Online Appendix

In this online appendix, we collect the materials omitted from the main text of the paper. The appendices are ordered according to where they are first referenced in the main text (Fang and Wang, 2014). In Appendix A we provide a table that summarizes the main notations in the paper for the convenience of the readers; in Appendix B we provide the details for a couple of intermediate steps in the derivation of Eq. (11); in Appendix C we provide the details of the omitted proof for Proposition 2; in Appendix D we provide some intuition for How β, $\tilde{\beta}$ and δ are distinguished in short-panel data when exclusion variables are available; in Appendix E we describe how we can replicate the proof of Proposition 2 in Magnac and Thesmar (2002) as a special case of our analysis when $\beta = \tilde{\beta} = 1$; in Appendix F we derive the asymptotic distribution of the estimator; in Appendix G, we report the estimation results without a priori restrictions that $\beta \in [0,1]$ and $\tilde{\beta} \in [\beta, 1]$; and in Appendix H we discuss the extension to the context of finite horizon models with non-stationary state transitions.

A Summary of Notations

Table A1 summarizes the main notations in the paper for the convenience of the readers.

Notation	Interpretation	Equation/Definition
$u_i(x)$	Deterministic payoff from choosing alternative i when state vector is x	
$u_i^*(x)$	Payoff, including the choice-specific shock, from choosing alternative i when state vector is $x : u_i^*(x) \equiv u_i(x) + \varepsilon_i$	Eq. (1)
β	The present-bias factor	Definition (1)
$\tilde{\beta}$	The partial naivety parameter	
δ	The standard discount factor	Definition (1)
$V_i(x)$	Perceived choice-specific long-run value function	Eq. (8)
$V(x)$	Perceived long-run value function	Eq. (4)
$W_i(x)$	Current choice-specific value function	Eq. (3)
$Z_i(x)$	Choice-specific value function of the next-period self as perceived by the current self	Eq. (5)
$\tilde{\sigma}(x, \boldsymbol{\varepsilon})$	Perceived continuation strategy profile for a partially naive agent	Definition (2)
$\sigma^*(x, \boldsymbol{\varepsilon})$	Perception-perfect strategy profile for a partially naive agent	Definition (3)

Table A1: Summary of Key Notations.

B Derivation of Eq. (11): Details

Here we provide some intermediate steps in the derivation of Eq. (11):

$$\begin{aligned}
V(x) &= \mathrm{E}_{\boldsymbol{\varepsilon}}\left[V_{\tilde{\sigma}(x,\boldsymbol{\varepsilon})}(x)+\varepsilon_{\tilde{\sigma}(x,\boldsymbol{\varepsilon})}\right] \\
&= \mathrm{E}_{\boldsymbol{\varepsilon}}\left[Z_{\tilde{\sigma}(x,\boldsymbol{\varepsilon})}(x)+\varepsilon_{\tilde{\sigma}(x,\boldsymbol{\varepsilon})}+\left(1-\tilde{\beta}\right)\delta\sum_{x'\in\mathcal{X}}V(x')\pi(x'|x,\tilde{\sigma}(x,\boldsymbol{\varepsilon}))\right] \\
&= \mathrm{E}_{\boldsymbol{\varepsilon}}\max_{i\in\mathcal{I}}\left[Z_i(x)+\varepsilon_i\right]+\left(1-\tilde{\beta}\right)\delta\mathrm{E}_{\boldsymbol{\varepsilon}}\sum_{x'\in\mathcal{X}}V(x')\pi(x'|x,\tilde{\sigma}(x,\boldsymbol{\varepsilon})) \\
&= \mathrm{E}_{\boldsymbol{\varepsilon}}\max_{i\in\mathcal{I}}\left[Z_i(x)+\varepsilon_i\right]+\left(1-\tilde{\beta}\right)\delta\sum_{j\in\mathcal{I}}\left[\tilde{P}_j(x)\sum_{x'\in\mathcal{X}}V(x')\pi(x'|x,j)\right]
\end{aligned}$$

where the first equality is just copying (9); the second equality follows from (10); and the third equality follows from (6) and thus

$$\mathrm{E}_{\boldsymbol{\varepsilon}}\left[Z_{\tilde{\sigma}(x,\boldsymbol{\varepsilon})}(x)+\varepsilon_{\tilde{\sigma}(x,\boldsymbol{\varepsilon})}\right]=\mathrm{E}_{\boldsymbol{\varepsilon}}\max_{i\in\mathcal{I}}\left[Z_i(x)+\varepsilon_i\right];$$

and the fourth equality follows from the fact that

$$\begin{aligned}
\mathrm{E}_{\boldsymbol{\varepsilon}}\sum_{x'\in\mathcal{X}}V(x')\pi(x'|x,\tilde{\sigma}(x,\boldsymbol{\varepsilon})) &= \sum_{j\in\mathcal{I}}\left[\Pr\left(\tilde{\sigma}(x,\boldsymbol{\varepsilon})=j\right)\sum_{x'\in\mathcal{X}}V(x')\pi(x'|x,j)\right] \\
&= \sum_{j\in\mathcal{I}}\left[\tilde{P}_j(x)\sum_{x'\in\mathcal{X}}V(x')\pi\left(x'|x,j\right)\right].
\end{aligned}$$

C Proof of Proposition 2

Proof. If the data set is generated by the assumed data generating process for some primitives $\left(\mathbf{u}^*,\beta^*,\tilde{\beta}^*,\delta^*\right)$, the observed choice probabilities $\tilde{\mathbf{P}}$ and transition probabilities $\boldsymbol{\Pi}$ must satisfy the system of $I\times X$ nonlinear equations defined by (25). When there are variables that satisfy the exclusion restrictions, the data must also satisfy the additional $I\times|\mathcal{X}_e|\times|\mathcal{X}_r|$ equations requiring that $u_i(x_r,x_e)=u_i(x_r)$ for each $i\in\mathcal{I}\backslash\{0\}$, each $x_e\in\mathcal{X}_e$ and each $x_r\in\mathcal{X}_r$. Denote the augmented system of equations as :

$$\widetilde{\mathcal{G}}\left(\mathbf{u},\beta,\tilde{\beta},\delta;\left(\tilde{\mathbf{P}},\boldsymbol{\Pi}\right)\right)=0, \tag{C1}$$

where $\left(\tilde{\mathbf{P}},\boldsymbol{\Pi}\right)$ indicates the data set. This new system of equations has $I\times X+I\times|\mathcal{X}_e|\times|\mathcal{X}_r|$ equations in $I\times X+3$ unknowns $\left(\mathbf{u},\beta,\tilde{\beta},\delta\right)$.

