Simulating bout-and-pause patterns with reinforcement learning

Kota Yamada

Keio University & LeapMind Inc.

Atsunori Kanemura

LeapMind Inc.

Author Note

Correspondence should be addressed to K. Yamada, Department of Psychology, Keio University, Tokyo 108-8345, Japan (Email: haroldthebarrel.yk@gmail.com).

Abstract

Animal responses occur according to a specific temporal structure composed of two states, where a bout is followed by a long pause until the next bout. Such a bout-and-pause pattern has three components: the bout length, the within-bout response rate, and the bout initiation rate. Previous studies have investigated how these three components are affected by experimental manipulations. However, it remains unknown what underlying mechanisms cause bout-and-pause patterns. In this article, we propose two mechanisms and examine computational models developed based on reinforcement learning. The first mechanism is choice—an agent makes a choice between operant and other behaviors. The second mechanism is cost—a cost is associated with the changeover of behaviors. These two mechanisms are extracted from past experimental findings. Simulation results suggested that both the choice and cost mechanisms are required to generate bout-and-pause patterns and if either of them is knocked out, the model does not generate bout-and-pause patterns. We further analyzed the proposed model and found that it reproduced the relationships between experimental manipulations and the three components that have been reported by previous studies. In addition, we showed alternative models can generate bout-and-pause patterns as long as they implement the two mechanisms.

Keywords: bout-and-pause patterns, reinforcement learning, choice, cost

Simulating bout-and-pause patterns with reinforcement learning

Introduction

Animals engage in various activities in their daily lives. For humans, they may be working, studying, practicing sports, or playing video games. For rats, they may be grooming, foraging, or escaping from a predator. Although specific activities are different between different species, common behavioral features are often observed.

Bout-and-pause patterns are one of the behavioral features commonly observed in many species. Activities engaged by an animal do not occur uniformly through time but often have short periods in which a burst of engaged responses is observed. For example, in an operant conditioning experiment, a rat presses a lever repeatedly in a short period and then it stops lever pressing. After a moment, the rat starts lever pressing again. The rat switches between the lever pressing behavior and the no lever pressing behavior again and again throughout the experiment. Such a temporal structure comprising of short-period response bursts and long pauses is observed in various species and activities; for example, email and letter communication by humans (Barabasi, 2005), foraging by cows (Tolkamp & Kyriazakis, 1999), and walking by Drosophila (Sorribes, Armendariz, Lopez-Pigozzi, Murga, & de Polavieja, 2011).

Shull, Gaynor, and Grimes (2001) showed that bout-and-pause patterns can be described with a broken-stick shape in the log-survivor plot of interresponse times (IRTs), which are characterized by a bi-exponential probability model. If IRTs follow a single exponential distribution, then the log-survivor plot shows a straight line. If IRTs follow a mixture exponential distribution called a bi-exponential model, the log-survior plot shows a broken-stick shape composed of two straight lines that have different slopes. Killeen, Hall, Reilly, and Kettle (2002) found that lever pressing by rats is well described with a bi-exponential model, suggesting that this behavior has a bout-and-pause pattern. If IRTs follow a bi-exponential distribution, there are two different types of responses; within-bout responses, which have short IRTs, and between-bout responses, which have long IRTs. Each response type has its own exponential distribution in a bi-exponential model. Killeen et al. (2002) formulated the

bi-exponential model as follows:

$$p(IRT = \tau) = (1 - q)\omega e^{-\omega\tau} + qbe^{-b\tau}, \tag{1}$$

where the first term describe IRTs of within-bout responses and the second term describes IRTs of between-bout responses. This model has three free parameters: q, ω, and b, each of which corresponds to a different component in bout-and-pause patterns. First, q denotes the mixture ratio of the two exponential distributions in the model and it corresponds to the length of a bout. The bout length is the number of responses contained in one bout. Second, ω denotes the rate parameter for the exponential distribution of within-bout IRTs and it corresponds to the within-bout response rate. Finally, b denotes the rate parameter for the exponential distribution of between-bout IRTs and it corresponds to the bout initiation rate. These three model parameters define the overall response rate. They are also called bout components.

The bout length, the within-bout response rate, and the bout initiation rate are affected by motivational and schedule-type manipulations (Brackney, Cheung, Neisewander, & Sanabria, 2011; Podlesnik, Jimenez-Gomez, Ward, & Shahan, 2006; Shull, 2004; Shull et al., 2001; Shull, Grimes, & Bennett, 2004; Tanno, 2016). Motivational manipulations include the reinforcement rate, the response-reinforcement contingency, and the deprivation level. An example of schedule-type manipulations is adding a small variable ratio (VR) schedule in tandem to a variable interval (VI) schedule. Table 1 summarizes existing findings on the relationships between experimental manipulations and the two bout components, where we omit the within-bout response rate since we will not study this bout component because of a model complexity issue and insufficient experimental findings on it. The bout length was reported to be affected by manipulations as follows:

- It increases or stays the same as the reinforcement rate increases (Shull et al., 2001; Shull et al., 2004).
- It increases or stays the same as the deprivation level increases (Podlesnik et al., 2006; Shull, 2004; Shull et al., 2001).

- It decreases or stays the same by extinction (Brackney et al., 2011; Cheung, Neisewander, & Sanabria, 2012).
- It increases by tandem VR (Shull et al., 2001; Shull et al., 2004). When a VI schedule is followed by a small VR (tandem VI VR), an animal stays in a bout longer and emits more responses in each bout.

The bout initiation rate was reported to be:

- It increases as the reinforcement rate increases (Shull et al., 2001; Shull et al., 2004).
- It increases as the deprivation level increases (Podlesnik et al., 2006; Shull, 2004; Shull et al., 2001).
- It decreases by extinction (Brackney et al., 2011; Cheung et al., 2012).
- It decreases or stays the same by tandem VR (Brackney et al., 2011). Brackney et al. (2011) showed that if we add a small VR schedule in tandem to a VI schedule, the bout initiation rate decreased slightly.

Although the previous studies have investigated the relationships between some experimental manipulations and the bout components, we still do not know how to construct a model that generate bout-and-pause patterns based on the experimental findings. Smith, McLean, Shull, Hughes, and Pitts (2014) showed experimentally that choice and cost play important roles in organizing responses into bout-and-pause patterns. When pigeons were trained under a single schedule, the log-survivor plot did not show a broken-stick shape (Bennett, Hughes, & Pitts, 2007; Bowers, Hill, & Palya, 2008). Smith et al. (2014) trained pigeons under a concurrent VI VI schedule with and without a changeover delay (COD). When pigeons were trained under the concurrent VI VI schedule without a COD, the log-survivor plot still did not show a broken stick, resulting in the a straight line. However, under the concurrent VI VI schedule with a COD, the log-survivor plot showed a broken stick, indicating that bout-and-pause patterns were clearly observed. From these experimental observations, we extracted the following three facts. 1) When animals engage only in one response in a given situation, bout-and-pause patterns are not observed. 2) If animals can choose responses from two

alternatives without a COD, bout-and-pause patterns are still not observed.
3) Considering 1) and 2), we conclude that bout-and-pause patterns are organized only when animals have two (or more) possible alternatives under a given situation (i.e., choice is available) and there is a COD between the start of engagement and a reinforcement (i.e., cost is associated with a changeover). These facts are interesting but they remain to be inductive and we still do not have constructive explanation that generate bout-and-pause patterns. Existing studies on bout-and-pause patterns have investigated to describe the phenomena rather than to provide constructive models. Although many models have been proposed (Killeen et al., 2002; Matsui, Yamada, Sakagami, and Tanno, 2018; Tanno, 2016), they are descriptive and did not answer the question of "what mechanisms shape responses into bout-and-pause patterns?"

