1 (2016) 1-48 Submitted 4/00; Published 10/00

DRAFT: MANUSCRIPT IN PREPARATION

Overcoming Temptation: Incentive Design For Intertemporal Choice

Michael C. Mozer MICHAEL.MOZER@COLORADO.EDU
Shruthi Sukumar SHRUTHI.SUKUMAR@COLORADO.EDU
Camden Elliott-Williams CAMDEN.ELLIOTTWILLIAMS@COLORADO.EDU
Department of Computer Science
University of Colorado
Boulder, CO 80309-0430, USA

Shabnam Hakimi SHABNAM.HAKIMI@COLORADO.EDU
Institute of Cognitive Science
University of Colorado Boulder, CO 80309-0345, USA

Adrian F. Ward ADRIAN.WARD@MCCOMBS.UTEXAS.EDU
McCombs School of Business
University of Texas
Austin, TX 78705, USA

Editors:

Abstract

Individuals are often faced with temptations that can lead them astray from long-term goals. We're interested in developing interventions that steer individuals toward making good initial decisions and then maintaining those decisions over time. In the realm of financial decision making, a particularly successful approach is the prize-linked savings account: individuals are incentivized to make deposits by tying deposits to a periodic lottery that awards bonuses to the savers. Although these lotteries have been very effective in motivating savers across the globe, they are a one-size-fits-all solution. We investigate whether customized bonuses can be more effective. We formalize a delayed-gratification task as a Markov decision problem and characterize individuals as rational agents subject to temporal discounting, costs associated with effort, and moment-to-moment fluctuations in willpower. Our theory is able to explain key behavioral findings in intertemporal choice. We created an online delayed-gratification game in which the player scores points by choosing a queue to wait in and patiently advancing to the front. Data collected from the game is fit to the model, and the instantiated model is then used to optimize predicted player performance over a space of incentives. We demonstrate that customized incentive structures can improve goal-directed decision making.

Should you go hiking today or work on that manuscript? Should you have a slice of cake or stick to your diet? Should you upgrade your flat-screen TV or contribute to your retirement account? Individuals are regularly faced with temptations that lead them astray from long-term goals. These temptations all reflect an underlying challenge in behavioral control that involves choosing between actions leading to small but immediate rewards and actions leading to large but delayed rewards. We introduce a formal model of this *delayed gratification* decision task, extending the Markov decision framework to incorporate the

psychological notion of willpower, and using formal models to optimize behavior by designing incentives to assist individuals in achieving long-term goals.

Consider the serious predicament with retirement planning in the United States. Only 55% of working-age households have retirement account assets—whether an employer-sponsored plan or an IRA—and the median account balance for near-retirement households is $14,500. Even considering households' net worth, 2/3 fall short of conservative savings targets based on age and income (Rhee and Boivie, 2015). Furthermore, 40% of every dollar contributed to the accounts of savers under age 55 simultaneously flows out of the retirement systems, not counting loans to oneself (Argento et al., 2015). In 2013, the US government and nonprofits spent $670M on financial education, yet financial literacy accounts for a miniscule 0.1% of the variance in financial outcomes (Fernandes et al., 2014).

One technique that has been extremely successful in encouraging savings, primarily in Europe and the developing world but more recently in the US as well, is the *prize linked savings account (PLSA)* (Kearney et al., 2010). The idea is to pool a fraction of the interest from all depositors to fund a prize awarded by periodic lotteries. Just as ordinary lotteries entice individuals to purchase tickets, the PLSA encourages individuals to save. Disregarding the fact that lotteries function in part because individuals overvalue low-probability gains (Kahneman and Tversky, 1979), the core of the approach is to offer savers the prospect of short-term payoffs in exchange for them committing to the long term. Although the account yields a lower interest rate to fund the lottery, the PLSA increases the net expected account balance due to greater commitment to participation.

The PLSA is a one-size-fits-all solution. A set of incentives that that work well for one individual or one subpopulation may not be optimal for another. In this article, we investigate approaches to customizing incentives to an individual or a subpopulation with the aim of achieving greater adherence to long-term goals and ultimately, better long-term outcomes for the participants. Our approach involves: (1) building a model to characterize the behavior of an individual or group, (2) fitting the model with behavioral data, (3) using the model to determine an incentive structure that optimizes outcomes, and (4) validating the model by showing better outcomes with model-derived incentives than with alternative incentive structures.

1. Intertemporal Choice

Intertemporal choice involves decisions that produce gains and losses at different points in time. How an individual interprets delayed consequences influences the utility or value associated with a decision. When consequences are discounted with the passage of time, decision making is biased toward more immediate gains (and more distant losses). The *delay discounting* task is often used to study intertemporal choice (Green and Myerson, 2004). Individuals are asked to choose between two alternatives, e.g., $1 today versus $\$X$ in Y days. By identifying the X that yields subjective indifference for a given Y, one can estimate an individual's discounting of future outcomes. Discount rates vary across individuals yet show stability over extended periods of time (Kirby, 2009).

This paradigm involves a single, hypothetical decision and reveals the intrinsic future value of an outcome. However, it does not address the temporal dynamics of behavior during a delay period. Once an initial decision is made to wait for a large reward, some scenarios

permit an individual to abandon the decision *at any instant* in favor of the small immediate reward. For example, in the classic marshmallow test (Mischel and Ebbesen, 1970), children are seated at a table with a single marshmallow. They are allowed to eat the marshmallow, but if they wait while the experimenter steps out of the room, they will be offered a second marshmallow when the experimenter returns. In this *delayed gratification* task, children must continually contemplate whether to eat the marshmallow or wait for two marshmallows. Their behavior depends not only on the hypothetical discounting of future rewards but on an individual's *willpower*—their ability to maintain focus on the larger reward and not succumb to temptation before the experimenter returns. Defection at any moment eliminates the possibility of the larger reward.

The marshmallow test achieved renown not only because it turns out to be predictive of later life outcomes (Mischel et al., 1989), but because it is analogous to many situations involving delayed gratification. Like the marshmallow test, some of these situations have an unspecified time horizon (e.g., exercise, waiting for an elevator, spending during retirement). However, others have a known horizon (e.g., avoiding snacks before dinner, saving for retirement, completing a college degree). Our work addresses the case of a known or assumed horizon.

Beyond whether the horizon is known or not, delayed-gratification tasks may also be characterized in terms of the number of opportunities to obtain the delayed reward. The marshmallow test is *one shot*, but many true-to-life scenarios have an *iterative* nature. For example, in retirement planning, the failure to contribute to the account one month does not preclude contributing the next month. Another intuitive example involves allocating time within a work day. One must choose between tasks that are relatively quick and provide a moment of satisfaction (e.g., answering email, cleaning a desk top) and those that are more effortful but also yield a greater sense of accomplishment (e.g., editing a paper for submission to a journal, reading a research article). Our work addresses both one-shot and iterated delayed-gratification tasks. For such tasks, we're interested in developing personalized interventions that assist individuals both in making good initial decisions and in maintaining those decisions over time.

2. Theories of Intertemporal Choice

Nearly all previous conceptualizations of intertemporal choice have focused on the shape of the discounting function and the initial 'now versus later' decision, not the time course. One exception is the work of McGuire and Kable (2013) who frame failure to postpone gratification as a rational, utility-maximizing strategy when the time at which future outcomes materialize is uncertain. Our theory is complementary in providing a rational account in the known time horizon situation.