We now show that our results follow from the Transversality Theorem (see Proposition 8.3.1, Mas-Colell, 1985):

Theorem 1 (Transversality Theorem, Proposition 8.3.1 in Mas-Colell, 1985.). *Let $F: N\times B\to R^m, N\subset R^n, B\subset R^s$ be C^r with $r>\max\{n-m,0\}$. Suppose that 0 is a regular value of F; that is, $F(x,b)=0$ implies rank $\partial F(x,b)=m$. Then, except for a set of $b\in B$ of Lebesgue measure zero, $F_b:N\to R^m$ has 0 as a regular value.*

It is useful to be explicit about the mapping between notations in the Transversality Theorem stated above and the corresponding terms in our application:

Transversality Theorem	This Paper
$F\left(x,b\right)$	$\tilde{\mathcal{G}}\left(\mathbf{u},\beta,\tilde{\beta},\delta;\left(\tilde{\mathbf{P}},\mathbf{\Pi}\right)\right)$
$x\in N\subset R^{n}$	$\left(\mathbf{u},\beta,\tilde{\beta},\delta\right)\in R^{I\times X}\times(0,1]^{3}\subset R^{I\times X+3}$
n	Number of unknowns: $I\times X+3$
$b\in B\subset R^{s}$	$\left(\tilde{\mathbf{P}},\mathbf{\Pi}\right)\in\Delta^{(I+1)X}\times\left(\overbrace{\Delta^{X}\times...\times\Delta^{X}}^{X\text{ copies}}\right)^{I+1}\subset R^{IX+(I+1)X(X-1)}$
s	$IX+\left(I+1\right)X\left(X-1\right)$
m	Number of equations in $\tilde{\mathcal{G}}$: $I\times X+I\times\lvert\mathcal{X}_{e}\rvert\times\lvert\mathcal{X}_{r}\rvert$
$F_{b}:N\rightarrow R^{m}$	$\tilde{\mathcal{G}}\left(\cdot;\left(\tilde{\mathbf{P}},\mathbf{\Pi}\right)\right):R^{I\times X}\times(0,1]^{3}\rightarrow R^{I\times X+I\times\lvert\mathcal{X}_{e}\rvert\times\lvert\mathcal{X}_{r}\rvert}$

Note that in the last line of the above table, $\tilde{\mathcal{G}}\left(\cdot;\left(\tilde{\mathbf{P}},\mathbf{\Pi}\right)\right)$ is simply the system $\tilde{\mathcal{G}}\left(\mathbf{u},\beta,\tilde{\beta},\delta;\left(\tilde{\mathbf{P}},\mathbf{\Pi}\right)\right)$ but with $\left(\tilde{\mathbf{P}},\mathbf{\Pi}\right)$ considered as parameters instead of arguments.

First, all functions in our system $\tilde{\mathcal{G}}$ are continuously differentiable, so the first requirement that $F\in C^{r}$ is satisfied trivially. Second, we need to check that "0 is a regular value of F; that is, $F\left(x,b\right)=0$ implies rank $\partial F\left(x,b\right)=m$." It is important to note that in writing out the $\partial F\left(x,b\right)$, we need to take derivatives of F with respect to both x and b. In our application, $\partial\tilde{\mathcal{G}}\left(\mathbf{u},\beta,\tilde{\beta},\delta;\left(\tilde{\mathbf{P}},\mathbf{\Pi}\right)\right)$ is thus a matrix of dimensions $\left\{\overbrace{I\times X+I\times|\mathcal{X}_{e}|\times|\mathcal{X}_{r}|}^{m}\right\}\times\left\{\overbrace{\left(I\times X+3\right)}^{n}+\overbrace{IX+\left(I+1\right)X\left(X-1\right)}^{s}\right\}$. Note that $|\mathcal{X}_{e}|+|\mathcal{X}_{r}|=X$, thus the number of rows in $\partial\tilde{\mathcal{G}}\left(\mathbf{u},\beta,\tilde{\beta},\delta;\left(\tilde{\mathbf{P}},\mathbf{\Pi}\right)\right)$, which is $\overbrace{I\times X+I\times|\mathcal{X}_{e}|\times|\mathcal{X}_{r}|}^{m}$, is smaller than (and potentially a lot smaller than) the number of columns $\overbrace{\left(I\times X+3\right)}^{n}+\overbrace{IX+\left(I+1\right)X\left(X-1\right)}^{s}$. The "beauty" of the Transversality Theorem is that we only need to check that $\partial\tilde{\mathcal{G}}\left(\mathbf{u},\beta,\tilde{\beta},\delta;\left(\tilde{\mathbf{P}},\mathbf{\Pi}\right)\right)$ has a rank of m (the smaller number) whenever $\tilde{\mathcal{G}}\left(\mathbf{u},\beta,\tilde{\beta},\delta;\left(\tilde{\mathbf{P}},\mathbf{\Pi}\right)\right)=0$. This can be verified in the same way that we verify that $\partial_{\mathbf{u}}\mathcal{G}(\mathbf{u}^{*};\cdot)$ has full rank whenever $\mathcal{G}(\mathbf{u}^{*};\cdot)=0$.

Given these, we can appeal to the Transversality Theorem and conclude that "except for a set of $b\in B$ of Lebesgue measure zero, $F_{b}:N\rightarrow R^{m}$ has 0 as a regular value", which in our application means that, "except for a set of $\left(\tilde{\mathbf{P}},\mathbf{\Pi}\right)\in\Delta^{(I+1)X}\times\left(\overbrace{\Delta^{X}\times...\times\Delta^{X}}^{X\text{ copies}}\right)^{I+1}\subset R^{IX+(I+1)X(X-1)}$ with Lebesgue measure 0, $\tilde{\mathcal{G}}\left(\cdot;\left(\tilde{\mathbf{P}},\mathbf{\Pi}\right)\right)$ has 0 as a regular value."