Kulubekova and McDowell (2008) examined a computational model aimed to be an account of bout-and-pause patterns based on the principle of selection by consequences developed by J. J. McDowell (2004), but they did not explain what mechanisms bring bout-and-pause response patterns happen. Furthermore, it was not conclusive that their model reproduced bout-and-pause patterns, which can be seen from the following two arguments. First, as they discussed, the log-survivor plot generated by their model was similar but different from the existing experimental findings and did not show a clear distinction between within- and between-bout IRTs. Second, they fitted a bi-exponential model to the simulated IRTs but estimated parameters, q and b, were far from the values that live animals have. These facts suggest a reasonable question that their model might lack essential parts behind bout-and-pause patterns, the choice and cost, and it was the cause of their inconclusive results to reproduce bout-and-pause patterns.

In this article, we propose a computational model based on reinforcement learning that accounts for the constructive mechanism of bout-and-pause patterns. We assume that bout-and-pause patterns are generated by two mechanisms: a choice between operant and other behaviors and a cost that is required to a transition from one

behavior to another. We suppose that motivational manipulations affect only the choice mechanism and schedule-type manipulations affect the cost mechanism. To incorporate those two mechanisms, we design a three-state Markov transition model, which has an extra state in addition to the bout and pause states. We perform three simulation studies to analyze the proposed model. In Simulation 1, we introduce our model on the basis of the two different mechanisms, choice and cost. We show that the proposed model can reproduce bout-and-pause patterns by finding that the log-survivor plot shows a broken-stick shape. We compare three models: a dual model, a no cost model, and a no choice model. The dual model is composed of both the choice and cost mechanisms. The no cost model has only the choice mechanism and the no choice model has only the cost mechanism. Simulation results demonstrate that the dual model can reproduce bout-and-pause patterns but the other two models failed to reproduce them. It implies that both of choice and cost are required for animal responses to be organized into bout-and-pause patterns. In Simulation 2, we analyze the dual model in depth and report its behavior under various experimental settings to test if the dual model can reproduce the relationships between the experimental manipulations and the bout components discovered so far. Simulation results suggest that the dual model can reproduce them not only qualitatively but also quantitatively. In Simulation 3, we show that a two-state model can also reproduce bout-and-pause patterns even without the third state because it incorporates the two mechanisms. However, having the third state is useful for separating the effects of the choice and cost mechanisms. We speculate that real animals might have similar mechanisms that generate bout-and-pause patterns as the dual model, which can be a useful computational tool for studying animal behavior.

Simulation 1

Method: Simulation 1

Model. Our model is based on reinforcement learning (Sutton & Barto, 2018), whose core concepts and algorithms used in this paper is described in Appendix A. We designed a three-state Markov process for modeling bout-and-pause patterns

(Figure 1(a)). Two of the three states are "Operant" and "Others," in which the agent engages in the operant behavior or performs other behaviors, respectively. We call them Operant and Others instead of engagement or visit and disengagement or pause, thereby we emphasize that bout-and-pause patterns are results of a choice between the operant and other behaviors. In the third "Choice" state, the agent makes a decision between the operant and other behaviors. By having the Choice state in our model, we incorporate the knowledge that animals can choose their behavior from available options (e.g. grooming, exploration, and excretion) when they move freely during an experiment. The second knowledge that is observed in animal behavior is a cost required to make a transition from one behavior to another. Animals must decide whether to keep doing the same behavior or to make a transition, because a fast switching is not optimal if a transition incurs a cost. Figure 1(b) and (c) shows two knockout models, the no choice model and the no cost model, respectively. In each model, one of the two mechanisms from the dual model is removed. In the no choice model, an agent can choose only the operant behavior in a given situation. In the no cost model, no cost is required when a transition is made.

Here is how the agent travels in the proposed model. In the Choice state, the agent chooses either the operant or other behaviors. As a result of the choice, it moves from the Choice state to one of the Operant or Others states. It makes the choice based on the preference for each behavior, which is denoted by Q_{pref}. We will explain how to calculate Q_{pref} in the next paragraph. In the Operant state, the agent engages in the operant behavior, and, after every response, it decides whether to stay in the Operant state or to move back to the Choice state. It decides to stay or move based on Q_{cost}, which represents a transition cost to the Choice state, whose mathematical definition will be given later in this Model section. The Others state is the same as the Operant state except for that the agent performs other behaviors.

The preference Q_{pref} is a function that compares the operant and other behaviors when the agent makes a choice between them. The Q_{pref} function changes over time since it is updated based on the presence (or absence) of a reinforcer per bout. The

following equation describes the updating rule for Q_{pref}:

$$Q_{\text{pref}}^{(i)}(t+1) = \begin{cases} Q_{\text{pref}}^{(i)}(t) + \alpha_{\text{rft}}(r^{(i)}(t) - Q_{\text{pref}}^{(i)}(t)), & \text{if a reinforcer is presented,} \quad (2) \\ Q_{\text{pref}}^{(i)}(t) + \alpha_{\text{ext}}(0 - Q_{\text{pref}}^{(i)}(t)), & \text{otherwise,} \quad (3) \end{cases}$$

where t denotes time in session; α_{rft} and α_{ext} denotes the learning rates of reinforcement and extinction, respectively; r denotes the reinforcer value and we assume $r > 0$ when a reinforcer is present and $r = 0$ when a reinforcer is absent; and $i \in \{\text{Operant}, \text{Others}\}$ denotes each option, that is, $i = \text{Operant}$ if the operant behavior is chosen and if $i = \text{Others}$ if other behaviors are chosen. We omit superscript (i) and denote Q_{pref} when it can be any of $i = \text{Operant}$ or Others.

In the Choice state, the agent chooses either of the Operant or Others states according to the probability distribution calculated from the preferences for the two behaviors. The probability of transition to option $i \in \{\text{Operant}, \text{Others}\}$ is defined as follows:

$$p_i = \frac{\exp\{\beta Q_{\text{pref}}^{(i)}(t)\}}{\sum_{i \in \{\text{Operant,Others}\}} \exp\{\beta Q_{\text{pref}}^{(i)}(t)\}}, \quad (4)$$

where the softmax inverse temperature parameter β represents the degree to which a choice is focused on the highest-value option.

The cost Q_{cost} is a function that defines a barrier in making a transition from the performed behavior to the Choice state. We assumed that the cost is independent from the preference and depends only on the number of responses that are emitted to obtain a reinforcer from a bout initiation. When a reinforcer is presented, the cost function Q_{cost} is updated according to

$$Q_{\text{cost}}^{(i)}(t+1) = Q_{\text{cost}}^{(i)}(t) + \alpha_{\text{rft}}(\log x^{(i)}(t) - Q_{\text{cost}}^{(i)}(t)), \quad (5)$$

where x denotes the number of responses that are emitted to obtain a reinforcer in a bout . Then, x is initialized to 1 when the agent receives a reinforcer or comes back to the Choice state without a reinforcer. The other parameters are the same as Equations 2 and 3. The same (i)-omitting rule applies also to Q_{cost}.

If the agent is in either of the Operant or Others states, it makes a decision whether to stay in the same state or to go back to the Choice state. A decision is made

according to the probability of staying in the same state calculated from the cost and the preference for the state, which is defined as follows:

$$p_{\text{stay}}^{(i)} = \exp\left\{\frac{-1}{w_{\text{pref}}Q_{\text{pref}}^{(i)}(t) + w_{\text{cost}}Q_{\text{cost}}^{(i)}(t)}\right\}, \tag{6}$$

where w_{pref} and w_{cost} are positive weighting parameters for Q_{pref} and Q_{cost}, respectively. We assumed $w_{\text{cost}} > w_{\text{pref}}$ because schedule-type operations have stronger effects on the bout length than motivational manipulations.