There is a rich literature on treating human decision making from the framework of Markov decision processes (*MDP*s; e.g., Shen et al., 2014; Niv et al., 2012), but this research does not directly address intertemporal choice.

Kurth-Nelson and Redish (2010, 2012) have explored a reinforcement learning framework to model precommitment in decision making as a means of preventing impulsive defections. This interesting work focuses on the initial decision whether to precommit rather than the ongoing possibility of defection. To the best of our knowledge, we are the first to adopt an

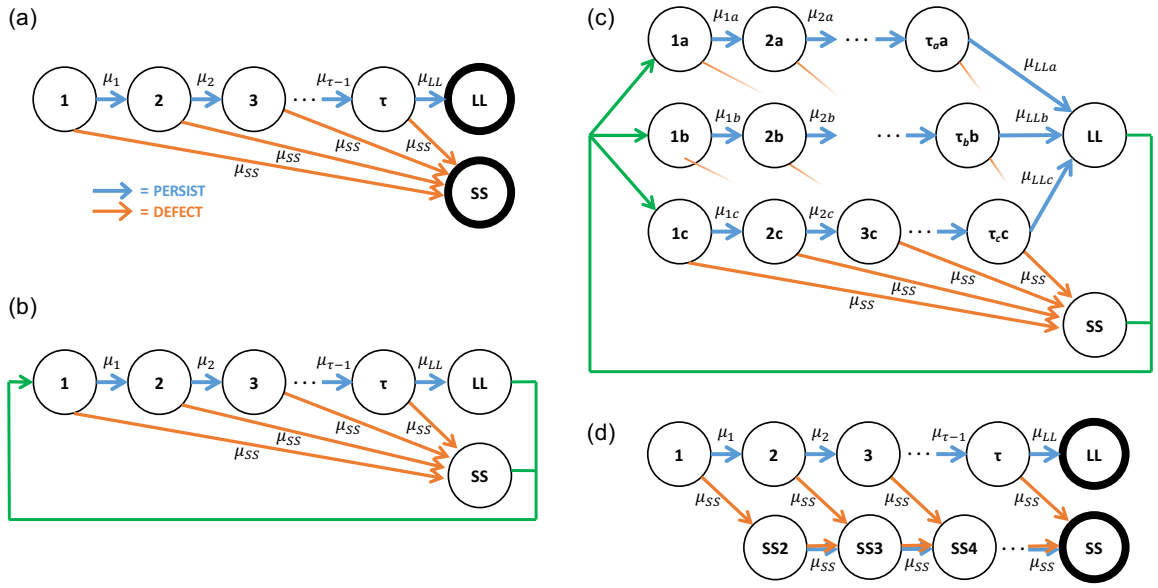

Figure 1: Finite-state environment formalizing (a) the one-shot delayed-gratification task; (b) the iterated delayed-gratification task; (c) the iterated delayed-gratification task with variable delays and LL outcomes; and (d) an efficient approximation to the iterated delayed-gratification task, suitable when episodes are independent of one another.

MDP perspective on intertemporal choice, a field which has relied primarily primarily on verbal, qualitative accounts.

One challenge to modeling behavior with MDPs, whether in the framework of reinforcement learning or dynamic programming, is that it is mathematically convenient to assume exponential discounting, whereas studies of human intertemporal choice support hyperbolic discounting (Frederick et al., 2002). Kurth-Nelson and Redish (2010) have proposed a solution to this issue by exploiting the fact that a hyperbolic function can be well approximated by a mixture of exponentials. In the work we present, we assume exponential discounting, but our work could readily be extended in the same manner as Kurth-Nelson and Redish (2010).

3. Formalizing Delayed-Gratification Tasks as a Markov Decision Problem

In this section, we formalize a delayed-gratification task as a Markov decision problem, which we will refer to as the *DGMDP*. We assume time to be quantized into discrete steps and we focus on situations with a known or assumed time horizon, denoted τ. At any step, the agent may DEFECT and collect a small reward, or the agent may PERSIST to the next step, eventually collecting a large reward at step τ. We use μ_{SS} and μ_{LL} to denote the *smaller sooner (SS)* and *larger later (LL)* rewards. Figure 1a shows a finite-state representation of the one-shot task with terminal states LL and SS that correspond to resisting and succumbing to temptation, respectively, and states for each time step between the initial and final times,

$t \in \{1, 2, ..., \tau\}$. Rewards are associated with state transitions. The possibility of obtaining *intermediate* rewards during the delay period is annotated via $\boldsymbol{\mu}_{1:\tau-1} \equiv \{\mu_1, ..., \mu_{\tau-1}\}$, which we return to later. With exponential discounting, rewards n steps ahead are devalued by a factor of γ^n, $0 \leq \gamma < 1$.

Given the DGMDP, an optimal decision sequence is trivially obtained by value iteration. However, this sequence is a poor characterization of human behavior. With no intermediate rewards ($\boldsymbol{\mu}_{1:\tau-1} = \mathbf{0}$), it takes one of two forms: either the agent defects at $t = 1$ or the agent persists through $t = \tau$. In contrast, individuals will often persist some time and then defect, and when placed into the same situation repeatedly, behavior is nondeterministic. For example, replicability on the marshmallow test is quite modest ($\rho < 0.30$: Mischel et al., 1988).

The discrepancy between human delayed-gratification behavior and the optimal decision-making framework might indicate an incompatibility. However, we prefer a *bounded rationality* perspective on human cognition according to which behavior is cast as optimal but subject to cognitive constraints. We claim two specific constraints.

1. Individuals exhibit moment-to-moment fluctuations in *willpower* based on factors such as sleep, hunger, mood, etc. Low willpower causes an immediate reward to seem more tempting, and high willpower, less tempting. We characterize willpower as a one-dimensional Gaussian process, $W = \{W_t\}$, with $w_1 \sim \text{Gaussian}(0, \sigma_1^2)$ and $w_t \sim \text{Gaussian}(w_{t-1}, \sigma^2)$. We suppose that willpower modulates an individual's subjective value of defecting at step t:

$$Q(t, w; \text{DEFECT}) = \mu_{\text{SS}} - w, \tag{1}$$

   where $Q(s; a)$ denotes the value associated with performing action a in state s, and the state space consists of the discrete step t and the continuous willpower level w.

2. Behavioral, economic, and neural accounts of decision making suggest that *effort* carries a cost, and that rewards are weighed against the effort required to obtain it (e.g., Kivetz, 2003). This notion is incorporated into the model via an effort cost, μ_{E} associated with persevering:

$$Q(t, w; \text{PERSIST}) = \begin{cases} \mu_{\text{E}} + \mu_t + \gamma \, \mathbb{E}_{W_{t+1}|W_t=w} V(t+1, w_{t+1}) & \text{for } t < \tau \\ \mu_{\text{LL}} & \text{for } t = \tau \end{cases} \tag{2}$$

$$\text{where} \quad V(t, w) \equiv \max_a \, Q(t, w; a). \tag{3}$$

With these two constraints, we will show that the model not only has adequate expressive power to fit behavioral data, but also has the explanatory power to predict experimental outcomes.

The one-shot DGMDP in Figure 1a can be extended to model the iterated task (Figure 1b), even when there is variability in the reward (μ_{LL}) or duration (τ) across *episodes* (Figure 1c).[1]

1. Figures 1b,c describe an indefinite series of episodes. If the total number of episodes or steps is constrained, as in any realistic scenario (e.g., an individual has eight hours in the work day to perform tasks like answering email), then the state must be augmented with a representation of remaining time. We dodge this complication by modeling situations in which the 'end game' is not approaching, e.g., only the first half of a work day.