However, "$\tilde{\mathcal{G}}\left(\cdot;\left(\tilde{\mathbf{P}},\mathbf{\Pi}\right)\right)$ has 0 as a regular value" means that, whenever $\tilde{\mathcal{G}}\left(\mathbf{u},\beta,\tilde{\beta},\delta\right)=0$,[1] $\partial_{\left(\mathbf{u},\beta,\tilde{\beta},\delta\right)}\tilde{\mathcal{G}}\left(\mathbf{u},\beta,\tilde{\beta},\delta\right)$ must have rank m (the number of equations). But this is impossible because in our application m is equal to $I\times X+I\times|\mathcal{X}_{e}|\times|\mathcal{X}_{r}|$, which is larger than the number of unknowns $I\times X+3$ under our identifying assumption that $I\times|\mathcal{X}_{e}|\times|\mathcal{X}_{r}|\geq4$. Therefore, generically $\tilde{\mathcal{G}}\left(\mathbf{u},\beta,\tilde{\beta},\delta;\left(\tilde{\mathbf{P}},\mathbf{\Pi}\right)\right)=0$ should have no solution except the true primitives $\left(\mathbf{u}^{*},\beta^{*},\tilde{\beta}^{*},\delta^{*}\right)$ that generated the data. □

[1] Here we supress the notation of $\left(\tilde{\mathbf{P}},\mathbf{\Pi}\right)$ to clarify that now we are only counting $\left(\mathbf{u},\beta,\tilde{\beta},\delta\right)$ as unknowns, and $\left(\tilde{\mathbf{P}},\mathbf{\Pi}\right)$ are simply parameters.

D Intuition for How β, $\tilde{\beta}$ and δ are Distinguished in Short Panel Data

Here we provide some intuition as to how $\left\langle \beta, \tilde{\beta}, \delta \right\rangle$ come to differentially affect the observed choice behavior of the current self depending on the values of the exclusion variables.[2]

Distinguishing β and δ. It may seem counter-intuitive that β and δ could be separately identified in a short two-period panel data set. To provide some intuition, let us consider the case that $\tilde{\beta} = \beta$. The question is: "Can we distinguish the behavior of an agent with exponential discounting rate $\hat{\delta} = \beta\delta$ from the behavior of a sophisticated time-inconsistent agent with preference (β, δ)?" Under stationarity assumption, if an agent has time consistent exponential discounting rate $\hat{\delta} = \beta\delta$, her expected continuation utilities are completely determined by the observed choice probabilities. To see this, observe that in equation (22), if one replaces $\tilde{\beta}$ by 1 and δ by $\hat{\delta}$, we will have

$$V(x) = Z_0(x) + \ln \left\{ \sum_{i \in \mathcal{I}} \exp\left[Z_i(x) - Z_0(x) \right] \right\},$$

which only depends on $D_i(x)$ when $\beta = \tilde{\beta}$.

However, for a sophisticated time-inconsistent agent with preference (β, δ), there is an incongruence between current self and her perceived future self regarding how they evaluate the future stream of payoffs. Though the current self has to defer to her next-period self in terms of the actual next-period choice that will be chosen, they disagree on how much weight to put on payoffs two-periods from now. It is this incongruence that leads to the last term in Eq. (22), which in turn breaks the tight link between observed choice probabilities and the continuation utilities.

As we demonstrated in Section 3.2.1, the continuation utilities will ultimately determine the identified values of $\langle \{ u_i(x) : i \in \mathcal{I} \backslash \{0\}, x \in \mathcal{X} \} \rangle$. Thus β and δ can be distinguished when there are exclusion variables because they require that $u_i(x)$ does not depend on the x_e components of the state vector.

Distinguishing β and $\tilde{\beta}$. To help provide intuition for why β could be distinguished from $\tilde{\beta}$, let us suppose that $\delta = 1$. First note that the ratio $\tilde{\beta}/\beta$ appears in term $Z_i(x)$ [see Eq. (20)]. This ratio regulates the incongruence between the current self's own behavior and her perception of the behavior of her future selves. Eq. (21) shows that if $\tilde{\beta}/\beta = 1$, then $Z_i(x) - Z_0(x)$ is uniquely determined by the observed $D_i(x)$; and thus the current self's perception about her future self's action is identical to her own action. This implies that $\tilde{\beta}/\beta = 1$ (and $\delta = 1$) will pin down completely the identified values of $\langle \{ u_i(x) : i \in \mathcal{I} \backslash \{0\}, x \in \mathcal{X} \} \rangle$ from the data, which could be refuted if the identified values of $u_i(x)$ do not satisfy the exclusion restrictions, i.e., $u_i(x)$ should not depend on the x_e components of the state vector.

E Exponential Discounting Special Case: $\beta = \tilde{\beta} = 1$

We here show that the non-linear equation system (21) and (24), for the special case of dynamic discrete choice models with exponential discounting (i.e., the case with $\tilde{\beta} = \beta = 1$), replicates the known results in the literature. In that case, (21) is reduced to the well-known relationship

$$Z_i(x) - Z_0(x) = \ln P_i(x) - \ln P_0(x); \tag{E2}$$

[2] In some sense, the exclusion state variables play the role of time delay in the experimental literature cited in Footnote 2 which relies on preference reversal to identify time inconsistent preferences.

that is, in standard models the difference in the choice probabilities for alternative i and the reference alternative 0 informs us about the difference in the value from choosing i relative to the value from choosing 0. This is of course also true when $\tilde{\beta} = \beta < 1$. The potential naivety we allow in our setup breaks this direct relationship between $P_i(x)/P_0(x)$ and $Z_i(x) - Z_0(x)$.