Simulation. In Simulation 1, we compared the three possible models; the dual model, the no choice model, and the no cost model. The dual model (Figure 1(a)) includes both the choice and cost mechanisms as we described in the Model section. The second model was the no choice model (Figure 1(b)), which has only the cost mechanism and it can be thought of as a model made by removing the choice mechanism from the dual model. In the no choice model, the agent only engages in the operant behavior. In other words, this model chooses only the operant behavior in the Choice state. The third model was the no cost model, which has only the choice mechanism without the cost mechanism. The no cost model chooses either operant or other behavior independent of the previous behavior; that is, according to this model, the agent does not continue to be in the same state and comes back to the Choice state after each response.

Simulation conditions were as follows. The schedule for the operant behavior was VI 120 s (0.5 reinforcer per min) without an inter-trial interval, and the schedule for the other behavior was FR 1. The maximum number of reinforcers in the Operant state was 1,000; that is, if the number of reinforcers reached 1,000, the simulation was terminated. The value of a reinforcer given by taking the operant behavior was $r^{(\text{Operant})} = 1.0$ and that by taking other behaviors was $r^{(\text{Others})} = 0.5$. The model parameters were α_{rft}, α_{ext}, β, w_{pref}, and w_{cost}. We set $\alpha_{\text{rft}} = 0.05$, $\alpha_{\text{ext}} = 0.01$, $\beta = 12.5$, $w_{\text{pref}} = 1.0$ and $w_{\text{cost}} = 3.5$. The response probabilities in the Operant and the Others states were fixed at 1/3 in each time step. Before the start of the simulation, we initialized the agent and the experimental environment. The initial values of $Q_{\text{pref}}^{(i)}$ and $Q_{\text{cost}}^{(i)}$ were both 0 and we created a VI table according to Fleshler and Hoffman (1962). We set the time step in

the simulation to be 0.1 s.

We show pseudocode of the model and simulation in Algorithm 1, where *NumResponses* means x in Equation 5 and the three *Behavior*() functions are defined in Algorithms 2, 3, and 4. We implemented the algorithm in Julia 1.0 and ran simulations on a computer with a 1.80 GHz Intel i7-8565 processor, 16 GB of RAM, and 1 TB of SSD, operating with Ubuntu 18.04 LTS. The same configuration was used also for Simulations 2 and 3. The Julia code is available at: https://github.com/echo0yasum1/simulating_bout_and_pause_pattern.

Results: Simulaton 1

Figure 2(a) shows event records of IRTs generated by each model and Figure 3 shows the model schemes with transition probabilities. The top panel of Figure 2(a) shows that the no choice model generated a dense repetition of only the operant behavior at a high rate without long pauses. From Figure 3, the probability the agent stayed in the Operant state was empirically 0.95. In the middle panel of Figure 2(a), the response rate under the no cost model was low and each response was separated by long pauses. From Figure 3, the probability of the agent choosing to transit to the Operant state was empirically 0.06 and the agent returned to the Choice state immediately after it responded. In the bottom panel of Figure 2(a), the agent with the dual model generated a repetitive pattern of responses with a high rate in a short period followed by a long pause. From Figure 3, the agent in the Choice state made a transition to the Operant state with a 0.12 probability and it stayed in the Operant state with a 0.71 probability.

Figure 2(b) show log-survivor plots to see whether they show a straight line or a broken stick. We used the IRTs from after the agent obtained 500 reinforcers to the end of the simulation. The log-survivor plots of the no choice model and the no cost model were described by one straight line whereas that of the dual model was described with a broken-stick shape. The no choice model has a steeper slope than the no cost model and is tangential to the curve of the dual model at the leftmost position. The no cost

Algorithm 1 Pseudocode of simulation

```
t ← 0, NumRewards ← 0, ResponseTimes ← {}, i ← Choice
while NumRewards < 1000 do
    t ← t + 0.1
    if i = Choice then
        ChoiceBehavior()
    end if
    if i = Operant and uniform(0, 1) ≤ 1/3 then
        OperantBehavior()
    end if
    if i = Others and uniform(0, 1) ≤ 1/3 then
        OthersBehavior()
    end if
end while
```

Algorithm 2 Definition of *ChoiceBehavior*()

```
Select a state i ∈ {Operant, Others} with probability defined by Equation 4
NumResponses ← 1
```

Algorithm 3 Definition of *OperantBehavior*()

```
Append t to ResponseTimes
NumResponses ← NumResponses + 1
Select a state i ∈ {Operant, Choice} with probability defined by Equation 6
if reward is presented then
    Update Q_pref^(Operant)(t) according to Equation 2
    Update Q_cost^(Operant)(t) according to Equation 5
    NumRewards ← NumRewards + 1
    NumResponses ← 1
end if
if reward is absent then
    if i = Choice then
        Update Q_pref^(Operant)(t) according to Equation 3
    end if
end if
```

Algorithm 4 Definition of *OthersBehavior*()

```
NumResponses ← NumResponses + 1
Update Q_pref^(Others)(t) according to Equation 2
Update Q_cost^(Others)(t) according to Equation 5
reward is presented according to FR 1
NumResponses ← 1
Select a state i ∈ {Others, Choice} with probability defined by Equation 6
```

model slope was slightly steeper than that of the dual model at the right side.

Discussion: Simulaton 1

Both the event records and log-survivor plots in Figure 2 imply that only the dual model generated bout-and-pause patterns and the other two models failed to reproduce bout-and-pause patterns. The event records in Figure 2(a) suggests that only the dual model exhibit bout-and-pause patterns. The log-survivor plot of only the dual model in Figure 2(b) showed not a straight but a broken-stick shape, which is an evidence that the underlying IRTs follow a bi-exponential distribution. Thus, only the dual model reproduced bout-and-pause patterns.

We posit both of the choice and cost mechanisms are necessary to organize responses into bout-and-pause patterns. The no choice model failed because it lacks the choice mechanism. Without the choice mechanism, the agent almost always stayed in the Operant state and responded at a high rate without pauses. The reason behind the failure of the no cost model was the knockout of the cost mechanism. When the cost of a changeover is zero, the agent easily return to the Choice state, resulting in sparing operant responses followed by long pauses. Similar behaviors were observed in pigeons under a concurrent VI VI schedule without COD (Smith et al., 2014). The choice and cost mechanisms contribute differently to generate bout-and-pause patterns; the choice mechanism generates pauses and the cost mechanism produces response bursts. Since the dual model has both the mechanisms, it reproduced bout-and-pause patterns.

Since we have a full control of the simulation environment and the agent in it, we can exclude the possibility of contamination by other factors. Smith et al. (2014)'s results implied that choice and cost are behind bout-and-pause patterns but it was not clear if other factors influence the formation of bout-and-pause patterns; this is an inherent limitation of experimental studies. It was not straightforward to draw conclusions like "these mechanisms are enough to generate bout-and-pause patterns" from the experimental findings that IRT distributions observed in pigeons followed a bi-exponential distribution under concurrent VI VI schedules with a COD. In contrast,

our constructive approach makes it clear that the two mechanisms are sufficient to reproduce bout-and-patterns, and this conclusion is hard to draw only from the experimental findings from Smith et al. (2014).

We suggest that what is important for generating bout-and-pause patterns is not the specific architecture of our model but the choice and cost mechanisms. Our model is composed of three states and five equations, and those equations are from one of most popular reinforcement algorithms called Q-learning. Even if such model architecture and algorithm are substituted with others, the new model will still reproduce bout-and-pause patterns if it involves the choice and cost. We do not reject other possible forms to implement the two mechanisms.

We also do not claim the uniqueness of our experimental settings. Although we employed an FR 1 schedule for the other behaviors, other schedules including VI should produce similar results.