Finally, it is straightforward to show that the solution to the iterated DGMDPs in Figures 1b or 1c is identical to the solution to the simpler and more tractable one-shot DGMDP in Figure 1d under certain constraints (see Supplementary Materials). Essentially, Figure 1d models the choice between the LL reward or a sequence of SS rewards matched in total number of steps, effectively comparing the reward rates for LL and SS, the critical variables in the iterated DGMDP.

To summarize, we have formalized one-shot and iterative delayed-gratification task with known horizon as a Markov decision problem with parameters $\Theta_{\text{task}} \equiv \{\tau, \mu_{\text{SS}}, \mu_{\text{LL}}, \boldsymbol{\mu}_{1:\tau-1}\}$, and a constrained rational agent parameterized by $\Theta_{\text{agent}} \equiv \{\gamma, \sigma_1, \sigma, \mu_{\text{E}}\}$. We now turn to solving the DGMDP and characterizing its properties.

3.1 Solving The Delayed-Gratification Markov Decision Problem (DGMDP)

The simple structure of the environment allows for a backward-induction solution to the Bellman equation (Equation 2). Although the continuous willpower variable precludes an analytical solution for $V(t, w)$, we construct a piecewise linear approximation (PLA) over w for each step t. To justify the PLA, consider the shape of $V(t, w)$. With high willpower ($w \to \infty$), the agent almost certainly persists to the LL reward and the function asymptotes at the discounted μ_{LL}. With low willpower ($w \to -\infty$), the agent almost certainly defects and the function approaches $\mu_{\text{SS}} - w$. Thus, both extrema of the value function are linear with known slope and intercept. At step τ, these two linear segments exactly define the value function. At $t < \tau$, there is an intermediate range within which small fluctuations in willpower can influence the decision and the expectation in Equation 2 yields a weighted mixture of the two extrema, which is well fit by a single linear segment—defined by its slope a_t and intercept b_t. With $V(t, w)$ expressed as a PLA, the expectation in Equation 2 becomes:

$$\begin{aligned}\mathbb{E}_{W_t|W_{t-1}=w} V(t, w_t) =& \Phi\left(z_t^-\right)(\mu_{\text{SS}} - w) + \left(\Phi\left(z_t^+\right) - \Phi\left(z_t^-\right)\right)(b_t + a_t w) \\ &+ \left(1 - \Phi\left(z_t^+\right)\right) c_t + \sigma\phi(z_t^-) + \sigma a_t \left(\phi(z_t^-) - \phi(z_t^+)\right),\end{aligned} \tag{4}$$

where $\Phi(.)$ and $\phi(.)$ are the cdf and pdf of a standard normal distribution, respectively, and the standardized segment boundaries are $z_t^- = \sigma^{-1}[(\mu_{\text{SS}} - b_t)/(a_t + 1) - w]$ and $z_t^+ = \sigma^{-1}[(c_t - b_t)/a_t - w]$. The backup is seeded with $z_\tau^- = z_\tau^+ = \sigma^{-1}(\mu_{\text{SS}} - \mu_{\text{LL}} - w)$ and $a_\tau = b_\tau = c_\tau = \mu_{\text{LL}}$. After each backup step, a Levenberg-Marquardt nonlinear least squares fit obtains a_{t-1} and b_{t-1}; c_{t-1}—the value of steadfast persistence—is obtained by propagating the discounted reward for persistence: $c_{t-1} = \mu_{\text{E}} + \mu_t + \gamma c_t$.

Figure 2a shows the value as a function of willpower at each step of an eight step DGMDP with an LL reward twice that of the SS reward, like the canonical marshmallow test. Both the exact value-function formulation (Equation 4) and the corresponding PLA are presented in colored and black lines, respectively. To ensure accuracy of the estimate and to eliminate an accumulation of estimation errors, we have also used a fine piecewise constant approximation in the intermediate region, yet the model output is almost identical.

Using the value function, we can characterize the agent's behavior in the DGMDP via the likelihood of defecting at various steps. With D denoting the defection step, we have the *hazard probability*

$$h_t \equiv P(D = t | D \geq t) \equiv P(W_t < w_t^* | W_1 \geq w_1^*, ..., W_{t-1} \geq w_{t-1}^*), \tag{5}$$

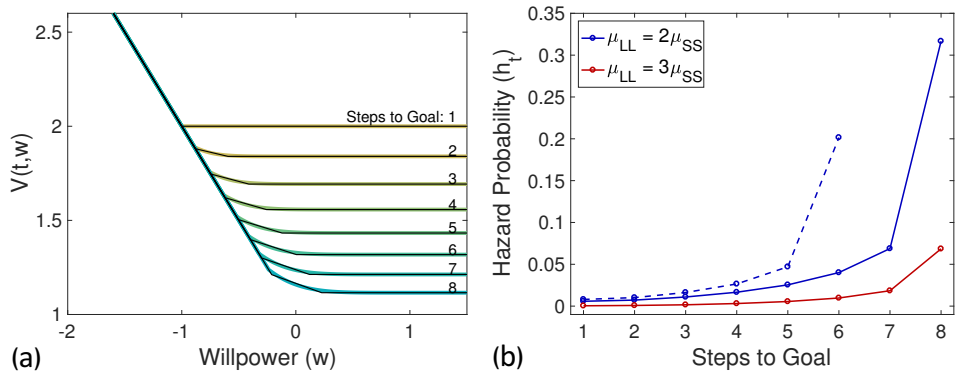

Figure 2: (a) Value function for a DGMDP with $\tau = 8$, $\sigma = .25$, $\sigma_1 = .50$, $\gamma = .92$, $\mu_{\text{E}} = \mu_t = 0$, $\mu_{\text{LL}} = 2$, $\mu_{\text{SS}} = 1$, exact (colored curves) and piecewise linear approximation (black lines). (b) Hazard functions for the parameterization in (a) (solid blue curve), with a higher level of LL reward (red curve), and with a shorter delay period, $\tau = 6$ (dashed blue curve).

where w^* is the willpower threshold that yields action indifference, $Q(t, w^*; \text{DEFECT}) = Q(t, w^*; \text{PERSIST})$. To represent the posterior distribution over willpower at each non-defection step, we initially used a particle filter but found a computationally more efficient and stable solution with quantile-based samples. We approximate the W_1 prior and ΔW with discrete, equal probability q-quantiles. We reject values for which defection occurs, and then propagate $W_{t+1} = W_t + \Delta W$ which results in up to q^2 samples, which we thin back to q-quantiles at each step. Using $q = 1000$ produces nearly identical results to selecting a much higher density of samples.

The solid blue curve in Figure 2b shows the hazard function for the DGMDP in Figure 2a. Defection rates drop as the agent approaches the goal. Defection rates also scale with the LL reward, as illustrated by the contrast between the solid blue and red curves. Finally, defection rates depend both on relative and absolute steps to goal: contrasting the solid and dashed blue curves, corresponding to $\tau = 8$ and $\tau = 6$, respectively, the defection rate at a given number of steps from the goal depends on τ. We will shortly show that human behavioral data exhibit this same qualitative property. Interestingly, the correlation in willpower from one step to the next is critical in obtaining this property. When willpower is independent from step to step, i.e., $w_t \sim \text{Gaussian}(0, \sigma^2)$, defection rates depend only on absolute steps to goal. Thus, moment-to-moment correlation in willpower is essential for modeling human behavior.