Moreover, when $\tilde{\beta} = \beta = 1$, (24) for $i = 0$ is reduced to (using the normalization that $u_0(x) = 0$) :

$$Z_0(x) = \delta \sum_{x' \in \mathcal{X}} Z_0(x') \pi(x'|x,0) + \delta \sum_{x' \in \mathcal{X}} \ln \left[\sum_{i \in \mathcal{I}} \frac{P_i(x')}{P_0(x')} \right] \pi(x'|x,0).$$

For simplicity, denote the $X \times 1$ vector $\{Z_0(x)\}_{x \in \mathcal{X}}$ as $\mathbf{Z}_0$; write the $X \times X$ matrix $\pi(x'|x,0)$ as $\mathbf{\Pi}_0$, and write the $X \times 1$ vector $\left\{ \ln \left[\sum_{i \in \mathcal{I}} \frac{P_i(x')}{P_0(x')} \right] \right\}_{x \in \mathcal{X}}$ as $\mathbf{m}$. The above equation can be written as

$$\mathbf{Z}_0 = \delta \mathbf{\Pi}_0 (\mathbf{Z}_0 + \mathbf{m}).$$

Thus,

$$\mathbf{Z}_0 = (\mathbf{I} - \delta \mathbf{\Pi}_0)^{-1} \delta \mathbf{\Pi}_0 \mathbf{m}.$$

Given this unique solution of $\mathbf{Z}_0$, (E2) immediately provides $Z_i(x)$ for all $i \in \mathcal{I}/\{0\}$ and all $x \in \mathcal{X}$. To obtain $u_i(x)$ for $i \in \mathcal{I} \backslash \{0\}$, note that (24) implies that

$$\mathbf{u}_i = \mathbf{Z}_i - \delta \mathbf{\Pi}_i \mathbf{Z}_0 - \delta \mathbf{\Pi}_i \mathbf{m}, \tag{E3}$$

where $\mathbf{Z}_i$ and $\mathbf{u}_i$ are $X \times 1$ vectors of $\{Z_i(x)\}_{x \in \mathcal{X}}$ and $\{u_i(x)\}_{x \in \mathcal{X}}$ respectively, $\mathbf{\Pi}_i$ is the $X \times X$ matrix $\pi(x'|x,i)$. Recall that in the standard exponential discounting model we have $Z_i(x) = V_i(x)$, thus we can conclude that , $\{\mathbf{u}_i\}_{i \in \mathcal{I} \backslash \{0\}}$ and $\{\mathbf{V}_i\}_{i \in \mathcal{I}}$ are identified once δ, G and $\{u_0(x)\}_{x \in \mathcal{X}}$ are fixed. This replicates the proof of Proposition 2 in Magnac and Thesmar (2002).[3]

F Asymptotic Distribution of the Estimator

This appendix contains the proof of Proposition 3. Recall that the objective function for our maximum pseudo-likelihood estimator is defined to be:

$$\widehat{\boldsymbol{\psi}} = \arg \max_{\{\boldsymbol{\psi}\}} \mathcal{L}(\text{data}; \boldsymbol{\psi}) = \Pi_{j=1}^{n} \Pi_{i=1}^{I} \Pi_{x \in \mathcal{X}} \hat{P}_i(x; \boldsymbol{\psi})^{D_{i,j}(x)}, \tag{F4}$$

where j stands for an individual and n is the sample size; $D_{i,j}(x)$ is an indicator with value 1 if individual j chooses alternative i when state variable is x, and 0 otherwise; and $\hat{P}_i(x; \boldsymbol{\psi})$ is the model's predicted choice probability for alternative i when state is x and parameter values are given by $\boldsymbol{\psi}$. Note from our discussion in Footnote 21 prior to Proposition 1, we know that in the neighborhood around the true parameter values $\boldsymbol{\psi}^*$, $\hat{P}_i(x; \boldsymbol{\psi})$ is continuous and differentiable with respect to $\boldsymbol{\psi}$.

The pseudo-maximum likelihood estimator defined in (F4) is consistent, asymptotically normal and efficient, under the usual technical assumptions for the asymptotic normality and efficiency of MLE estimators (see e.g, Hayashi, 2000). Indeed, to simplify the exposition, consider the case of two alternatives as in the empirical application implemented in the paper, i.e., when $I = 1$;[4] and consider a limiting argument as the number of individual observations per state goes to infinity, i.e. when $n \to \infty$ while keeping the set of states $\mathcal{X}$ fixed.

[3] The only difference is that our argument above indicates that $\mathbf{V}_0$ does not have to be fixed. It can be identified from the model.

[4] The proof can be trivially generalized for $I > 1$ at the cost of much more cumbsersome notation.

Note that we can rewrite the logarithm of likelihood function in Eq. (F4) as:

$$\log \mathcal{L} = \sum_{j=1}^{n} \sum_{x \in \mathcal{X}} \Big(D_{1,j}(x) \log \hat{P}\left(x; \boldsymbol{\psi}\right) + (1 - D_{1,j}(x)) \log(1 - \hat{P}\left(x; \boldsymbol{\psi}\right)) \Big), \tag{F5}$$

where for notational simplicity $\hat{P}\left(x;\boldsymbol{\psi}\right) \equiv \hat{P}_1\left(x;\boldsymbol{\psi}\right)$. The MLE estimator $\widehat{\boldsymbol{\psi}}$ satisfies the first-order condition $\frac{\partial log \mathcal{L}}{\partial \boldsymbol{\psi}}(\widehat{\boldsymbol{\psi}}) = 0$; thus we have:

$$\frac{1}{n} \sum_{j=1}^{n} \sum_{x \in \mathcal{X}} \left(\frac{D_{1,j}(x) - \hat{P}(x;\widehat{\boldsymbol{\psi}})}{\hat{P}(x;\widehat{\boldsymbol{\psi}})(1 - \hat{P}(x;\widehat{\boldsymbol{\psi}}))} \right) \frac{\partial \hat{P}(x;\widehat{\boldsymbol{\psi}})}{\partial \boldsymbol{\psi}} = 0. \tag{F6}$$