Simulation 2

Having demonstrated in Simulation 1 that the dual model successfully reproduced bout-and-pause patterns, in Simulation 2 we analyzed this model under various environments. The previous studies (Brackney et al., 2011; Podlesnik et al., 2006; Shull, 2004; Shull et al., 2001, 2002; Shull et al., 2004; Tanno, 2016) have applied various experimental manipulations to animals to understand bout-and-pause patterns, as summarized in Table 1. We applied manipulations to the agent in the model by changing environmental settings.

Method: Simulation 2

Using the dual model, we performed four experiments by manipulating only one of the four variables while keeping the other three variables the same as Simulation 1. The procedure of simulation was also the same as Simulation 1.

The four experimental manipulations are applied independently to each of the four variables: 1) the rate of reinforcement, 2) the deprivation level, 3) the presence of extinction, and 4) the schedule type. 1) We manipulated the rate of reinforcement by

varying mean intervals of the VI schedule. Mean intervals used in this simulation were VI 30 s, 120 s, and 480 s (2.0, 0.5, and 0.125 reinforcer per min). 2) We varied reward values obtained in the Operant state to control the deprivation level of the agent. Those values were 0.5, 1.0, and 1.5 to induce low deprivation, baseline, and high deprivation levels, respectively. The reward value that the agent received by taking other behaviors was the same as Simulation 1 throughout all the simulations. 3) To attenuate the engagement to the operant response, we switched the schedule from VI 120 s (0.5 reinforcer per min) to extinction after the agent obtained 1,000 reinforcers. The extinction phase finished when 3,600 s (36,000 time steps) elapsed. 4) We manipulated the schedule type by adding a small VR schedule in tandem to a variable time (VT) schedule. The mean interval of the VT schedule was fixed to 120 s and VR values were 0, 4, and 8.

When we analyzed the IRTs data from the extinction simulation, we used a dynamic bi-exponential model (Brackney et al., 2011), in which the model parameters, q, ω, and b, are time-dependent and Equation 1 is rewritten as follows:

$$p(IRT = \tau) = (1 - q_t)\omega_t e^{-\omega\tau} + q_t b_t e^{-b_t\tau}. \tag{7}$$

Extinction causes exponential decay of the model parameters according to the following equations:

$$1 - q_t = (1 - q_0)e^{-\gamma t}, \tag{8}$$

$$b_t = b_0 e^{-\delta t}, \tag{9}$$

where the parameters γ and δ denote the decay rates of q and b, respectively. Since the decay of any of the three model parameters q, b, and ω can cause extinction, we need to identify which of these parameters actually decayed during the extinction simulation. We excluded ω because it was fixed to 1/3 during the simulation. To identify whether one or both of the q and/or b parameters decayed, we compared three models, that is, the qb-decay, q-decay, and b-decay models. We calculated WAIC (widely applicable information criterion; Watanabe (2010)) for each model. The same configuration as Simulation 1 was used in Simulation 2.

To examine the molar relationship between the reinforcement rate and response rate, we fitted Herrnstein's hyperbola (Herrnstein, 1970) to the simulated data. We used its modern version (J. McDowell, 2005, Equation 6),

$$R = \frac{kr^a}{r^a + r_{\mathrm{e}}^a/b}, \tag{10}$$

where R is the response rate, r is the reinforcement rate, r_{e} is the external reinforcement rate, k is the total amount of behavior, and a and b are the exponent and bias parameters, respectively. Since the parametrization of term r_{e}^a/b is redundant, we did not fit r_{e} and b separately and estimated only r_{e}^a/b.

Results: Simulation 2

Figure 4 shows the log-survivor plots of IRTs from each of the four simulations. Figure 4(a) and (b) shows that manipulating the rate of reinforcement or the deprivation level changed the slope and intercept of the right limb. As the rate of reinforcement or the deprivation level increased, the slope of the right limb became steeper, indicating that the bout initiation rate became larger. The broken sticks in Figure 4(c) have different slopes and y-axis intercepts, suggesting that both the bout initiation rate and the bout length were changed. Figure 4(d) shows that adding the tandem VR schedule to the VT schedule affected only the y-axis intercept of the right limb without changing its slope. As the required response increased from the baseline to VR 4 or VR 8, the bout length became larger. However, the right limbs were not stable and we performed a fitting analysis described in the next paragraph.

Table 2 shows estimated parameters of the bi-exponential model, q, ω, and b in three simulations except for extinction. Parameter q increased as the reinforcement rate, the deprivation level, and the number of required responses increased. Parameter ω did not change in all manipulations. Parameter b increased as the rate of reinforcement and the deprivation level increased.

Figure 5 shows the relationship between the reinforcement rate and response rate in our model. The response rate increased with diminished gradients, converging to

$k = 186$. The other parameters were fitted to be $a = 2.15$, and $r_{\mathrm{e}}^{a}/b = 6.59$. The percentage of variance accounted for (%VAF) was 99.5.

In Figure 4(c), the total number of IRTs during the extinction phase was insufficient to reliably estimate the right limb. Then, we analyzed the dynamic bi-exponential model fitted to the IRTs during extinction. Table 3 shows the WAIC values for the three models. The smallest WAIC was attained by the *qb*-decay model, but the differences from the other models are not large and it is not conclusive which of the bout initiation rate and the bout length decayed during extinction.

Figure 6 shows the boxplots of Q_{pref} and Q_{cost} in the three simulations except for extinction, which are to be used for assessing how the changes in the bout components are mediated. We excluded the extinction simulation because we already knew that Q_{pref} causes the change of the bout components since Q_{cost} is fixed during the extinction phase. The top panel shows that Q_{pref} and Q_{cost} increased as the rate of reinforcement increased. The middle panel indicates that increasing the deprivation level moved Q_{pref} and Q_{cost} upward. From the bottom panel, we can see that adding tandem VR schedule increased Q_{cost} without affecting Q_{pref}.

Discussion: Simulation 2

In Simulation 2, we tested whether the dual model has the same characteristics as animals reported by the previous studies. We analyzed the model with four experimental manipulations: the rate of reinforcement, the deprivation level, the presence of extinction, and the schedule type. The rate of reinforcement, the deprivation level, and the presence of extinction affected the bout initiation rate and the bout length and adding the tandem VR schedule to the VT schedule affected only the bout length.

Table 4 summarizes the relationship between the experimental manipulations and the bout components observed in the dual model, which suggests that the behaviors of the dual model are consistent with the existing knowledge on animal behaviors. Furthermore, we made stable predictions to the cells with the question marks in Table 1. Our predictions are stable because our results can be easily reproduced and

tested using the same simulation code. In contrast, experimental studies with animals could report different conclusions. Although our model does not implement Herrnstein's hyperbola a priori, the molar relationship between the reinforcement rate and response rate is well described by the modern matching theory (Figure 5). Cheung et al. (2012) and Brackney, Cheung, Herbst, Hill, and Sanabria (2012) showed that the bout initiation rate and the bout length decayed during extinction. Table 3 shows that parameter selection for dynamic bi-exponentila model with WAIC but the differences between each model are small. However, the lowest WAIC model is consistent with previous studies. Therefore, the dual model satisfies at least the necessary conditions to be a model to be analyzed for the generation mechanism of bout-and-pause patterns.

Table 2 showed estimated parameters of the bi-exponential model in each simulation and they are consistent with the parameters of previous study with real animals.

How Q_{pref} and Q_{cost} depend on experimental manipulations can be understood according to the categorization of motivational and schedule-type manipulations proposed by Shull et al. (2001). In our simulations, manipulating any of the three motivational variables, i.e. the rate of reinforcement, the deprivation level, or extinction, changed Q_{pref} and Q_{cost}. The change of Q_{cost} was not a primary but a secondary effect because Q_{cost} was changed as a result of the increased Q_{pref}; with a higher Q_{pref}, the agent emits more responses. The schedule type manipulation affected only Q_{cost}. These changes of Q_{pref} and Q_{cost} are consistent with what was proposed by Shull et al. (2001).

