison between attributes (reward and time), rather than between options (LL and SS). If the benefit of receiving a larger reward exceeds the cost of waiting longer, she will choose LL; otherwise, she will choose SS. Let B denote the benefit of receiving a larger reward, Q denote the cost of waiting longer. The value of B can be simply represented by $u(x_l) - u(x_s)$. Following Scholten et al. (2014), I represent Q by

$$Q = \frac{\kappa}{\zeta_1} \ln \left(1 + \zeta_1 \left(\frac{w(t_l) - w(t_s)}{\zeta_2} \right)^{\zeta_2} \right)$$

where $w(t) = \ln(1 + \omega t)/\omega$. The parameter ω measures the magnitude in which time is distorted in the decision maker's mind; κ measures the relative importance of reducing waiting time compared with increasing reward magnitude; ζ_1, ζ_2 jointly determine the curvature of

changes in Q relative to $t_l - t_s$.[2] I assume the decision maker's perception of B and Q, denoted by $\tilde{B}$ and $\tilde{Q}$, is noisy. $\tilde{B} = B + \eta_B$, $\tilde{Q} = Q + \eta_Q$, where η_B and η_Q are noise terms. Again, assume η_B and η_Q are independent and follow $Gumble(0, \rho)$, then the probability that the decision maker chooses LL is

$$P\{\tilde{Q} \leq \tilde{B}\} = \frac{1}{1 + \exp\{-\frac{1}{\rho}(B - Q)\}}$$

For the heuristic model, I employ a decision tree algorithm called XGBoost (Chen and Guestrin, 2016). The intuition underlying XGBoost is that, the decision-maker uses a chain of if-then rules to make a choice, and repeats this process for several times, adding up the results of each iteration to make the final decision. This algorithm has been widely used in solving classfication problems (including predicting human risky choices, see Plonsky et al. 2019). To better fit the data, I tune the hyper-parameters of this algorithm via grid search. For *Ericson* data, the features I use are x_s, x_l, t_s, t_l, the absolute and relative differences between x_s and x_l, the absolute and relative differences between t_l and t_s, the interest rate of LL when SS is invested as principal. For *Chávez* data, given that $t_s = 0$, I omit t_s and the differences between t_s and t_l from the features.[3]

The attention-adjusted discounting factor is dependent on the decision maker's initial attention allocation. I test the model under the assumptions that initial attention allocation is exponential and uniform (I term the former as "*attention*", the latter as "*attention_uni*"). Along with the attention-adjusted discounting, I employ 8 other methods to draw discounting factors, which are:

1. **exponential**, denoted by "*expo*"

$$d(t) = \delta^t$$

[2]Scholten et al. (2014) use these two parameters $\{\zeta_1, \zeta_2\}$ to ensure that Q follow a S-shape curve in relation to $t_l - t_s$, and that the decision-maker's behavioral pattern can shift between sub-additivity and super-additivity.

[3]The features are extracted following the methods in Read et al. (2013) and Ericson et al. (2015).

where the parameter is δ and $\delta \in (0, 1]$.

2. **double exponential**, denoted by "*expo2*" (van den Bos and McClure, 2013)

$$d(t) = \omega\delta_1^t + (1 - \omega)\delta_2^t$$

where the parameters are $\{\delta_1, \delta_2, \omega\}$, and $\delta_1, \delta_2 \in (0, 1]$.

3. **hyperbolic**, denoted by "*hb*"

$$d(t) = \frac{1}{1 + kt}$$

where the parameter is k.

4. **dual-parameter hyperbolic**, denoted by "*hb2*" (Loewenstein and Prelec, 1992)

$$d(t) = \frac{1}{(1 + kt)^a}$$

where the parameters are $\{k, a\}$.

5. **magnitude-dependent hyperbolic**, denoted by "*hbmd*" (Gershman and Bhui, 2020)

$$d(t) = \frac{1}{1 + kt}, \quad k = \frac{1}{bu(x_t)}$$

where the parameter is b.

6. **quasi-hyperbolic**, denoted by "*quasihb*"(Laibson, 1997)

$$d(t) = \mathbf{1}\{t = 0\} + \beta\delta^t \cdot \mathbf{1}\{t > 0\}$$

where the parameters are $\{\beta, \delta\}$, and $\beta, \delta \in (0, 1]$.