Taking the first-order Taylor expansion of the above first-order condition around the true parameter vector $\boldsymbol{\psi}^*$, we have:

$$\begin{aligned}
0 &= \frac{1}{n} \sum_{j=1}^{n} \sum_{x \in \mathcal{X}} \left(\frac{D_{1,j}(x) - \hat{P}(x;\widehat{\boldsymbol{\psi}})}{\hat{P}(x;\widehat{\boldsymbol{\psi}})(1 - \hat{P}(x;\widehat{\boldsymbol{\psi}}))} \right) \frac{\partial \hat{P}(x;\widehat{\boldsymbol{\psi}})}{\partial \boldsymbol{\psi}} \\
&= \frac{1}{n} \sum_{j=1}^{n} \sum_{x \in \mathcal{X}} \left(\frac{D_{1,j}(x) - \hat{P}(x;\boldsymbol{\psi}^*)}{\hat{P}(x;\boldsymbol{\psi}^*)\left(1 - \hat{P}(x;\boldsymbol{\psi}^*)\right)} \right) \frac{\partial \hat{P}(x;\boldsymbol{\psi}^*)}{\partial \boldsymbol{\psi}} \\
&\quad + \frac{1}{n} \sum_{j=1}^{n} \sum_{x \in \mathcal{X}} \left[\begin{array}{c} \left(\frac{D_{1,j}(x) - \hat{P}(x;\boldsymbol{\psi}^*)}{\hat{P}(x;\boldsymbol{\psi}^*)\left(1 - \hat{P}(x;\boldsymbol{\psi}^*)\right)} \right) \frac{\partial^2 \hat{P}(x;\boldsymbol{\psi}^*)}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}'} \\ - \frac{\hat{P}(x;\boldsymbol{\psi}^*)^2 + D_{1,j}(x)\left(1 - 2\hat{P}(x;\boldsymbol{\psi}^*)\right)}{\hat{P}(x;\boldsymbol{\psi}^*)^2\left(1 - \hat{P}(x;\boldsymbol{\psi}^*)\right)^2} \frac{\partial \hat{P}(x;\boldsymbol{\psi}^*)}{\partial \boldsymbol{\psi}} \frac{\partial \hat{P}(x;\boldsymbol{\psi}^*)}{\partial \boldsymbol{\psi}'} \end{array} \right] (\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}^*).
\end{aligned} \tag{F7}$$

The Law of Large Numbers implies that, for each $x \in \mathcal{X}$, $\frac{1}{n}\sum_{j=1}^{n} D_{1,j}(x) \to \mathrm{E}(D_{1,j}(x)) = \hat{P}(x;\boldsymbol{\psi}^*)$. Thus, as $n \to \infty$,

$$\begin{aligned}
&\frac{1}{n} \sum_{j=1}^{n} \sum_{x \in \mathcal{X}} \left[\left(\frac{D_{1,j}(x) - \hat{P}(x;\boldsymbol{\psi}^*)}{\hat{P}(x;\boldsymbol{\psi}^*)\left(1 - \hat{P}\left(x;\boldsymbol{\psi}^*\right)\right)} \right) \frac{\partial^2 \hat{P}(x;\boldsymbol{\psi}^*)}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}'} - \frac{\hat{P}(x;\boldsymbol{\psi}^*)^2 + D_{1,j}(x)\left(1 - 2\hat{P}(x;\boldsymbol{\psi}^*)\right)}{\hat{P}(x;\boldsymbol{\psi}^*)^2\left(1 - \hat{P}(x;\boldsymbol{\psi}^*)\right)^2} \frac{\partial \hat{P}(x;\boldsymbol{\psi}^*)}{\partial \boldsymbol{\psi}} \frac{\partial \hat{P}(x;\boldsymbol{\psi}^*)}{\partial \boldsymbol{\psi}'} \right] \\
\to \quad & -\sum_{x \in \mathcal{X}} \frac{1}{\hat{P}(x;;\boldsymbol{\psi}^*)\left(1 - \hat{P}(x;;\boldsymbol{\psi}^*)\right)} \frac{\partial \hat{P}(x;\boldsymbol{\psi}^*)}{\partial \boldsymbol{\psi}} \frac{\partial \hat{P}(x;\boldsymbol{\psi}^*)}{\partial \boldsymbol{\psi}'} \equiv -H
\end{aligned} \tag{F8}$$

where H denoted the negative of the expected value of the Hessian matrix. Now, the Central Limit Theorem implies that

$$\frac{1}{\sqrt{n}} \sum_{j=1}^{n} \sum_{x \in \mathcal{X}} \left(\frac{D_{1,j}(x) - \hat{P}(x;\boldsymbol{\psi}^*)}{\hat{P}(x;\boldsymbol{\psi}^*)\left(1 - \hat{P}(x;\boldsymbol{\psi}^*)\right)} \right) \frac{\partial \hat{P}(x;\boldsymbol{\psi}^*)}{\partial \boldsymbol{\psi}} \to_d \mathcal{N}(0, J), \tag{F9}$$

where the Fisher information J is given by

$$J = \sum_{x \in \mathcal{X}} Var \left(\frac{D_{1,j}(x) - \hat{P}(x;;\boldsymbol{\psi}^*)}{\hat{P}(x;;\boldsymbol{\psi}^*)\left(1 - \hat{P}(x;;\boldsymbol{\psi}^*)\right)} \right) \frac{\partial \hat{P}(x)}{\partial \theta}(\theta).$$

Using (F7), (F8) and (F9), we thus have that:

$$\sqrt{n}(\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}^*) \to_d \mathcal{N}(0, H^{-1} J H^{-1}).$$

Finally, it follows from the information matrix equality (see Hayashi, 2000, p. 50) that $H = J$. Thus,

$$\sqrt{n}(\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}^*) \to_d \mathcal{N}(0, H^{-1}).$$

F.1 Standard Error Correction When Parameter is on the Boundary

In this subsection, we provide some details about the standard error correction for the parameter estimate of $\tilde{\beta}$, which we estimated to be on the boundary of the parameter space. Our correction procedure follows Andrews (1999, 2001) and in particular Moran (1971). In order to be as close as possible to the notations used in Theorem 1 in Moran (1971), we write $\tilde{\beta}_1 = 1-\tilde{\beta}$, so that the estimate of $\tilde{\beta}_1$, i.e., $\widehat{\tilde{\beta}_1}$, is on the boundary of zero. Let $K-1$ be the dimension of the remaining variables in $\boldsymbol{\psi}$.