The dual model is limited to reproduce only some of the previous findings. Here are three examples of limitations. First, our model is not designed for analyzing the addition of a tandem VR schedule to a VT schedule, by which Tanno (2016) and Matsui et al. (2018) found the change of the within-bout response rate, which was fixed in our model. Second, the value and the delay between a response and a reinforcer were fixed in our model. Brackney et al. (2011) and Podlesnik et al. (2006) considered a delayed reinforcement from a bout initiation causes the inverse correlation between the bout initiation rate and the bout length. This result can be reproduced if required

responses by tandem VR is very high (more than 32). Third, sometimes the bout length does not decrease during extinction (Brackney et al., 2011). Our dual model could not reproduce this result even if we changed the model parameters.

Simulation 3

In Simulation 3, we examined a two-state model that incorporates the choice and cost mechanisms to examine the possibility of alternative models, particularly a simpler one. We built a two-state model without the Choice state and ran simulations with it.

Method: Simulation 3

Figure 7 shows the two-state model comprising of the Operant and Others states. Although it does not have the Choice state, the choice mechanism is implemented as the transition between the Operant and the Others state. The probability of staying at the same state is defined as follows.

$$p(s_{t+1} = i | s_t = i) = \frac{\exp(w_{\text{pref}} Q^{(i)}_{\text{pref}}(t) + w_{\text{cost}} Q^{(i)}_{\text{cost}}(t))}{\sum_i \exp(w_{\text{pref}} Q^{(i)}_{\text{pref}}(t) + w_{\text{cost}} Q^{(i)}_{\text{cost}}(t))}, \tag{11}$$

where w_{pref} and w_{cost} are positive weights for Q_{pref} and Q_{cost}, respectively. Updating rules for Q_{pref} and Q_{cost} are the same as Equations 2, 3 and 5. The parameters of the two-state model were sought in the ranges shown in Table 6, which includes the parameter values used for the three-state, dual model. The following parameter settings were selected from the range: $\alpha_{\text{rft}} = 0.01$, $\alpha_{\text{ext}} = 0.01$, $w_{\text{pref}} = 4.0$, $w_{\text{cost}} = 3.5$, $r^{(\text{Operant})} = 1.0$, and $r^{(\text{Others})} = 0.5$.

To examine if the two-state model could generate bout-and-pause patterns and if it could be used for simulations with experimental manipulations, we performed simulation analysis. We varied the reinforcement rate as VI 30 s, VI 120 s, and VI 480 s, which were the same as the values used in Simulation 2.

Results: Simulation 3

Figure 8 shows the log-survivor plots of IRTs from the simulation of the two-state model with different values of the reinforcement rate. It showed broken-stick shapes and

the slopes and intercepts of the right limbs decreased as the reinforcement rate decreased.

Discussion: Simulation 3

Since the log-survivor plots of IRTs generated by the two-state model showed broken-stick curves, bout-and-pause patterns were reproduced. In addition, the change of the log survivor plots of the two-state model was consistent with experimental findings. Therefore, we can construct alternative models even without the explicit third state. Also, the two-state model implements the two mechanisms through Equation 11.

We consider the three-state, dual model has advantages in modeling and analyzing bout-and-pause patterns. First, in the three-state dual model, the effects of choice and cost are separated. It is clear in the dual model shown in Fig. 1(a) that the choice between the operant and other behaviors is made at the Choice state and whether the agent continues to stay in the same state is moderated by the cost mechanism at each of the Operant and Others states. This can be understood by Equation 4, which describes only the choice rule, and Equation 6, which calculates the stay probability based on the cost mechanism. However, in the two-state model, choice and stay are not well separated; in Equation 11, choice and cost are mixed and the behavior of the agent cannot be explained by only one of them. Second, the behavior of the two-state model when the reinforcement rate was manipulated was not close to the empirical results. Figure 4(a) showed the log-survivor plots of IRTs generated by the three-state dual model under VI 30 s, VI 120 s, and VI 480 s, and they are consistent with empirical results reported in Shull et al. (2004). Figure 8 showed the log-survivor plots of IRTs generated by the two-state model and the change in slopes and intercepts were smaller than empirical results.

General Discussion

In this paper, we have developed a computational model with reinforcement learning. The model was meant to explain how bout-and-pause patterns can be generated and we examined its validity by comparing computer simulations and

experimental findings. We hypothesized that two independent mechanisms, the choice between Operant and Others and the cost in the changeover of behaviors, are necessary to organize responses into bout-and-pause patterns. We demonstrated in Simulation 1 that the dual model reproduced bout-and-pause patterns under a VI schedule. Simulation 2 found that the relationships between various experimental manipulations and the bout components in our model were consistent. Simulation 3 found that two-state model incorporating the two mechanisms can also reproduce the bout-and-pause pattern. However, the third state has advantages in analyzing an agent behavior because it separates the effects of the choice and cost mechanisms. These results support our hypothesis that assumes that an agent transitioning between the three states driven by the choice and cost mechanisms organizes its responses into bout-and-pause patterns. This is our answer to "why bout-and-pause patterns are organized?"

Our constructive model reproduced the descriptive results reported by (Shull, 2004; Shull et al., 2001, 2002; Shull et al., 2004). Although our dual model does not explicitly include the bi-exponential model in Equation 1, IRTs generated by the dual model followed the bi-exponential model. This is an example of the first advantage of reinforcement learning discussed in the next paragraph.

Our model is developed based on reinforcement learning, which has two advantages. First, reinforcement learning models consist of simple rules. Instead of describing complicated behaviors of animals by abstract rules, we implement a simple model designed by decomposing animal behaviors into basic rules that are likely to be implemented also in real animals. This is a divide-and-conquer approach to develop a model. Second, reinforcement learning is a computational approach, where the model is described with mathematical expressions that are translated into programming code. Because of this, the computational approach is useful not only to attain reproducibility, but also to run simulations to analyze what happens inside the model. Internal states of a real animal are difficult to observe, but it is easy for a computational model. Such a computational approach is more quantitative than models described by natural

languages.

The fundamental difference between our model based on reinforcement learning and Kulubekova and McDowell (2008)'s model based on selection by consequences is that our model explicitly has the choice and cost mechanisms but Kulubekova and McDowell (2008)'s model is unclear about them. Their model did not generate a clear distinction between a burst of responses in a short period and long pauses that separate bursts, resulting in a dull bend of the log-survivor plot. Kulubekova and McDowell (2008) discussed that this divergence from live animals might be due to the lack of CODs in their model. Our model reproduced clear distinction between bursts and pauses (Figure 2), and this was because our model can change CODs through the cost mechanism. Another advantage of us over Kulubekova and McDowell (2008) is that they did not compare with alternative models but we tested our hypothesises of the choice and cost mechanisms by the knockout analysis Simulation 1.

Our model has at least two shortcomings, which are the range of parameters and the redundancy of the model. The parameters α_{rft}, α_{ext}, β, ω_{pref}, and ω_{cost} in our model have not been optimized to fit to behavioral data from real animals. The evidence that supports our parameter selection is that our model quantitatively reproduced bout-and-pause patterns. Second, although our model has five parameters, real animals may have fewer parameters. To verify our model for these two points, it would be useful to compare empirical data from real animals and our computational model.

Standing on our model proposed in this paper, we can extend our research to many directions to explain more various aspects of bout-and-pause patterns. Here we discuss two of them. The following two paragraphs are devoted to this topic.