7. **quasi-hyperbolic plus fixed delay cost**, denoted by "*quasihb_fc*" (Benhabib et al., 2010)

$$d(t) = \mathbf{1}\{t = 0\} + (\beta\delta^t - \frac{c}{u(x_t)}) \cdot \mathbf{1}\{t > 0\}$$

where the parameters are $\{\beta, \delta, c\}$, and $\beta, \delta \in (0, 1]$.

8. **homogeneous costly empathy**, denoted by "*hce*" (Noor and Takeoka, 2022)

$$d_t = \kappa_t u(x_t)^{\frac{1}{m}}$$

where κ_t is decreasing in t. I set $\kappa_t = \delta^t$, where the parameters are $\{m, \delta\}$ and $\delta \in (0, 1]$.

Besides, I employ 2 types of utility functions to obtain $u(.)$: exponential or CARA utility, where $u(x) = 1 - e^{-\gamma x}$; power utility, where $u(x) = x^\gamma$. In each utility function, γ is the parameter, $\gamma \in (0, \infty)$. For parameters in discounting function, except for those explicitly marked as having a domain between 0 and 1, the domain of all other parameters is $(0, \infty)$. In model fitting, if a parameter has a lower bound of 0, I set its lower bound to 0.001; if a parameter has a upper bound of infinity, I set its upper bound to 100.

I use the maximum likelihood method to estimate the parameters, and apply L-BFGS-B method for optimization. As the solutions of L-BFGS-B are sensitive to initial points and often converge to local optima, I use the basin-hopping algorithm to achieve global optimization. [4] Finally, I compare the goodness of fit and out-of-sample performance of 20 discounted utility models, 2 trade-off models, and 1 heuristic model on the two datasets.

[4]The basin-hopping algorithm runs a local optimizer for several times. After each iteration, the solution randomly drifts to a new point. This new point is taken as the initial point for the next iteration. The algorithm compares the solution of the next iteration with the original solution, and is more likely to accept

4 Result

4.1 Results for *Ericson* data

Table 1 shows the goodness of fit for each model in cross-validation. The heuristic model has the highest accuracy rate, the lowest log loss, and the lowest MAE. The trade-off (*trade*) model with power utility performs the lowest MSE. The magnitude-dependent hyperbolic (*hbmd*) model with power utility ranks the second or third in all evaluation metrics. On the test sample, these three models also perform the best in MSE, MAE, log loss and accuracy rate (see Table 2). To test the correlation between these models, I randomly draw 1,000 choice questions from *Ericson* data, and set the choices predicted by the heuristic model as labels, letting the other models to predict them. The accuracy rate for *trade* with power utility is the highest, which is 95.7%, and the second is *hbmd* with power utility (95.0%).

Table 1: Cross-Validation Results on Ericson Data

model	**utility**	**mse**	**mae**	**log_loss**	**accuracy**
heuristic	–	0.298	0.298	0.581	0.702
trade	power	0.204	0.407	0.595	0.693
hbmd	power	0.206	0.413	0.602	0.692
quasihb_fc	power	0.208	0.418	0.606	0.685
quasihb	power	0.209	0.417	0.607	0.687
expo2	power	0.210	0.420	0.609	0.685
attention_uni	power	0.211	0.422	0.611	0.679
hb2	power	0.211	0.422	0.611	0.682
hb	power	0.211	0.422	0.612	0.681
expo	power	0.211	0.422	0.612	0.681
hce	power	0.211	0.422	0.612	0.681

the better solution between them (note there is still some probability of accepting an inferior solution). The magnitude of drifting is dependent on a stepwise parameter, which I set as 0.5; the probability of accepting the inferior solution is dependent on a temper parameter, which I set as 1.0. I also set the maximum number of iterations as 500.