Let h_{ij} denote the (i,j)-element of Hessian H, where the Hessian is adjusted appropriately to reflect the change in variables from $\tilde{\beta}$ to $\tilde{\beta}_1 = 1-\tilde{\beta}$. Let σ_{ij} denotes the (i,j)-element of H^{-1}. Then, under the MLE assumptions 1-7 in Moran (1971), the asymptotic distribution of $\sqrt{n}(\widehat{\boldsymbol{\psi}}-\boldsymbol{\psi}^*)$ converges to the mixture of two distributions F_1 and F_2 :

$$\sqrt{n}(\widehat{\boldsymbol{\psi}}-\boldsymbol{\psi}^*) \to_d \frac{1}{2}F_1 + \frac{1}{2}F_2. \tag{F10}$$

In (F10), F_1 is K-dimensional truncated multivariate Normal distribution defined on $\{(0,+\infty),(-\infty,+\infty)^{K-1}\}$ whose density is equal to twice the density of a multivariate Normal distribution with mean zero and covariance matrix of H^{-1}. F_2 is a $(K-1)$-dimensional distribution concentrated on the the subspace $\{0,(-\infty,+\infty)^K-1\}$ such that the joint distribution of $\sqrt{n}(\widehat{\psi_k}-\psi_k^*)$ for $k=2,\ldots,K+1$ is that of

$$\sqrt{n}(\widehat{\psi_k}-\psi_k^*) = \sum_{j=2}^{K} \tilde{\sigma}_{i,j-1}y_j,$$

where the vector y follows multivariate Normal distribution with mean zero and variance matrix H, and the distribution of y is being taken conditional on the inequality $\sigma_{11}^{-1}\sum\sigma_{1s}y_s \leq 0$ where $\tilde{\sigma}_{ij}$ is defined as the (i,j)- element of the inverse of the matrix obtained by removing the first column and the first row of matrix H. See Moran (1971, p. 444) for more details.

G Estimation Results without A Priori Restrictions on β and $\tilde{\beta}$

In this section, we report the estimation results without a priori restrictions that $\beta \in [0,1]$ and $\tilde{\beta} \in [\beta,1]$.[5]

In Table G2 we report the estimation results when $\tilde{\beta}$ is relaxed to $(0,1]$. In comparison to the results reported in Table 6 in the main text when $\tilde{\beta}$ is constrained to be $[\beta,1]$, the estimates do not change much except for standard errors, which is not surprising since $\tilde{\beta}$ reaches 1 in the original estimation.

In Table G3 we report the estimation results when β and $\tilde{\beta}$ are both relaxed to be $(0,2]$. Again, the parameter estimates are very close to the original specifications reported in Table 6 of the main text. Even though we do obtain estimates of $\tilde{\beta}$ that exceed 1, we can not formally reject the hypothesis that $\tilde{\beta}=1$ (see Panel C of Table G3).

H Finite Horizon with Non-Stationary State Transition

In this appendix we discuss the identification strategy for a dynamic discrete choice model for (potentially naive) hyperbolic discounters with finite horizon and non-stationary state transitions. The time period in this case goes from t, the current period, until T, the end of horizon. In this

[5] We thank an anonymous referee for suggesting these relaxations.

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Instantaneous Utility Function Parameters						
Bad Health	-0.434***	-0.724***	-0.138	-0.913***	-0.335***	-0.472***
	(0.147)	(0.216)	(0.087)	(0.163)	(0.082)	(0.131)
Log Income	1.177***	1.167***	1.346***	1.153***	1.265***	1.280***
	(0.016)	(0.016)	(0.009)	(0.034)	(0.011)	(0.009)
Constant	-0.811***	-0.928***	-2.732***	-0.926***	-1.722***	-2.014***
	(0.117)	(0.238)	(0.097)	(0.096)	(0.060)	(0.124)
Panel B: Time Preference Parameters						
δ	0.681***	0.792***	0.741***	0.947***	0.759***	0.764***
	(0.318)	(0.206)	(0.103)	(0.089)	(0.258)	(0.164)
β	0.679***	0.791	0.679***	0.508***	0.578***	0.762
	(0.282)	(0.584)	(0.341)	(0.133)	(0.246)	(0.503)
$\tilde{\beta}$	1.000***	1.000***	0.984***	1.000***	1.000***	1.000***
	(0.452)	(0.278)	(0.253)	(0.094)	(0.340)	(0.130)
Panel C: Hypothesis Tests						
$\text{H}_0 : \beta = 1$	Reject	Reject	Reject	Reject	Reject	Reject
$\text{H}_0 : \tilde{\beta} = \beta$	Reject	Reject	Reject	Reject	Reject	Reject
Exclusion Variables:						
White	Yes	Yes	Yes	Yes	Yes	Yes
Age	Yes	Yes	Yes	Yes	Yes	Yes
Married	Yes	Yes	Yes	Yes	Yes	Yes
HighSchool	Yes	Yes	Yes	No	No	Yes
Insurance	Yes	Yes	Yes	Yes	Yes	No
Mother70	Yes	No	No	No	No	Yes
MotherHighSchool	Yes	No	Yes	No	Yes	Yes
Father70	No	Yes	No	Yes	No	No
FatherHighSchool	No	Yes	No	No	No	No

Table G2: Parameter Estimates for the Instantaneous Utility Function and Time Preference Parameters Under Six Different Sets of Exclusive Restriction Variables with $\tilde{\beta} \in [0, 1]$.

Notes: (1). The last panel indicates the exclusive restriction variables used in the specification in that column, with “Yes” meaning the variable is used, and “No” otherwise; (2). Standard errors for parameter estimates are in parenthesis, and *, **, *** represents statistical significance at 10%, 5% and 1% respectively; (3). For hypothesis tests reported in Panel C, all are rejected with p-value less than 0.01.