First, our results were retrospective to data from previous behavioral experiments and the proposed model was not tested by its prediction ability to unseen data. Our model can suggest a new experiment that could add new knowledge about how manipulating CODs affect animals' behavior under a concurrent VI VI schedule. Smith et al. (2014) pointed out that employing asymmetrical CODs in a concurrent VI VI schedule could produce behaviors under a single response schedule. Our modeling is

consistent to what Smith et al. (2014) pointed out but approaches from a different direction. We consider that, even if an animal is under the single schedule, it makes choices between the operant behavior and other behaviors; this is implemented to our simulation as concurrent VI FR 1. We used an FR 1 schedule for other behaviors in our simulations, but we can change it from FR 1 to a VI schedule so that the whole schedule becomes a concurrent VI VI schedule. In our model, the cost for the operant behavior, defined by $Q_{\text{cost}}^{(\text{Operant})}$, affects actions of the agent only in the Operant state, without affecting those in the Others state; similarly, the cost for other behaviors influences the agent to its actions only in the Others state. Therefore, according to our model, it is expected that, in a concurrent VI VI schedule, if the experimenter varies CODs for one schedule, the behavior of the animal changes only for the varied schedule without affecting the behavior for the other schedule. It will be interesting to conduct such experiments with real animals to reveal actual effects of CODs on the behavior under concurrent VI VI schedules. In such a way, our model can bridge between animal behaviors observed in concurrent schedules and single schedules by offering a unified framework.

The second approach is verification based on neuroscientific knowledge. Even if the model can correctly predict unseen data from behavioral experiments, it is not guaranteed that animals employ the same the model. To explore real mechanisms that animals implement, it would be effective to compare the internal variables of the model with neural activities measured from real animals during behavioral experiments. Possible experiments are to perform knock out experiments by inducing lesions at specific areas of the brain that should be active during the experiments, or to activate or deactivate specific neurons during the experiment.

Reinforcement learning can be employed to model and explain animal behaviors other than bout-and-pause patterns, since it is a general framework where an agent learns optimal behaviors in a given environment through trial-and-error (Sutton & Barto, 2018). Such a reinforcement learning framework agrees well with the three-term contingency in behavior analysis. There are three essential elements in reinforcement

learning; a state, an action, and a reward. The state is what agent observe and is information about the environment. The action is a behavior that the agent takes in a given state. The reward is what the agent obtains as the result of the action. These three elements are similar to a discriminative stimulus, a response, and an outcome. This similarity would allow behavior analysts to employ reinforcement learning in their research. For example, Sakai and Fukai (2008) employed actor-critic reinforcement learning to modeling the matching law. We hope more computational studies will be performed to expand methods of behavioral science.

Acknowledgment

We are grateful to the editor and reviewers for giving us constructive comments, by which this study has been improved very much.

References

Barabasi, A.-L. (2005). The origin of bursts and heavy tails in human dynamics. Nature, 435(7039), 207.

Bennett, J. A., Hughes, C. E., & Pitts, R. C. (2007). Effects of methamphetamine on response rate: a microstructural analysis. Behavioural Processes, 75(2), 199–205.

Bowers, M. T., Hill, J., & Palya, W. L. (2008). Interresponse time structures in variable-ratio and variable-interval schedules. Journal of the Experimental Analysis of Behavior, 90(3), 345–362.

Brackney, R. J., Cheung, T. H., Herbst, K., Hill, J. C., & Sanabria, F. (2012). Extinction learning deficit in a rodent model of attention-deficit hyperactivity disorder. Behavioral and Brain Functions, 8(1), 59.

Brackney, R. J., Cheung, T. H., Neisewander, J. L., & Sanabria, F. (2011). The isolation of motivational, motoric, and schedule effects on operant performance: a modeling approach. Journal of the Experimental Analysis of Behavior, 96(1), 17–38.

Cheung, T. H., Neisewander, J. L., & Sanabria, F. (2012). Extinction under a behavioral microscope: isolating the sources of decline in operant response rate. Behavioural Processes, 90(1), 111–123.

Fleshler, M. & Hoffman, H. S. (1962). A progression for generating variable-interval schedules. Journal of the Experimental Analysis of Behavior, 5(4), 529–530.

Herrnstein, R. J. (1970). On the law of effect 1. Journal of the experimental analysis of behavior, 13(2), 243–266.

Killeen, P. R., Hall, S. S., Reilly, M. P., & Kettle, L. C. (2002). Molecular analyses of the principal components of response strength. Journal of the Experimental Analysis of Behavior, 78(2), 127–160.

Kulubekova, S. & McDowell, J. J. (2008). A computational model of selection by consequences: log survivor plots. Behavioural Processes, 78(2), 291–296.

Matsui, H., Yamada, K., Sakagami, T., & Tanno, T. (2018). Modeling bout–pause response patterns in variable-ratio and variable-interval schedules using hierarchical Bayesian methodology. Behavioural Processes, 157, 346–353.

McDowell, J. J. (2004). A computational model of selection by consequences. Journal of the Experimental Analysis of Behavior, 81(3), 297–317.

McDowell, J. (2005). On the classic and modern theories of matching. Journal of the Experimental Analysis of Behavior, 84(1), 111–127.

Podlesnik, C. A., Jimenez-Gomez, C., Ward, R. D., & Shahan, T. A. (2006). Resistance to change of responding maintained by unsignaled delays to reinforcement: a response-bout analysis. Journal of the Experimental Analysis of Behavior, 85(3), 329–347.

Sakai, Y. & Fukai, T. (2008). The actor-critic learning is behind the matching law: matching versus optimal behaviors. Neural Computation, 20(1), 227–251.

Shull, R. L. (2004). Bouts of responding on variable-interval schedules: effects of deprivation level. Journal of the Experimental Analysis of Behavior, 81(2), 155–167.

Shull, R. L., Gaynor, S. T., & Grimes, J. A. (2001). Response rate viewed as engagement bouts: effects of relative reinforcement and schedule type. Journal of the Experimental Analysis of Behavior, 75(3), 247–274.

Shull, R. L., Gaynor, S. T., & Grimes, J. A. (2002). Response rate viewed as engagement bouts: Resistance to extinction. Journal of the Experimental Analysis of Behavior, 77(3), 211–231.

Shull, R. L., Grimes, J. A., & Bennett, J. A. (2004). Bouts of responding: the relation between bout rate and the rate of variable-interval reinforcement. Journal of the Experimental Analysis of Behavior, 81(1), 65–83.

Smith, T. T., McLean, A. P., Shull, R. L., Hughes, C. E., & Pitts, R. C. (2014). Concurrent performance as bouts of behavior. Journal of the Experimental Analysis of Behavior, 102(1), 102–125.

Sorribes, A., Armendariz, B. G., Lopez-Pigozzi, D., Murga, C., & de Polavieja, G. G. (2011). The origin of behavioral bursts in decision-making circuitry. PLoS Computational Biology, 7(6), e1002075.

Sutton, R. S. & Barto, A. G. (2018). Reinforcement learning: an introduction. MIT press.

Tanno, T. (2016). Response-bout analysis of interresponse times in variable-ratio and variable-interval schedules. Behavioural Processes, 132, 12–21.

Tolkamp, B. J. & Kyriazakis, I. (1999). To split behaviour into bouts, log-transform the intervals. Animal Behaviour, 57(4), 807–817.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. Journal of Machine Learning Research, 11(Dec), 3571–3594.

Table 1

Previous findings from animal experiments on the relationships between manipulations and bout components. The three cells marked with "?" do not have agreement within the previous reports.

	Motivational			Schedule type
	Reinforcement rate	Deprivation level	Extinction	Tandem VR
Bout length	$\nearrow$ or $\rightarrow$?	$\nearrow$ or $\rightarrow$?	$\searrow$ or $\rightarrow$?	$\nearrow$
Bout initiation rate	$\nearrow$	$\nearrow$	$\searrow$	$\searrow$ or $\rightarrow$?

Table 2

Estimated parameters of the bi-exponential model in simulations.