Table 1: Cross-Validation Results on Ericson Data

model	**utility**	**mse**	**mae**	**log_loss**	**accuracy**
attention	power	0.215	0.431	0.621	0.673
trade	cara	0.218	0.437	0.628	0.666
attention	cara	0.228	0.456	0.648	0.638
attention_uni	cara	0.229	0.458	0.650	0.631
hbmd	cara	0.229	0.458	0.650	0.631
quasihb	cara	0.229	0.459	0.651	0.630
quasihb_fc	cara	0.229	0.459	0.651	0.630
expo2	cara	0.229	0.458	0.651	0.630
expo	cara	0.229	0.459	0.651	0.629
hce	cara	0.229	0.459	0.651	0.629
hb2	cara	0.229	0.459	0.651	0.629
hb	cara	0.229	0.459	0.651	0.629

Note: **mae** denotes mean absolute error, **mse** denotes mean squared error

Table 2: Out-of-Sample Test Results on *Ericson* Data

model	**utility**	**mse**	**mae**	**log_loss**	**accuracy**	**pred_ll**
heurstic	–	0.200	0.400	0.586	0.702	0.290
hbmd	power	0.204	0.411	0.596	0.696	0.237
trade	power	0.202	0.405	0.591	0.695	0.249
quasihb_fc	power	0.207	0.412	0.603	0.694	0.248
quasihb	power	0.207	0.422	0.604	0.694	0.248
expo	power	0.210	0.427	0.609	0.690	0.306
hce	power	0.210	0.428	0.609	0.686	0.260
hb	power	0.209	0.424	0.607	0.686	0.260
attention_uni	power	0.211	0.422	0.611	0.678	0.157

Table 2: Out-of-Sample Test Results on *Ericson* Data

model	**utility**	**mse**	**mae**	**log_loss**	**accuracy**	**pred_ll**
attention	power	0.215	0.431	0.623	0.673	0.142
trade	cara	0.217	0.434	0.626	0.668	0.120
attention	cara	0.231	0.457	0.654	0.629	0.091
attention_uni	cara	0.231	0.459	0.654	0.624	0.087
hbmd	cara	0.231	0.459	0.654	0.623	0.085
hb2	cara	0.233	0.454	0.658	0.620	0.046
hb	cara	0.231	0.460	0.655	0.618	0.082
hce	cara	0.231	0.460	0.655	0.618	0.080
quasihb	cara	0.231	0.460	0.655	0.618	0.078
quasihb_fc	cara	0.231	0.460	0.655	0.618	0.078
expo	cara	0.231	0.460	0.655	0.618	0.080
expo2	cara	0.240	0.447	0.679	0.616	0.013
expo2	power	0.385	0.386	4.767	0.614	0.000
hb2	power	0.386	0.386	6.562	0.614	0.000

Note: **mae** denotes mean absolute error, **mse** denotes mean squared error, **pred_ll** denotes the ratio of LL in predicted choices.

Now we focus on attention-adjusted models. First, it is notable that *hbmd* can be viewed as a special case of the attention-adjusted models. [5] Magnitude-dependent hyperbolic discounting plus power utility is identical to attention-adjusted discounting under uniform initial attention allocation plus log utility. We can conclude that log utility function fits the data better than power and CARA functions under attention-adjusted model settings.

Second, the reason why attention-adjusted models with power or CARA utility underpeform than some other models can be (at least partially) attributed to the distribution of observations on *Ericson* data. Note that the rewards for SS follow a power-law-like distribution in

[5]Note that under the assumption that the initial attention is uniformly allocated, the discounting factor in attention-adjusted model is $1/(1+kt)$, where $k = e^{-u(x_t)/\lambda}$. Setting $u(x)/\lambda = \ln\beta + \gamma \ln x$, we can get the magnitude-dependent hyperbolic model with power utility.

Ericson data. The data contains a substantial proportion of very small rewards, compared with the reward range: 33% of the x_s are no larger than 5 USD, whereas only 15% are between 1,000 USD and 10,000 USD. The derivatives of exponential and power functions (with a positive power) at a small number are much smaller than that of a log function. Thus, for a small reward, neither CARA utility nor power utility of it can significantly differ from the utility of its near neighbors, and this makes it more difficult for a model to learn the decision makers' choices on the given data.