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Instantaneous Utility Function Parameters						
Bad Health	-0.258***	-0.313***	-0.143**	-0.123***	-0.224***	-0.569***
	(0.080)	(0.087)	(0.073)	(0.053)	(0.052)	(0.097)
Log Income	0.585***	0.723***	0.606***	0.531	0.434***	0.230
	(0.175)	(0.244)	(0.266)	(0.295)	(0.155)	(0.144)
Constant	-1.354***	-0.845***	-1.557***	-1.458***	-1.437***	-2.289***
	(0.122)	(0.121)	(0.131)	(0.125)	(0.120)	(0.087)
Panel B: Time Preference Parameters						
δ	0.752***	0.701***	0.669***	0.610	0.820***	0.742***
	(0.267)	(0.137)	(0.259)	(0.338)	(0.299)	(0.373)
β	0.680***	0.528***	0.500	0.571***	0.625***	0.718***
	(0.259)	(0.241)	(0.298)	(0.208)	(0.103)	(0.123)
$\tilde{\beta}$	1.050***	1.117***	1.219***	1.055***	1.042***	1.250***
	(0.196)	(0.435)	(0.135)	(0.310)	(0.179)	(0.185)
Panel C: Hypothesis Tests						
$\text{H}_0 : \beta = 1$	Reject	Reject	Reject	Reject	Reject	Reject
$\text{H}_0 : \tilde{\beta} = \beta$	Reject	Reject	Reject	Reject	Reject	Reject
$\text{H}_0 : \tilde{\beta} = 1$	Not Reject	Not Reject	Not Reject	Not Reject	Not Reject	Not Reject
Exclusion Variables:						
White	Yes	Yes	Yes	Yes	Yes	Yes
Age	Yes	Yes	Yes	Yes	Yes	Yes
Married	Yes	Yes	Yes	Yes	Yes	Yes
HighSchool	Yes	Yes	Yes	No	No	Yes
Insurance	Yes	Yes	Yes	Yes	Yes	No
Mother70	Yes	No	No	No	No	Yes
MotherHighSchool	Yes	No	Yes	No	Yes	Yes
Father70	No	Yes	No	Yes	No	No
FatherHighSchool	No	Yes	No	No	No	No

Table G3: Parameter Estimates for the Instantaneous Utility Function and Time Preference Parameters Under Six Different Sets of Exclusive Restriction Variables with $\beta \in [0, 2]$ and $\tilde{\beta} \in [0, 2]$.

Notes: (1). The last panel indicates the exclusive restriction variables used in the specification in that column, with "Yes" meaning the variable is used, and "No" otherwise; (2). Standard errors for parameter estimates are in parenthesis, and *, **, *** represents statistical significance at 10%, 5% and 1% respectively; (3). For hypothesis tests reported in Panel C, all are rejected with p-value less than 0.01.

section, we will use superscripts $t, t+1, t+2, ..., T$ to denote the time period for all the components in our analysis for clarification. The discussion in this section will be shorter and less detailed than that in Section 2, as it will soon be clear that the identification strategy for this non-stationary case is essentially the same as the one for the stationary case, with only minor modification. However, we should emphasize that the identification for this case requires, not surprisingly, that we have access to a panel data set that covers the entire finite horizon.

First, define the current choice-specific value function, $W_i^t\left(x^t\right)$, as follows:

$$W_i^t(x_t) = u_i^t(x_t) + \beta\delta \sum_{x_{t+1}\in\mathcal{X}} V^{t+1}(x_{t+1})\pi(x_{t+1}|x_t, i)dx^{t+1}. \tag{H11}$$

Then, define the choice-specific value function of the next-period self as *perceived* by the current self, $Z_i^{t+1}\left(x_{t+1}\right)$, as follows:

$$Z_i^{t+1}\left(x_{t+1}\right) = u_i^{t+1}\left(x_{t+1}\right) + \tilde{\beta}\delta \sum_{x_{t+2}\in\mathcal{X}} V^{t+2}\left(x_{t+2}\right)\pi(x_{t+2}|x_{t+1}, i). \tag{H12}$$

Given $Z_i^{t+1}\left(x_{t+1}\right)$, we know that the current self's perception of her future self's choice, i.e., $\tilde{\sigma}$ as defined in Definition 2 is simply

$$\begin{aligned}\tilde{\sigma}\left(x_{t+1}, \boldsymbol{\varepsilon}_{\mathbf{i}}^{\mathbf{t+1}}\right) &= \max_{i\in\mathcal{I}}\left[u_i^{t+1}\left(x_{t+1}\right) + \varepsilon_{i,t+1} + \tilde{\beta}\delta \sum_{x_{t+2}\in\mathcal{X}} V^{t+2}(x_{t+2})\pi(x_{t+2}|x_{t+1}, i)\right] \\ &= \max_{i\in\mathcal{I}}\left[Z_i^{t+1}\left(x_{t+1}\right) + \varepsilon_{i,t+1}\right].\end{aligned}$$

Let us define the probability of choosing alternative j by the the next period self as perceived by the current period self, $\tilde{P}_j\left(x_{t+1}\right)$, when next period state is x_{t+1} :

$$\begin{aligned}\tilde{P}_j^{t+1}\left(x_{t+1}\right) &= \Pr\left[\tilde{\sigma}\left(x_{t+1}, \boldsymbol{\varepsilon}_{t+1}\right) = j\right] \\ &= \Pr\left[Z_j^{t+1}\left(x_{t+1}\right) + \varepsilon_{j,t+1} \geq Z_{j'}^{t+1}\left(x_{t+1}\right) + \varepsilon_{j',t+1}, \forall j' \neq j\right].\end{aligned}$$

Further denote

$$V_i^{t+1}\left(x_{t+1}\right) = u_i^{t+1}\left(x_{t+1}\right) + \delta \sum_{x_{t+2}\in\mathcal{X}} V^{t+2}(x_{t+2})\pi(x_{t+2}|x_{t+1}, i). \tag{H13}$$