Manipulation	Condition	ω	b	q
Rate of reinforcement	VI 30 (2.0 per min)	3.08	0.23	0.17
	VI 120 (0.5 per min)	3.06	0.09	0.27
	VI 480 (0.125 per min)	3.19	0.03	0.31
Deprivation level	High deprivation	3.04	0.24	0.17
	Baseline	3.15	0.08	0.26
	Low deprivation	3.17	0.03	0.31
Tandem VT 120 VR x	VR 0	3.07	0.09	0.26
	VR 4	3.23	0.08	0.19
	VR 8	3.11	0.08	0.16

Table 3

Parameter selection for the dynamic bi-exponential model with WAIC. The lower WAIC, the better the model.

Model	WAIC
qb-decay	1.936
b-decay	1.940
q-decay	1.980

Table 4

The behavior of the dual model. The cells marked with "" indicates the consistency with the animal findings shown in Table 1. The cells marked with "†" were the cells with "?" in Table 1.*

	Motivational			Schedule type
	Reinforcement rate	Deprivation level	Extinction	Tandem VR
Bout length	$\nearrow^{*}$	$\nearrow^{*}$	$\searrow^{\dagger}$	$\nearrow^{*}$
Bout initiation rate	$\nearrow^{*}$	$\nearrow^{*}$	$\searrow^{*}$	$\rightarrow^{\dagger}$

Table 5

The dependency of Q_{pref} and Q_{cost} to experimental manipulations in the dual model.

	Motivational			Schedule type
	Reinforcement rate	Deprivation level	Extinction	Tandem VR
Q_{pref}	$\nearrow$	$\nearrow$	$\searrow$	$\rightarrow$
Q_{cost}	$\nearrow$	$\nearrow$	$\rightarrow$	$\nearrow$

Table 6

Parameter range of the two-state model.

Parameter	Min	Max	Step
α_{rft}	0.01	0.2	0.01
α_{ext}	0.01	0.2	0.01
w_{pref}	1.0	6.0	0.1
w_{cost}	1.0	6.0	0.1

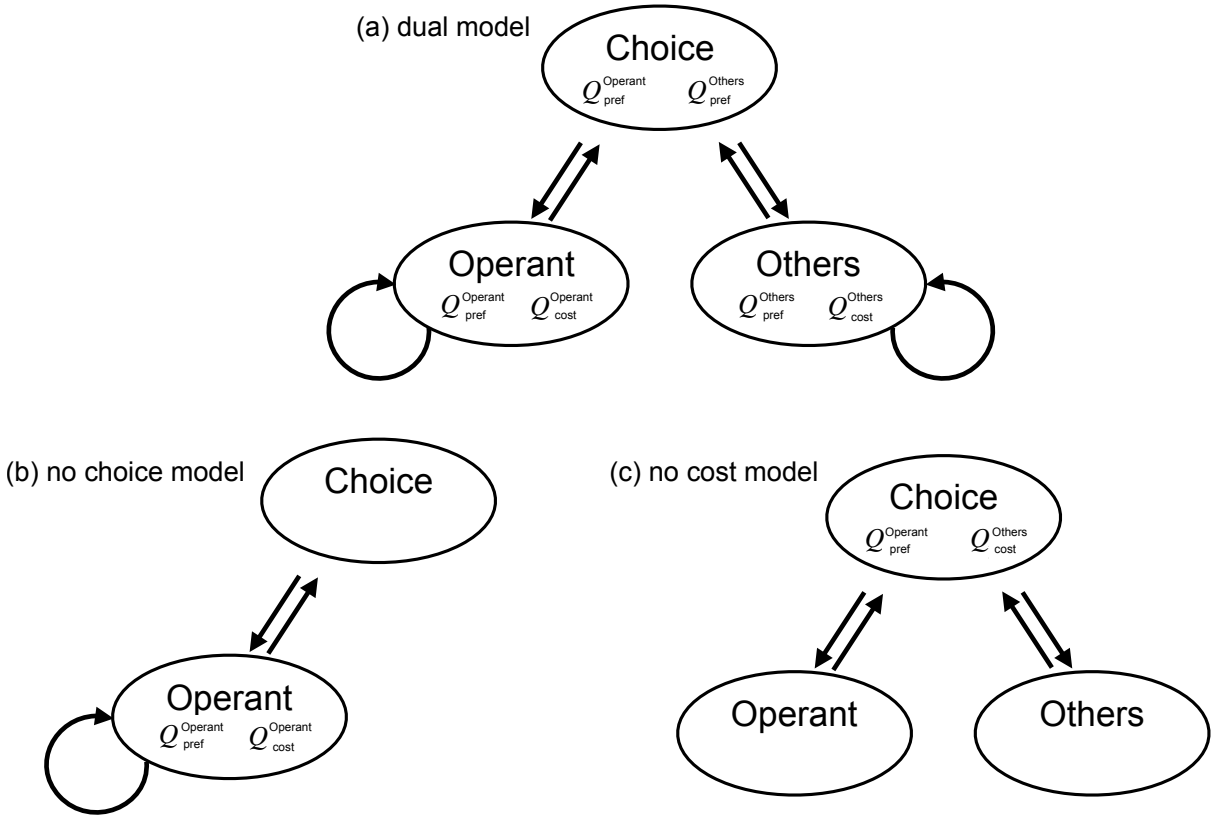

Figure 1. Model schemes of the dual model, the no choice model and the no cost model. (a) The model scheme of the dual model. The upper node, the bottom left node, and the bottom right node correspond to the Choice state, the Operant state, and the Others state respectively. Each arrow denotes the transition from one state to others. (b) The model scheme of the no choice model. In this model, the Others state is omitted. (c) The model scheme of the no cost model. In this model, self-transitions in the Operant and Others state is omitted.

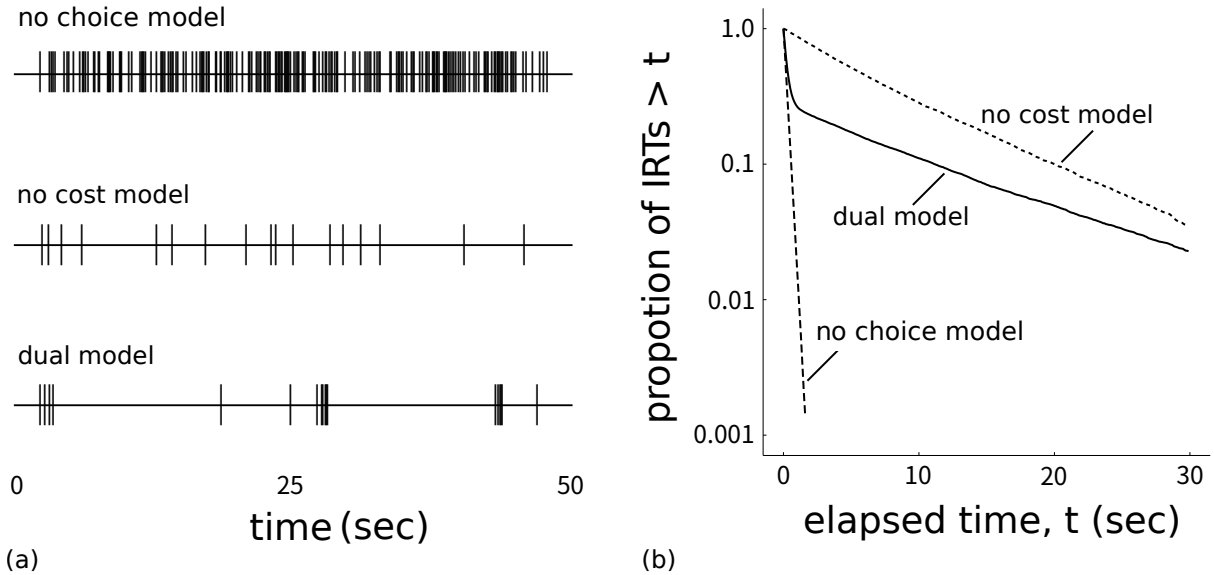

Figure 2. (a) Response event records in the (top) no choice, (middle) no cost, and (bottom) dual models in the 50 s period just after 500 reinforcers were presented (event records were stable after 500 reinforcer presentations). Each vertical line denotes one response. (b) Log-survivor plots of the three models drawn by using all the IRTs after 500 reinforcers.