3.2 Behavioral Phenomena Explained

We consider the solution of the DGMDP as a rational theory of human cognition. It is meant to explain both an individual's initial choice ("Should I open a retirement account?") as well as the temporal dynamics of sustaining that choice ("Should I withdraw the funds to buy a car?")

Our theory explains two key phenomena in the literature. First, failure on a DG task is sensitive to the relative magnitudes of the SS and LL rewards (Mischel, 1974). Figure 2b presents hazard functions for two reward magnitudes. The probability of obtaining the LL reward is greater with $\mu_{\text{LL}}/\mu_{\text{SS}} = 3$ than with $\mu_{\text{LL}}/\mu_{\text{SS}} = 2$. Figure 2b can also accommodate the finding that environmental reliability and trust in the experimenter affect outcomes in the marshmallow test (Kidd et al., 2012): in unreliable or nonstationary environments, the expected LL reward is lower than the advertised reward, and the DGMDP is based on reward expectations. Second, a reanalysis of data from a population of children performing the marshmallow task shows a declining hazard rate over the task period of 7 minutes (McGuire and Kable, 2013). The rapid initial drop in the empirical curve looks remarkably like the curves in Figure 2b. One might interpret this phenomenon as a *finish-line effect*: the closer one gets to a goal, the greater is the commitment to achieve the goal. However, the model suggests that this behavior arises not from abstract psychological constructs but because of correlations in willpower over time: if an individual starts down the path to an LL reward, the individual's willpower at that point must be high. The posterior willpower distributions reflect the elimination of individuals with low momentary willpower, which contributes to the declining hazard rate. Also contributing is the exponential increase in value of the discounted LL reward as the agent advances through the DGMDP. McGuire and Kable (2013) explain the empirical hazard function via a combination of uncertainty in the time horizon and time-fluctuating discount rates. Our theory shows that these strong assumptions are not necessary, and our theory can address situations with a well delineated horizon such as retirement saving. Additionally, our theory aims to move beyond population data and explain the granular dynamical behavior of an individual.

4. Optimizing Incentives

With a computational theory of the DG task in hand, we now explore a mechanism-design approach (Nisan and Ronen, 1999) aimed at steering individuals toward improved long-term outcomes. We ask whether we can provide incentives to rational value-maximizing agents that will increase their expected reward subject to constraints on the incentives.

We focus on an investment scenario roughly analogous to a prize-linked savings account (PLSA). Suppose an individual has x dollars which they can deposit into a bank account earning interest at rate r, compounded annually. At the start of each year, they decide whether to continue saving (PERSIST) or to withdraw and spend their *entire* savings with interest accumulated thus far (DEFECT).[2] Our goal to assist them in maximizing the profit they reap over $\tau - 1$ years from their initial investment. Our incentive mechanism is a a schedule of lotteries. We refer to expected lottery distributions as *bonuses*, even though they are funded through the interest earned by a population of individuals, like the prizes of the PLSA.

With μ_t denoting the bonus awarded in year t and $\boldsymbol{\mu}_{1:\tau-1}$ denoting the set of scheduled bonuses, our goal as mechanism designers is to identify the schedule that maximizes the

2. Although this all-or-none withdrawal of savings is not entirely realistic, it reduces the decision space to correspond with the FSM in Figure 1a. Were we to allow intermediate levels of withdrawal, the simulation would yield intermediate benefits of incentives.

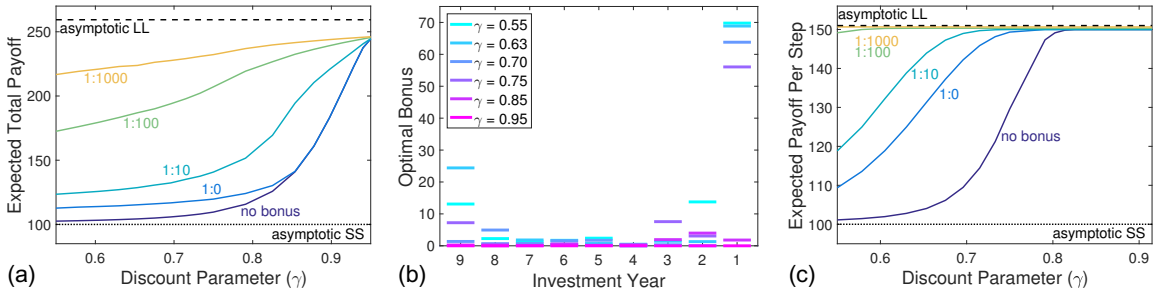

Figure 3: Bonus optimization for an agent with $\sigma_1 = 50$, $\sigma = 30$, $\mu_{\mathrm{E}} = 0$, and $\gamma \in [0.55, 0.95]$. (a) Expected payoff for the one-shot DGMDP for various bonus scenarios, including no bonus and optimal bonuses with lottery odds 1:0, 1:10, 1:100, and 1:1000. In these simulations, the interest-accrual scheme is used to constrain bonuses and payoffs. (b) Optimal bonus amounts at each step for various γ and lottery 1:0 (certain win), on the scale of an $x = 100$ initial pool of funds. (c) Expected payoff per time step for the iterated DGMDP for various bonus scenarios. In these simulations, the bonus-limits scheme is used to constrain bonuses and payoffs.

expected net accumulation from an individual's investment:

$$\boldsymbol{\mu}^*_{1:\tau-1} = \mathrm{argmax}_{\boldsymbol{\mu}_{1:\tau-1}} \sum_{t=1}^{\tau} P(D = t|\boldsymbol{\mu}_{1:\tau-1}) \left[b_t + \sum_{t'=1}^{t-1} \mu_{t'} \right], \tag{6}$$

where b_t is the amount banked at the start of year t, with $b_1 = x$ and $b_{t+1} = (1 + r)(b_t - \mu_t)$, and D is the year of defection, where $D = 1$ represents immediate defection and $D = \tau$ represents the the account reaching maturity. Defection probabilities are obtained from the theory (Equation 5).

To illustrate this approach, we conducted a simulation with $\gamma \in [0.55, 0.95]$, $\tau = 10$, $r = 0.1$, and $x = 100$, comparing an agent's expected accumulation without bonuses and with optimal bonuses. Optimization is via direct search using the simplex algorithm over unconstrained variables $p_t \equiv \mathrm{logit}(\mu_t/b_t)$, representing the proportion of the bank being distributed as a bonus.

We first consider the case of deterministic bonuses: the agent receives bonus μ_t in year t with certainty. Figure 3a shows the expected payoff as a function of an agent's discount factor γ for the scenario with no bonuses (purple curve) versus optimal bonuses awarded with probability 1.0 (light blue curve, labeled with the odds of a bonus being awarded, '1:0'). For reference, the asymptotic SS and LL payoffs are shown with dotted and dashed lines, respectively.

With high discounting, this simulation yields a modest ($\sim 10\%$) improvement in an individual's expected accumulation by providing bonuses at the end of the early years and going into the final year (Figure 3b). Bonuses are recommended only when the gain from encouraging persistence beats the loss of interest on an awarded bonus. With low discounting, the model optimization recommends no bonuses. Thus, the simulation recommends different incentives to individuals depending on their discount factors.

Now consider a lottery such as that conducted for the PLSA. If individuals operate based on expected returns, an uncertain lottery with odds 1:α and payoff $(\alpha + 1)\mu_t$ would be equivalent to a certain payoff of μ_t. However, as characterized by prospect theory Kahneman and Tversky (1979), individuals overweight low probability events. Using median parameter estimates from cumulative prospect theory Tversky and Kahneman (1992) to infer subjective probabilities on lotteries with 1:10, 1:100, and 1:000 odds, we optimize bonuses for these cases.[3] As depicted by the three upper curves in Figure 3a, lotteries such as the PLSA can significantly boost the benefit of incentive optimization.