According to the definition of $V\left(\cdot\right)$ as given by (2), $V^{t+1}\left(x_{t+1}\right)$ is simply the expected value of $\left[V_i^{t+1}\left(x_{t+1}\right) + \varepsilon_i^{t+1}\right]$ where i is the chosen alternative according to $\tilde{\sigma}\left(x_{t+1}, \boldsymbol{\varepsilon}_{t+1}\right)$. Thus it must satisfy the following relationship:

$$V^{t+1}\left(x_{t+1}\right) = \mathrm{E}_{\boldsymbol{\varepsilon}^{\mathbf{t+1}}}\left[V_{\tilde{\sigma}(x_{t+1},\boldsymbol{\varepsilon}_{t+1})}^{t+1}\left(x_{t+1}\right) + \varepsilon_{\tilde{\sigma}(x_{t+1},\boldsymbol{\varepsilon}_{t+1})}^{t+1}\right]. \tag{H14}$$

Now note from (H12) and (H13), we have

$$V_i^{t+1}\left(x_{t+1}\right) = Z_i^{t+1}\left(x_{t+1}\right) + \left(1 - \tilde{\beta}\right)\delta \sum_{x_{t+2}\in\mathcal{X}} V^{t+2}(x_{t+2})\pi(x_{t+2}|x_{t+1}, i). \tag{H15}$$

The relationship in (H15) is crucial as it allows us to rewrite (H14) as:

$$\begin{aligned}
V^{t+1}(x_{t+1}) &= \mathrm{E}_{\boldsymbol{\varepsilon}_{t+1}}\left[V^{t}_{\tilde{\sigma}(x_{t+1},\boldsymbol{\varepsilon}_{t+1})}(x_{t+1})+\varepsilon_{\tilde{\sigma}(x_{t+1},\boldsymbol{\varepsilon}_{t+1}),t}\right] \\
&= \mathrm{E}_{\boldsymbol{\varepsilon}_{t+1}}\left[\begin{array}{c} Z_i^{t+1}(x_{t+1})+\varepsilon_{\tilde{\sigma}(x_{t+1},\boldsymbol{\varepsilon}^{\mathbf{t+1}}),t} \\ +\left(1-\tilde{\beta}\right)\delta\sum_{x_{t+2}\in\mathcal{X}}V^{t+2}(x_{t+2})\pi(x_{t+2}|x_{t+1},\tilde{\sigma}(x_{t+1},\boldsymbol{\varepsilon}_{t+1}))\end{array}\right] \\
&= \mathrm{E}_{\boldsymbol{\varepsilon}_{t+1}}\max_{i\in\mathcal{I}}\left[Z_i^{t+1}(x_{t+1})+\varepsilon_{i,t+1}\right] \\
&\quad +\left(1-\tilde{\beta}\right)\delta\mathrm{E}_{\boldsymbol{\varepsilon}_{t+1}}\sum_{x_{t+2}\in\mathcal{X}}V^{t+2}(x_{t+2})\pi(x_{t+2}|x_{t+1},\tilde{\sigma}(x_{t+1},\boldsymbol{\varepsilon}_{t+1})) \\
&= \mathrm{E}_{\boldsymbol{\varepsilon}_{t+1}}\max_{i\in\mathcal{I}}\left[Z_i^{t+1}(x_{t+1})+\varepsilon_{i,t+1}\right] \\
&+ \left(1-\tilde{\beta}\right)\delta\sum_{j\in\mathcal{I}}\tilde{P}_j^{t+1}(x_{t+1})\sum_{x_{t+2}\in\mathcal{X}}V^{t+2}(x_{t+2})\pi(x_{t+2}|x_{t+1},j) \qquad \text{(H16)}
\end{aligned}$$

The probability of observing action i being chosen at a given state variable x, in this non-stationary case, is still:

$$P_i^t(x_t)=\Pr\left[W_i^t(x_t)+\varepsilon_{i,t}>\max_{j\in\mathcal{I}\backslash\{i\}}\left\{W_j^t(x)+\varepsilon_{j,t}\right\}\right]=\frac{\exp\left[W_i^t(x_t)\right]}{\sum_{j=0}^{I}\exp\left[W_j^t(x_t)\right]}, \qquad \text{(H17)}$$

where $P_i^t(x_t)$ is the current-period self's equilibrium choice probabilities and will be observed in the data.

The relationship between W_i and Z_i can no longer be described as in Equation 13, however, we still have

$$\mathrm{E}_{\boldsymbol{\varepsilon}_{t+1}}\max_{i\in\mathcal{I}}\{Z_i^{t+1}(x_{t+1})+\varepsilon_{i,t+1}\}=\ln\left\{\sum_{i\in\mathcal{I}}\exp\left[Z_i^{t+1}(x_{t+1})\right]\right\}, \qquad \text{(H18)}$$

and

$$\tilde{P}_j^{t+1}(x_{t+1})=\frac{\exp\left[Z_j^{t+1}(x_{t+1})\right]}{\sum_{i=0}^{I}\exp\left[Z_i^{t+1}(x_{t+1})\right]}. \qquad \text{(H19)}$$

Using (H18) and (H19), we can rewrite (H16) as

$$\begin{aligned}
V^{t+1}(x_{t+1}) &= \ln\left\{\sum_{i\in\mathcal{I}}\exp\left[Z_i^{t+1}(x_{t+1})\right]\right\} \\
&+\left(1-\tilde{\beta}\right)\delta\sum_{j\in\mathcal{I}}\frac{\exp\left[Z_j^{t+1}(x_{t+1})\right]}{\sum_{i=0}^{I}\exp\left[Z_i^{t+1}(x^{t+2})\right]}\sum_{x_{t+2}\in\mathcal{X}}V(x_{t+2})\pi(x_{t+2}|x_{t+1},j). \qquad \text{(H20)}
\end{aligned}$$

In this non-stationary case with finite horizon, at $t=T$ where the continuation value is zero, we have that

$$Z_i^T=u_i^T=V_i^T$$

which leads to:

$$V^T=\ln\sum_{i\in\mathcal{I}}\exp[Z_i^T]=\ln\sum_{i\in\mathcal{I}}\exp[u_i^T]. \qquad \text{(H21)}$$

Equations (H21) and (H12) combined give us Z_i^{T-1}, which can be further combined with equations (H19) and (H20) for $t=T-1$ to give us V^{T-1}. We can keep doing backward induction in this way until we reach V^{t+1}, which can be used in (H11) to derive W_i^t, the current choice specific continuation value function. Equation (H17) shows the relationship between the observed choice pattern from the data and W_i^t.