Third, the last column in Table 2, which reports the ratio of LL in predicted choices, also gives us a hint on why attention-adjusted models underperform. This kind of models seem underestimating the participants' tendency to choose LL in *Ericson* data. For example, *attention_uni* with power utility predicts only 15.7% of the choices are LL, whereas the heuristic model and *hbmd* with power utility predict 29% and 23.7% of the choices are LL respectively. To some extent, this can also be attributed to the fact that there are many small rewards in the data. Considering *hbmd* and *attention_uni* models, the discounting factor of each can be formatted as $1/(1+kt)$, where $k = e^{-u(x)/\lambda}$ for *attention_uni* and $k = (\beta u(x))^{-1}$ for *hbmd*. When x approaches to 0, under both CARA and power utility settings, the k in *attention_uni* approaches to 1, while the k in *hbmd* approaches to infinity. Thus, the utility of a small reward under *hbmd* may be discounted more heavily than that under *attention_uni*. In such a case, suppose a decision maker still prefers a delayed reward to an immediate reward, the *hbmd* model should conclude the decision maker is more patient in comparison to *attention_uni*.

To check if sample distribution affects the performance of models, I conduct two follow-up analyses on *Ericson* data. Table 3 shows the performance of the seven models with the highest out-of-sample prediction accuracy for each follow-up analysis. In the first analysis, I set 0.01 USD as reward unit, and replace x_s and x_l by their logarithms (keeping other features equal)[6], then re-train and test the models. The results for the first analysis are presented in Panel A of Table 3. The highest accuracy is for the heuristic model, and its accuracy rate rises from 70.2% in 2 to 71.1%. The accuracy of *attention_uni* model with

[6]The distributions for log magnitude of rewards are similar to normal distributions.

power utility rises from 61.1% in 2 to 70.8%. In the second analysis, I omit the observations with $x_s \leq 5$ USD, retaining 66% of the observations for model fitting. The results for the second analysis are presented in Panel B of Table 3. The highest accuracy is for the heuristic model, and its accuracy rate rises from to 74.9%. The accuracy of *attention_uni* model with power utility rises to 74.8%.

In each analysis, there is an increase in the performance of attention-adjusted models relative to other models. Meanwhile, the attention-adjusted models with power utility outperform other models except the heuristic model. I also set the predicted choices from heuristic models as the labels and apply other fitted models to predicting it. In the first analysis, *attention* and *attention_uni* with power utility perform the best in prediction, which have an accuracy of 93.2% and 92.4%. In the second analysis, the models that predict the best are *quasihb_fc* and *quasihb* with power utility, with an accuracy of 94.2% and 93.3%. The *attention_uni* model with power utility has an accuracy of 91.1%. Therefore, when reward magnitude is much greater than 0 and only few extreme values are contained in the data, the attention-adjusted models seem a good substitute to the heuristic model for researchers who want to save parameters.

Table 3: Out-of-Sample Test Results for Follow-Up Analyses

model	**utility**	**mse**	**mae**	**log_loss**	**accuracy**	**pred_ll**
Panel A						
heurstic	–	0.194	0.394	0.574	0.711	0.265
attention	power	0.201	0.407	0.590	0.709	0.234
attention_uni	power	0.201	0.408	0.590	0.708	0.231
trade	power	0.198	0.402	0.583	0.701	0.231
hbmd	power	0.203	0.413	0.595	0.701	0.295
trade	cara	0.202	0.407	0.592	0.698	0.269
attention_uni	cara	0.205	0.412	0.597	0.691	0.161
Panel B						
heurstic	–	0.177	0.368	0.535	0.749	0.250

Table 3: Out-of-Sample Test Results for Follow-Up Analyses

model	**utility**	**mse**	**mae**	**log_loss**	**accuracy**	**pred_ll**
attention_uni	power	0.185	0.384	0.556	0.748	0.213
quasihb	power	0.191	0.400	0.568	0.748	0.234
expo	power	0.187	0.380	0.560	0.747	0.256
hb	power	0.189	0.398	0.565	0.747	0.256
hbmd	power	0.185	0.384	0.555	0.746	0.217
trade	power	0.182	0.382	0.549	0.746	0.216

Note: *Panel A* shows the results for using 0.01 USD as reward unit and taking log magnitude of rewards. *Panel B* shows the results for omitting the observations in which the rewards for SS are no larger than 5 USD. Each panel present the seven models with the highest accuracy rate. **mae** denotes mean absolute error, **mse** denotes mean squared error, **pred_ll** denotes the ratio of LL in predicted choices.