Lotteries and interest accrual are not suitable for all delayed-gratification tasks. For instance, one would not wish to encourage a dieter by offering a lottery for a 50-gallon tub of ice cream or the promise of a massive all-one-can-eat desert buffet at the conclusion of the diet. To demonstrate the flexibility of our framework, we posit a *bonus-limit* scheme as an alternative to the *interest-accrual* scheme in which up to n_b bonuses of fixed size can be awarded and the optimization determines the time steps at which they are awarded. We conducted a simulation with the iterated DGMDP (Figure 1d) using $\gamma \in [0.55, 0.95]$, $\tau = 10$, awarding of $n_b \leq 4$ bonuses each of value 50, $\mu_{\text{SS}} = 100$, and $\mu_{\text{LL}} = 150\tau - 50n_b$. Multiple bonuses could be awarded in the same step, but bonuses were limited such that no defection could achieve a reward rate greater than μ_{SS}. This set up anticipates human experiments that we report later in the article.

Figure 3c shows expected payoff per step, ranging from 100 from the SS reward to 150 for the LL reward, for the no-bonus condition (purple curve) and conditions with lotteries having odds 1:0, 1:10, 1:100, and 1:1000. As with the alternative DGMDP formulation with a single-shot task and the interest-based framework, optimization of bonuses achieves benefits which depend on γ and lottery odds.

5. Experiments

We have argued that our modeling framework is flexible enough to describe a variety of delayed gratification tasks, both one shot and interative, with variable payoff and incentive structures. This framework provides a potential explanation of human cognition, under the conjecture that individuals can be cast as bounded rational agents who seek to maximize their payoffs given cognitive constraints such as discounting and fluctuations in willpower. If this conjecture is supported, the framework should allow us to determine incentives that will shape behavioral outcomes.

Typically, support for a model is obtained by comparing it to alternatives and arguing that one model is better on grounds of parsimony or predictive power. With no existing models suited to explaining the moment-to-moment dynamics of behavior, our strategy instead is to show first that the model is consistent with behavior by fitting model parameters to behavioral data, and second, that the fitted, fully constrained model can make strong predictions concerning the outcomes of subsequent experiments.

To collect behavioral data, we created a simple online delayed-gratification game in which players score points by waiting in a queue, much as diners score delicious foods by waiting their turn at a restaurant buffet (Figure 4a). The upper queue is short, having only one

3. According to prospect theory, the 1:10, 1:100, and 1:1000 lotteries yield overweighting by factors of 1.86, 5.50, and 14.40, respectively.

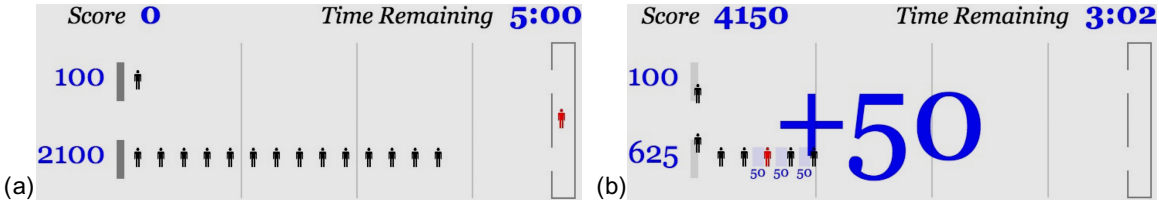

Figure 4: The queue-waiting game. (a) The player (red icon) is in the vestibule, prior to choosing a queue. Queues advance right to left. Points awarded per queue are displayed left of the queue. (b) A snapshot of the game taken while the queues advance. As described in the text, this condition includes bonuses at certain positions in the long queue. Point increments are flashed as they are awarded.

position, and delivers a 100 point reward when the player is serviced. The lower queue is long, having τ positions, and delivers a $100\tau\rho$ point reward when the player is serviced. The *reward-rate ratio*, ρ, is either 1.25 or 1.50 in our experiments. The player starts in a vestibule (right side of screen) and selects a queue with the up and down arrow keys. The game updates at 2000 msec intervals, at which point the player's request is processed and the queues advance (from right to left). Upon entering the short queue, the player is immediately serviced. Upon entering the long queue, the player immediately advances to the next-to-last position as the queue shuffles forward. With every tick of the game clock, the player may hit the left-arrow key to advance in the long queue or the up-arrow key to defect to the short queue. If the player takes no action, the simulated participants behind the player jump past. When the player defects to the short queue, the player is immediately serviced. When points are awarded, the screen flashes the points and a cash register sound is played, and the player returns to the vestibule and a new *episode* begins. In our initial experiments, the long-queue length τ is uniformly drawn from $\{4, 6, 8, 10, 12, 14\}$ for each epsiode.

Note that the reward rate (points per action) for either queue does not depend on the long-queue length. Because of this constraint, each episode is functionally decoupled from following episodes. That is, the optimal action for the current episode will not depend on upcoming episodes.[4] Due to this fact and the time-constrained nature of the game, the iterated DGMDP in Figure 1d is appropriate for describing a rational player's understanding of the game. This DGMDP focuses on reward *rate* and treats a defection as if the player continues to defect until τ steps are reached, each step delivering the small reward. In contrast to Figure 1c, Figure 1d is not concerned with the interdependence of episodes. The vestibule in Figure 4a corresponds to state 1 in Figure 1d and lower queue position closest to the service desk to state τ. Note the left-to-right reversal of the two Figures, which has often confused the authors of this article.

Participants were recruited to play the game for five minutes via Amazon Mechanical Turk. In our analyses of player behavior, we remove the first and last thirty seconds of play. At the start, players are learning the game actions; at the end, players may not have sufficient time to traverse the long queue and defection is the optimal strategy. Participants are paid

4. A dependence does occur in the final seconds of the game, where the player may not have sufficient time to complete the long queue. We handle this case by discarding data toward the end of the game.

$0.80 to play and are awarded a score-based bonus. They are required to perform at least one action every ten seconds or the experiment terminates and their data are rejected.

5.1 Experiment 1: Varying Reward Magnitude

In Experiment 1, we manipulated the reward-rate ratio. Twenty different participants were tested for each $\rho \in \{1.25, 1.50\}$. Figure 5a shows the reward accumulation by individual participants in the two conditions as a function of time within the session. The two dashed black lines represent the reward that would be obtained by deterministically performing the SS or LL action at each tick of the game clock. (Participants are not required to act every tick, but they are warned after 7 sec and rejected after 14 sec if they fail to act.) The traces show that some participants had a strong preference for the short queue, others had a nearly perfect preference for the long queue, and still others alternated between strategies. The variability in strategy over time within an individual suggests that they did not simply lock into a fixed, deterministic action sequence.

For each participant, each queue length, and each of the τ positions in a queue, we compute the fraction of episodes in which the participant defects at the given position. We average these proportions across participants and then compute empirical hazard curves. Figure 5b shows hazard curves for each of the six queue lengths and the two ρ conditions. The $\rho = 1.50$ curves are lighter and are offset slightly to the left relative to the $\rho = 1.25$ curves to make the pair more discriminable. The Figure presents both human data—asterisks connected by dotted lines—and simulation results—circles connected by solid lines. Focusing on the human data for the moment, initial-defection rates rise slightly with queue length and are greater for $\rho = 1.25$ than for $\rho = 1.50$. We thus see robust evidence that participants are sensitive to game conditions.