4.2 Results for *Chávez* data

For *Chávez* data, I use MXD as reward unit and days of delay as t_l. Note the rewards range from 11 MXD to 85 MXD, and for either SS or LL, the distribution of rewards approximates a uniform distribution. Therefore, Table 4 shows the goodness of fit for each model in cross-validation. Similar with Table 1, the *trade* model with power utility performs the lowest MSE, and the heuristic model performs the best in MAE, log loss, and accuracy. Though, the goodness of fit of attention-adjusted models with power utility are close to the *trade* model in all evaluation metrics. For example, the log loss for *attention* with power utility (0.486) is only 0.003 higher than that for *trade* with power utility (0.483). Table 5 shows the out-of-sample performance of each model. The heuristic model performs the best in all evaluation metrics, following *trade*, *attention*, and *attention_uni* with power utility. The performance between these models are also close. For example, the accuracy rate of heuristic model is 76.7%, while those of *trade*, *attention*, and *attention_uni* with power utility are 76.6%, 76.3%, 76.3%. Setting the predicted choices by heuristic model as the labels and letting the other models to predicting them, we obtain an accuracy of 96.2% for *trade* with

power utililty (the highest), and an accuracy of 95.9% for attention-adjusted models with power utility (the second highest).

In summary, I find the heuristic model outperforms than the other models in predicting human choices on Ericson and *Chávez* data. Both Intertemporal trade-off and attention adjustment (given that *hbmd* is a special case of attention-adjusted model) can be viewed as good candidates of mechansim for explaining the choices predicted by the heuristic model. Remarkably, apart from the heuristic model, the *trade* model has the largest number of parameters, which has 6. The *attention_uni* and *attention* model have only 3 and 4 parameters to be fitted, and the minimum number of parameters among the fitted models is 3. Thus, for people who want to save parameters or draw the predictions on a normative theory of intertemporal choice, the attention adjustment mechanism can be a choice.

Table 4: Cross-Validation Results on Chávez Data

model	**utility**	**mse**	**mae**	**log_loss**	**accuracy**
heuristic	–	0.221	0.221	0.480	0.779
trade	power	0.157	0.314	0.483	0.782
hb2	power	0.158	0.315	0.485	0.775
expo2	power	0.158	0.315	0.485	0.780
attention	power	0.158	0.317	0.486	0.782
hbmd	power	0.158	0.317	0.486	0.774
quasihb_fc	power	0.158	0.317	0.487	0.781
trade	cara	0.159	0.317	0.487	0.772
attention_uni	power	0.158	0.318	0.488	0.782
hb	power	0.159	0.320	0.489	0.774
hbmd	cara	0.159	0.319	0.489	0.774
quasihb	power	0.159	0.321	0.491	0.777
expo	power	0.160	0.322	0.491	0.774
hce	power	0.160	0.322	0.491	0.774
attention	cara	0.162	0.324	0.494	0.746

Table 4: Cross-Validation Results on Chávez Data

model	**utility**	**mse**	**mae**	**log_loss**	**accuracy**
attention_uni	cara	0.166	0.331	0.501	0.730
hb	cara	0.165	0.331	0.503	0.747
quasihb	cara	0.167	0.332	0.506	0.741
quasihb_fc	cara	0.168	0.337	0.509	0.732
hce	cara	0.169	0.339	0.511	0.721
expo	cara	0.172	0.345	0.516	0.701
hb2	cara	0.174	0.348	0.519	0.689
expo2	cara	0.175	0.348	0.522	0.683

Note: **mae** denotes mean absolute error, **mse** denotes mean squared error

Table 5: Out-of-Sample Test Results on Chávez Data

model	**utility**	**mse**	**mae**	**log_loss**	**accuracy**	**pred_ll**
heurstic	–	0.163	0.322	0.500	0.767	0.295
trade	power	0.165	0.316	0.504	0.766	0.258
attention	power	0.165	0.326	0.505	0.763	0.332
attention_uni	power	0.165	0.325	0.506	0.763	0.332
hb	power	0.167	0.327	0.508	0.757	0.332
quasihb_fc	power	0.206	0.448	0.604	0.757	0.332
quasihb	power	0.180	0.397	0.543	0.757	0.332
hce	power	0.167	0.335	0.510	0.757	0.332
hbmd	power	0.165	0.327	0.505	0.757	0.332
hbmd	cara	0.167	0.329	0.509	0.757	0.332
hb2	power	0.167	0.341	0.509	0.757	0.332
expo	power	0.167	0.334	0.510	0.757	0.332
trade	cara	0.248	0.498	0.689	0.700	0.222

Table 5: Out-of-Sample Test Results on Chávez Data

model	**utility**	**mse**	**mae**	**log_loss**	**accuracy**	**pred_ll**
attention	cara	0.179	0.356	0.535	0.685	0.037
attention_uni	cara	0.191	0.414	0.567	0.685	0.037
hb2	cara	0.184	0.344	0.545	0.667	0.000
hb	cara	0.207	0.322	0.622	0.667	0.000
expo2	power	0.333	0.333	3.764	0.667	0.000
expo2	cara	0.326	0.333	1.605	0.667	0.000
hce	cara	0.202	0.316	0.634	0.667	0.000
quasihb	cara	0.222	0.333	0.650	0.667	0.000
expo	cara	0.203	0.314	0.653	0.667	0.000
quasihb_fc	cara	0.198	0.336	0.582	0.667	0.000

Note: **mae** denotes mean absolute error, **mse** denotes mean squared error, **pred_ll** denotes the ratio of LL in predicted choices.

Reference

Benhabib, J., Bisin, A., and Schotter, A. (2010). Present-bias, quasi-hyperbolic discounting, and fixed costs. *Games and Economic Behavior*, 69(2):205–223.

Caplin, A., Dean, M., and Leahy, J. (2022). Rationally Inattentive Behavior: Characterizing and Generalizing Shannon Entropy. *Journal of Political Economy*, 130(6):1676–1715.

Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, San Francisco California USA. ACM.

Chávez, M. E., Villalobos, E., Baroja, J. L., and Bouzas, A. (2017). Hierarchical Bayesian modeling of intertemporal choice. *Judgment and Decision Making*, 12(1):19–28. Publisher: Cambridge University Press.

Ericson, K. M., White, J. M., Laibson, D., and Cohen, J. D. (2015). Money Earlier or Later? Simple Heuristics Explain Intertemporal Choices Better Than Delay Discounting Does. *Psychological Science*, 26(6):826–833.

Gershman, S. J. and Bhui, R. (2020). Rationally inattentive intertemporal choice. *Nature Communications*, 11(1):3365. Number: 1 Publisher: Nature Publishing Group.

Laibson, D. (1997). Golden Eggs and Hyperbolic Discounting. *The Quarterly Journal of Economics*, 112(2):443–478.

Loewenstein, G. and Prelec, D. (1992). Anomalies in Intertemporal Choice: Evidence and an Interpretation. *The Quarterly Journal of Economics*, 107(2):573–597.

Matějka, F. and McKay, A. (2015). Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model. *American Economic Review*, 105(1):272–298.

Maćkowiak, B., Matějka, F., and Wiederholt, M. (2023). Rational Inattention: A Review. *Journal of Economic Literature*, 61(1):226–273.

Noor, J. and Takeoka, N. (2022). Optimal Discounting. *Econometrica*, 90(2):585–623.

Plonsky, O., Apel, R., Ert, E., Tennenholtz, M., Bourgin, D., Peterson, J. C., Reichman, D., Griffiths, T. L., Russell, S. J., Carter, E. C., Cavanagh, J. F., and Erev, I. (2019). Predicting human decisions with behavioral theories and machine learning. Publisher: arXiv Version Number: 1.

Read, D., Frederick, S., and Scholten, M. (2013). DRIFT: An analysis of outcome framing in intertemporal choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39(2):573–588.

Scholten, M. and Read, D. (2010). The Psychology of Intertemporal Tradeoffs. *Psychological Review*, 117(3):925–944.

Scholten, M., Read, D., and Sanborn, A. (2014). Weighing Outcomes *by* Time or *Against* Time? Evaluation Rules in Intertemporal Choice. *Cognitive Science*, 38(3):399–438.

van den Bos, W. and McClure, S. M. (2013). Towards a General Model of Temporal Discounting: General Model of Time Discounting. *Journal of the Experimental Analysis of Behavior*, 99(1):58–73.

Wulff, D. U. and van den Bos, W. (2018). Modeling Choices in Delay Discounting. *Psychological Science*, 29(11):1890–1894.