To model the data, we set the DGMDP parameters (Θ_{task}) based on the game configuration. We obtain least-squares fits to the four agent parameters (Θ_{agent}): discount rate $\gamma = 0.957$, initial and delta willpower spreads $\sigma_1 = 81.3$, and $\sigma = 21.3$, and effort cost $\mu_{\text{E}} = -52.1$. The latter three parameters can be interpreted using the scale of the SS reward, $\mu_{\text{SS}} = 100$ points. Although the model appears to fit the pattern of data quite well, the model has four parameters and the data can essentially be characterized by four qualitative features: the mean rate of initial defection, the modulation of the initial-defection rate based on queue length and on ρ, and the curvature of the hazard function. The model parameters have no direct relationship to these features of the curves, but the model is flexible enough to fit many empirical curves. Consequently, we are cautious in making claims for the model's validity based solely on the fit to Experiment 1. We note, however, that we investigated a variant of the model in which willpower is uncorrelated across steps, and it produces qualitatively the *wrong* prediction: it yields curves whose hazard probability depends only on the steps to the LL reward. In contrast, the curves of the correlated-willpower account depend primarily on the distance from the initial state, t, but secondarily on distance to the LL reward, $\tau - t$.

5.2 Experiments 2 and 3: Modulating Effort

To obtain additional support for the theory, we modified the queue-waiting game such that the player must work harder and experiences more frustration in reaching the front of the long queue. By increasing the required effort, we may test whether model parameters fit to

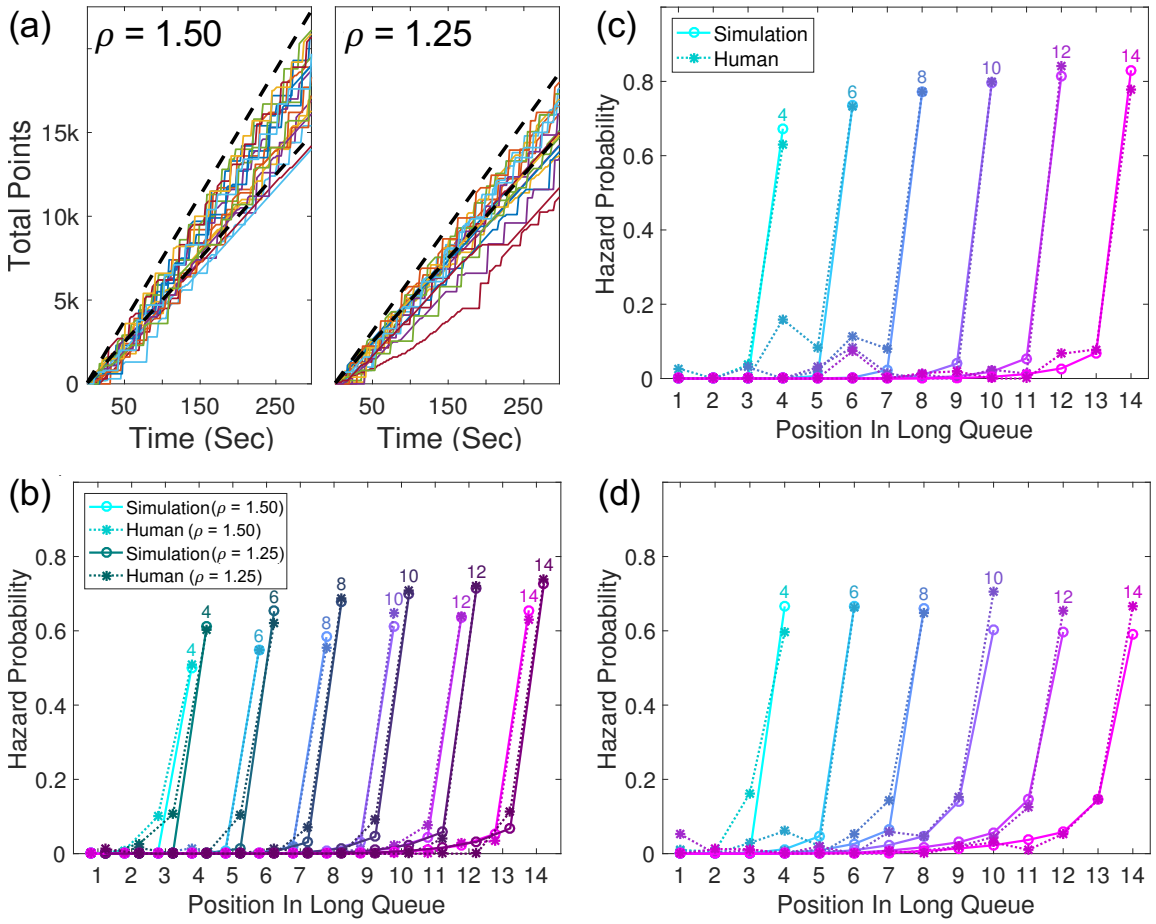

Figure 5: (a) Game points accumulated by individual participants over time in Experiment 1. (b) Hazard curves in Experiment 1 for 6 line lengths and two reward-rate ratios. Human data shown with asterisks and dashed lines, model fits with circles and solid lines. (c) Hazard curves for Experiment 2, with only one free model parameter, (d) Hazard curves for Experiment 3, with no free model parameters.

Experiment 1 will also fit new data, changing only the effort parameter, μ_{E}. To increase the required effort, the long queue advanced only every other clock tick in an apparently random fashion. Nonetheless, the player must press the advance key every tick to move with the queue, thus requiring exactly two keystrokes for each action in the game FSM (Figure 1d). The game clock in Experiment 2 updated every 1000 msec, twice the rate as in Experiment 1, and thus the overall timing was unchanged. We tested only the reward-rate ratio $\rho = 1.50$.

Figure 5c shows hazard curves for Experiment 2. Using Experiment 1 parameter settings for γ, σ_1, and σ, we fit only the effort parameter, obtaining $\mu_{\mathrm{E}} = -99.7$, which is fortuitously twice the value obtained in Experiment 1. Model fits are superimposed over the human data. To further test the theory's predictive power, we froze all four parameters and ran an Experiment 3 in which we introduced a smattering of 50 and 75 point bonuses along

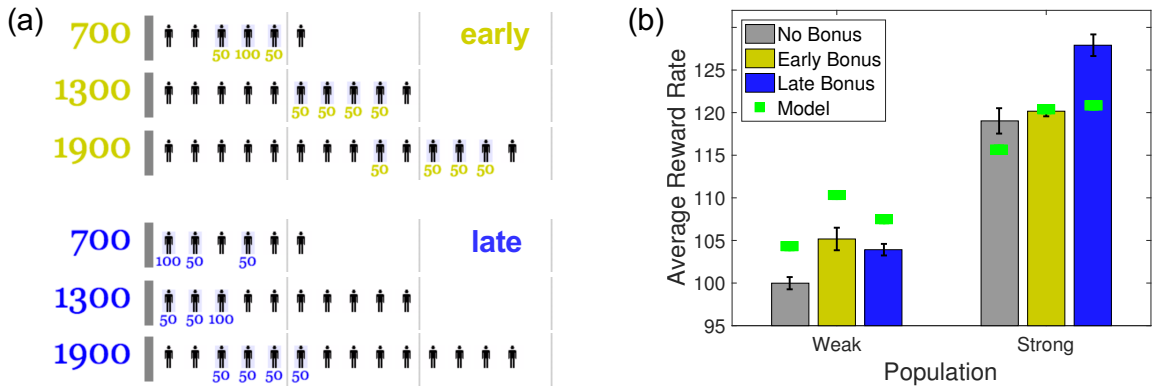

Figure 6: Experiment 4. (a) Model-predicted optimal bonus sequences—early (yellow) and late (blue) bonuses for weak and strong participants, respectively. (b) Average reward rate for weak and strong subpopulations and three bonus conditions. Error bars are ±1 SEM, corrected for between-subject variance (Masson and Loftus, 2003).

the path to the LL reward, adjusting the front-of-queue reward such that the reward-rate ratio $\rho = 1.50$ was attained when traversing the entire queue (see example in Figure 4b). Using the fully constrained model from Experiment 2, the fit obtained for Experiment 3 was quite good (Figure 5d). The model might underpredict long-queue initial defections, but it captures the curvature of the hazard functions due to the presence of bonuses.

5.3 Experiment 4: Customized Bonuses

In Experiment 4, we tested the effect of bonuses customized to a subpopulation. To situate this Experiment, we reviewed the Experiment 2 data to examine inter-participant variability. We stratified the 30 participants in Experiment 2 based on their mean reward rate per action. This measure reflects quality of choices and does not penalize individuals who are slow. With a median split, the *weak* and *strong* groups have average reward rates of 103 and 132, respectively. Theoretically, rates range from 0 (always switching between lines and never advancing) to 100 (deterministically selecting the short queue) to 150 (deterministically selecting the long queue). We fit the hazard curves of each group to a customized γ, leaving unchanged the other parameters previously tuned to the population. We obtained excellent fits to the distinctive hazard functions with $\gamma_{\text{strong}} = 0.999$ and $\gamma_{\text{weak}} = 0.875$.

We then optimized bonuses for each group for various line lengths. As in Figure 3c, we searched over a bonus space consisting of all arrangements of up-to four bonuses, each worth fifty points, allowing multiple bonuses at the same queue position.[5] We subtracted 200 points from the LL reward, maintaining a reward-rate ratio of $\rho = 1.50$ for completing the long queue. We constrained the search such that no mid-queue defection strategy would lead to

5. We avoided the interest-accrual scheme for bonuses because it could lead to variable reward rates among episodes in an interated DGMDP, which would introduce dependencies that invalidate treating the iterated DGMDP in Figure 1c as equivalent to that in Figure 1d.

$\rho > 1$. A brute-force optimization yields bonuses *early* in the queue for the weak group, and bonuses *late* in the queue for the strong group (Figure 6a).

Experiment 4 tested participants on three line lengths—6, 10, and 14—and three bonus conditions—early, late, and no bonuses. (The no-bonus case was as in Experiment 2.) The 54 participants who completed Experiment 4 were median split into a weak and a strong group based on their reward rate on no-bonus episodes only. Consistent with the model-based optimization, the weak group performs better on early bonuses and the strong group on late bonuses (the yellow and blue bars in Figure 6b). Importantly, there is a 2×2 interaction between group and early versus late bonus ($F(1, 51) = 11.82$, $p = .001$) indicating a differential effect of bonuses on the two groups. Figure 6b also shows model predictions based the parameterization determined from Experiment 2. The model has a perfect rank correlation with the data, and correctly predicts that both bonus conditions will facilitate performance, despite the objectively equal reward rate in the bonus and no-bonus conditions. That bonuses should improve performance is nontrivial: the persistence induced by the bonuses must overcome the tendency to defect because the LL reward is lower (as we observed in Experiment 1 with $\rho = 1.25$ versus $\rho = 1.50$).

6. Discussion

In this article, we developed a formal theoretical framework to modeling the dynamics of intertemporal choice. We hypothesized that the theory is suitable to modeling human behavior. We obtained support for the theory by demonstrating that it explains key qualitative behavioral phenomena (Section 2.2) and predicts quantitative outcomes from a series of behavioral experiments (Section 4). Although our first experiment merely suggests that the theory has the flexibility to fit behavioral data post hoc, each following experiment used parametric constraints from the earlier experiments, leading to strong predictions from the theory that match behavioral evidence. The theory allows us to design incentive mechanisms that steer individuals toward better outcomes (Section 3), and we showed that this idea works in practice for customizing bonuses to subpopulations playing our queue-waiting game. The theory and the behavioral evidence both show a non-obvious and non-intuitive statistical interaction between the subpopulations and various incentive schemes. Because the theory has just four free parameters, it is readily pinned down to make strong, make-or-break predictions. Furthermore, it should be feasible to fit the theory to individuals as well as to subpopulations. With such fits comes the potential for maximally effective, truly individualized approaches to guiding intertemporal choice.

This research program is still far from demonstrating utility in incentivizing individuals to persevere toward long-term goals such as losing weight or saving for retirement. It remains unclear whether intertemporal choice on a long time scale will have the same dynamics as on the short time scale of our queue-waiting game. However, the finding that reward-seeking behavior on the time scale of eye movements can be related to reward-seeking behavior on the time scale of weeks and months (Shadmehr et al., 2010; Wolpert and Landy, 2012) leads us to hope for temporal invariance.

Acknowledgments

This research was supported by NSF grants DRL-1631428, SES-1461535, SBE-0542013, SMA-1041755, and seed summer funding from the Institute of Cognitive Science at the University of Colorado. We thank Ian Smith and Brett Israelson for design and coding of the experiments.

References

R. Argento, V. L. Bryant, and J. Sabelhaus. Early withdrawals from retirement accounts during the great recession. *Contemporary Economic Policy*, 33:1–16, 2015.

D. Fernandes, J. G. Lynch, Jr., and R. G. Netemeyer. Financial literacy, financial education, and downstream financial behaviors. *Management Science*, 60:1861–1883, 2014.

Shane Frederick, George Loewenstein, and Ted O'Donoghue. Time discounting and time preference: A critical review. *Journal of Economic Literature*, 40(2):351–401, June 2002. doi: 10.1257/002205102320161311.

L. Green and J. Myerson. A discounting framework for choice with delayed and probabilistic rewards. *Psychological Bulletin*, 130:769–792, 2004.

D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47:263–292, 1979.

Melissa Schettini Kearney, Peter Tufano, Jonathan Guryan, and Erik Hurst. Making savers winners: An overview of prize-linked savings products. Working Paper 16433, National Bureau of Economic Research, October 2010. URL `http://www.nber.org/papers/w16433`.

C. Kidd, H. Palmeri, and R. N. Aslin. Rational snacking: Young children's decision-making on the marshmallow task is moderated by beliefs about environmental reliability. *Cognition*, 126:109–114, 2012. doi: doi:10.1016/j.cognition.2012.08.004.

K N Kirby. One-year temporal stability of delay-discount rates. *Psychological Bulletin & Review*, 16:457–462, 2009.

R Kivetz. The effects of effort and intrinsic motivation on risky choice. *Marketing Science*, 22:477–502, 2003.

Z. Kurth-Nelson and A. D. Redish. A reinforcement learning model of precommitment in decision making. *Frontiers in Behavioral Neuroscience*, 4, 2010. doi: http://doi.org/10.3389/fnbeh.2010.00184.

Z. Kurth-Nelson and A. D. Redish. Don't let me do that! models of precommitment. *Frontiers in Neuroscience*, 6, 2012. doi: http://doi.org/10.3389/fnins.2012.00138.

Michael EJ Masson and Geoffrey R Loftus. Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology*, 57(3):203—220, 2003.

J. T. McGuire and J. W. Kable. Rational temporal predictions can underlie apparent failure to delay gratification. *Psychological Review*, 120:395–410, 2013.

W. Mischel. Processes in delay of gratification. *Advances in Experimental Social Psychology*, 7:249–292, 1974. doi: doi:10.1016/S0065-2601(08)60039-8.

W. Mischel and E. B. Ebbesen. Attention in delay of gratification. *Journal of Personality and Social Psychology*, 16:329–337, 1970.

W Mischel, Y Shoda, and P K Peake. The nature of adolescent competencies predicted by preschool delay of gratification. *Journal of Personality & Social Psychology*, 54:687–696, 1988.

W Mischel, Y Shoda, and MI Rodriguez. Delay of gratification in children. *Science*, 244(4907):933–938, 1989. ISSN 0036-8075. doi: 10.1126/science.2658056. URL `http://science.sciencemag.org/content/244/4907/933`.

Noam Nisan and Amir Ronen. Algorithmic mechanism design (extended abstract). In *Proceedings of the Thirty-first Annual ACM Symposium on Theory of Computing*, STOC '99, pages 129–140, New York, NY, USA, 1999. ACM. ISBN 1-58113-067-8. doi: 10.1145/301250.301287. URL `http://doi.acm.org/10.1145/301250.301287`.

Y. Niv, J.A. Edlund, P. Dayan, and J.P. O'Doherty. Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *The Journal of Neuroscience*, 32:551–562, 2012.

N. Rhee and I. Boivie. The continuing retirement savings crisis. Technical report, National Institute on Retirement Security, March 2015. URL `http://www.nirsonline.org/storage/nirs/documents/RSC%202015/final_rsc_2015.pdf`.

Reza Shadmehr, Jean Jacques Orban de Xivry, Minnan Xu-Wilson, and Ting-Yu Shih. Temporal discounting of reward and the cost of time in motor control. *Journal of Neuroscience*, 30(31):10507–10516, August 2010.

Yun Shen, Michael J. Tobia, Tobias Sommer, and Klaus Obermayer. Risk sensitive reinforcement learning. *Neural Computation*, 26:1298–1328, 2014.

A. Tversky and D. Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5:279–323, 1992.

D. M. Wolpert and M. S. Landy. Motor control is decision making. *Current Opinions in Neurobiology*, 22:996–1003, 2012.

1 (2016) 1-48 Submitted 4/00; Published 10/00

Supplementary Materials
Overcoming Temptation: Incentive Design For Intertemporal Choice

Editors:

Consider the value function for a special case where the willpower does not fluctuate, i.e., $\sigma^2 = 0$ and where intermediate rewards are not provided, i.e., $\mu_i = 0$ for $i \in \{1...\tau - 1\}$. In this case, I can show that the solution to the DGMDP in Figure 1b is identical to the solution to the DGMDP in Figure 1d.

We need to extend this result to the following more general cases, roughly in order of challenge:

- Allow for nonzero intermediate rewards
- Allow for the case of Figure 1c where $\mu_{\text{LLa}}/\tau_a = \mu_{\text{LLb}}/\tau_b$ for all a and b,
- Allow for the case where $\sigma^2 > 0$

1. Proof of $\sigma^2 = 0$ and $\mu_i = 0$ case

In Figure 1b, the value of state 1 is defined by the Bellman equation as:

$$V(1) = \max(\mu_{\text{SS}} + \gamma V(1), \gamma^{\tau-1}[\mu_{\text{LL}} + \gamma V(1)]) \tag{1}$$

We can solve for $V(1)$ if the first term is larger:

$$V_{\text{SS}}(1) = \frac{1}{1-\gamma}\mu_{\text{SS}}. \tag{2}$$

We can solve for $V(1)$ if the second term is larger:

$$V_{\text{LL}}(1) = \frac{\gamma^{\tau-1}}{1-\gamma^{\tau}}\mu_{\text{LL}}. \tag{3}$$

Now consider Figure 1d, whose Bellman equation can be simplified to:

$$V(1) = \max\left(\sum_{i=0}^{\tau-1}\gamma^i\mu_{\text{SS}}, \gamma^{\tau-1}\mu_{\text{LL}}\right) \tag{4}$$

$$= \max\left(\frac{1-\gamma^{\tau}}{1-\gamma}\mu_{\text{SS}}, \gamma^{\tau-1}\mu_{\text{LL}}\right) \tag{5}$$

$$= (1-\gamma^{\tau})\max\left(\frac{1}{1-\gamma}\mu_{\text{SS}}, \frac{\gamma^{\tau-1}}{1-\gamma^{\tau}}\mu_{\text{LL}}\right). \tag{6}$$

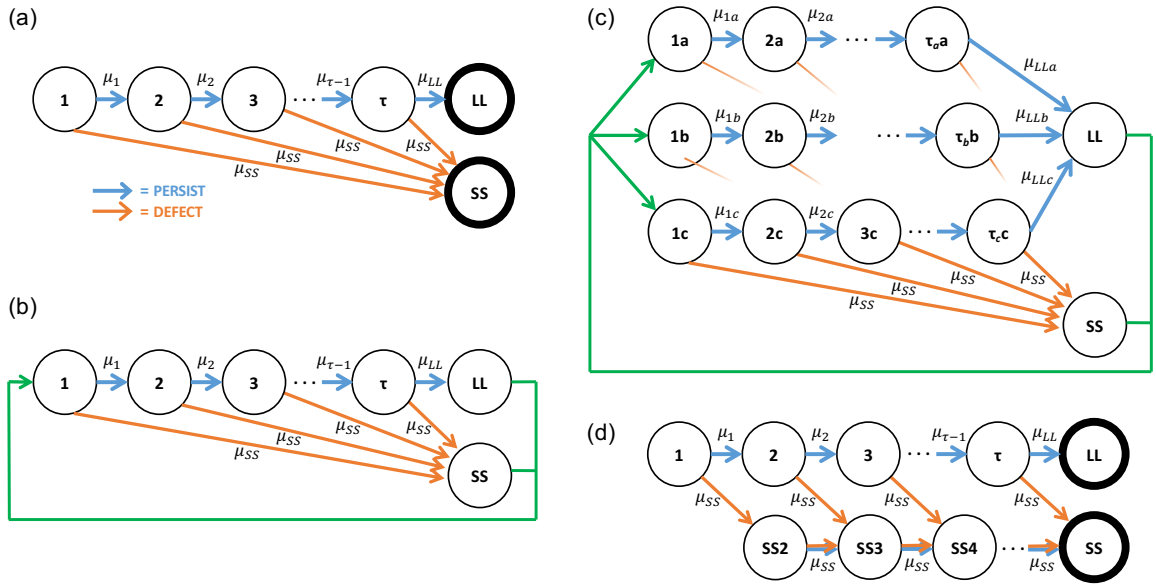

Figure 1: Finite-state environment formalizing (a) the one-shot delayed-gratification task; (b) the iterated delayed-gratification task; (c) the iterated delayed-gratification task with variable delays and LL outcomes; and (d) an efficient approximation to the iterated delayed-gratification task, suitable when episodes are independent of one another.

Note that the two terms inside the max function of Equation 6 are identical to the values in Equations 2 and 3, and thus the value functions for Figures 1b and 1d are identical up to a scaling constant.