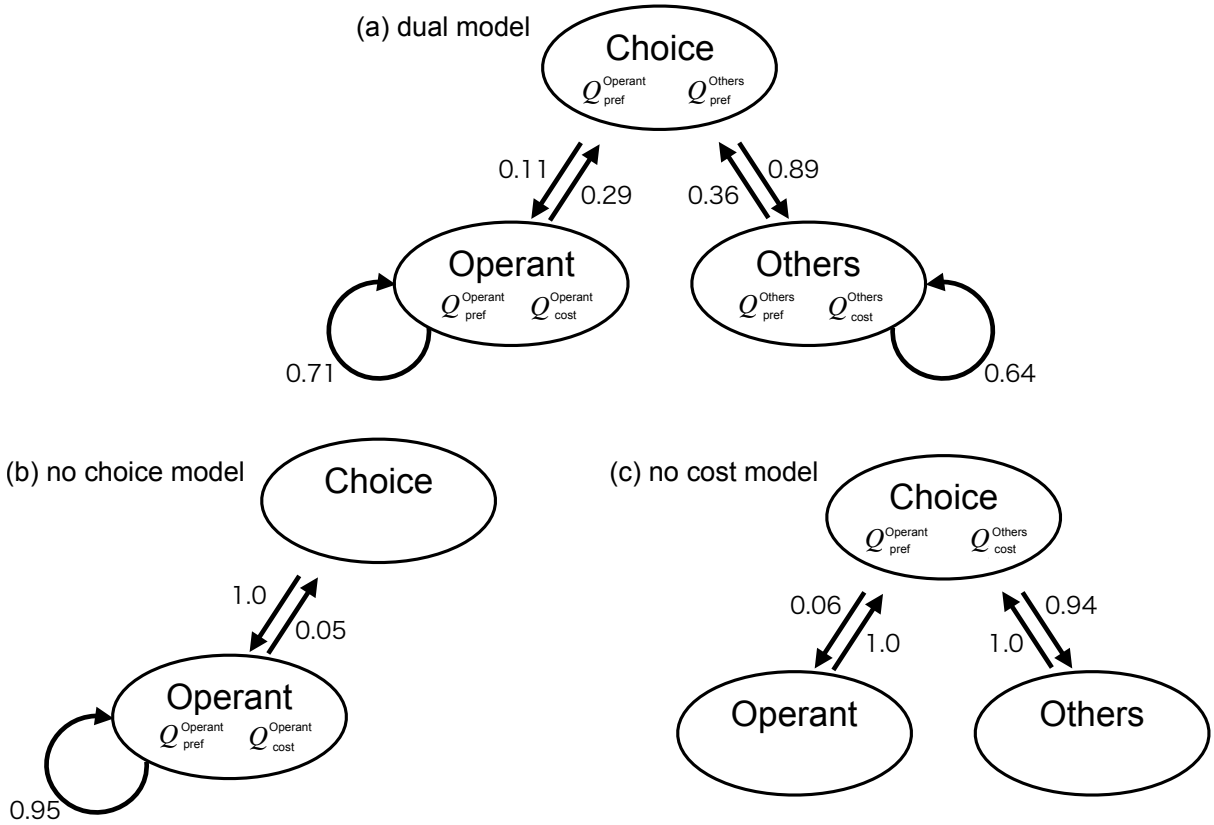

Figure 3. The transition probabilities between the three states that were calculated from the simulation data after the agent obtained 500 reinforcers.

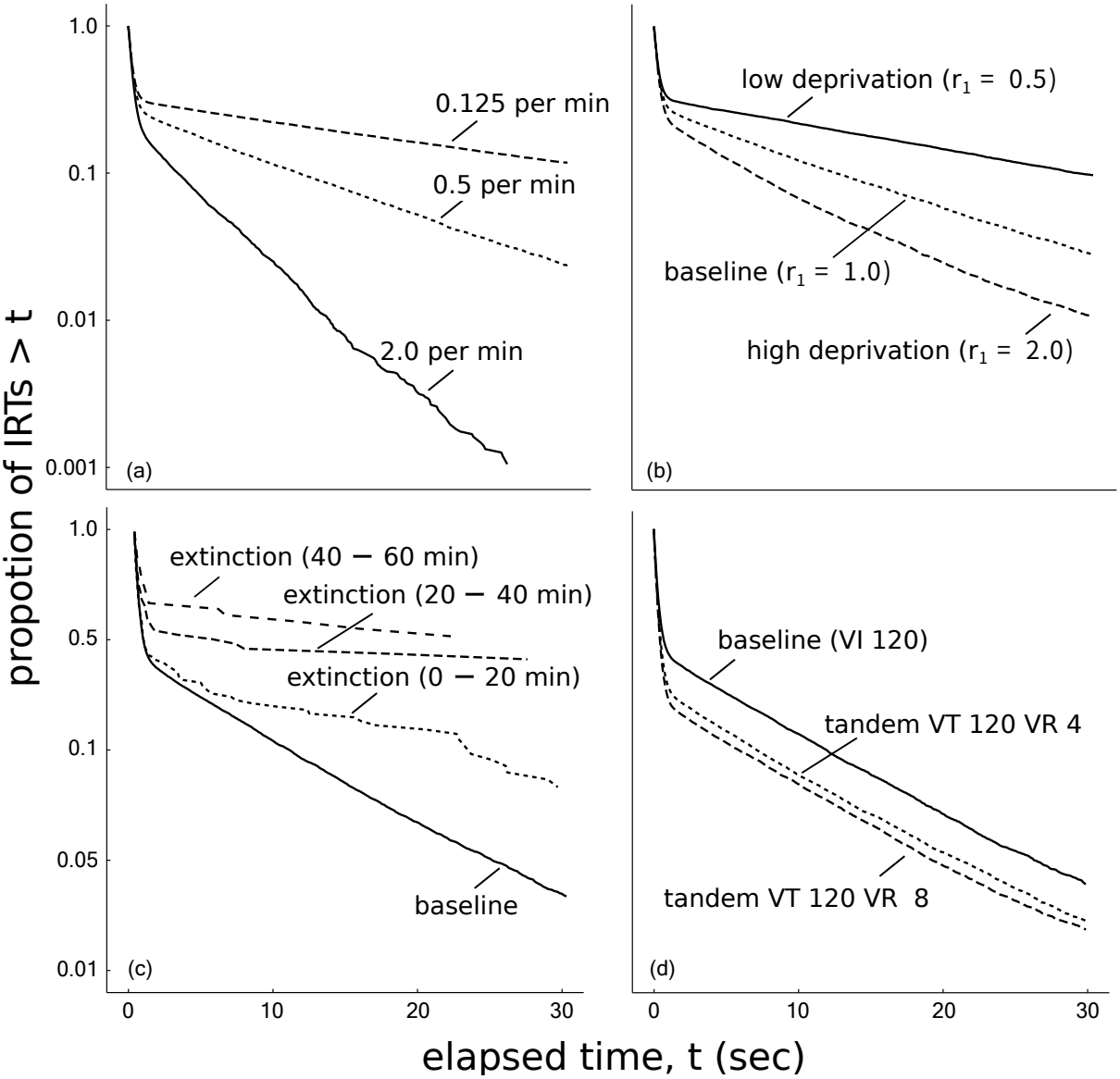

Figure 4. The log-survivor plots of IRTs generated under the manipulation of (a) the rate of reinforcement, (b) the deprivation level, (c) the schedule type, and (d) the presence of extinction drawn by all data after the agent obtained 500 reinforcers.

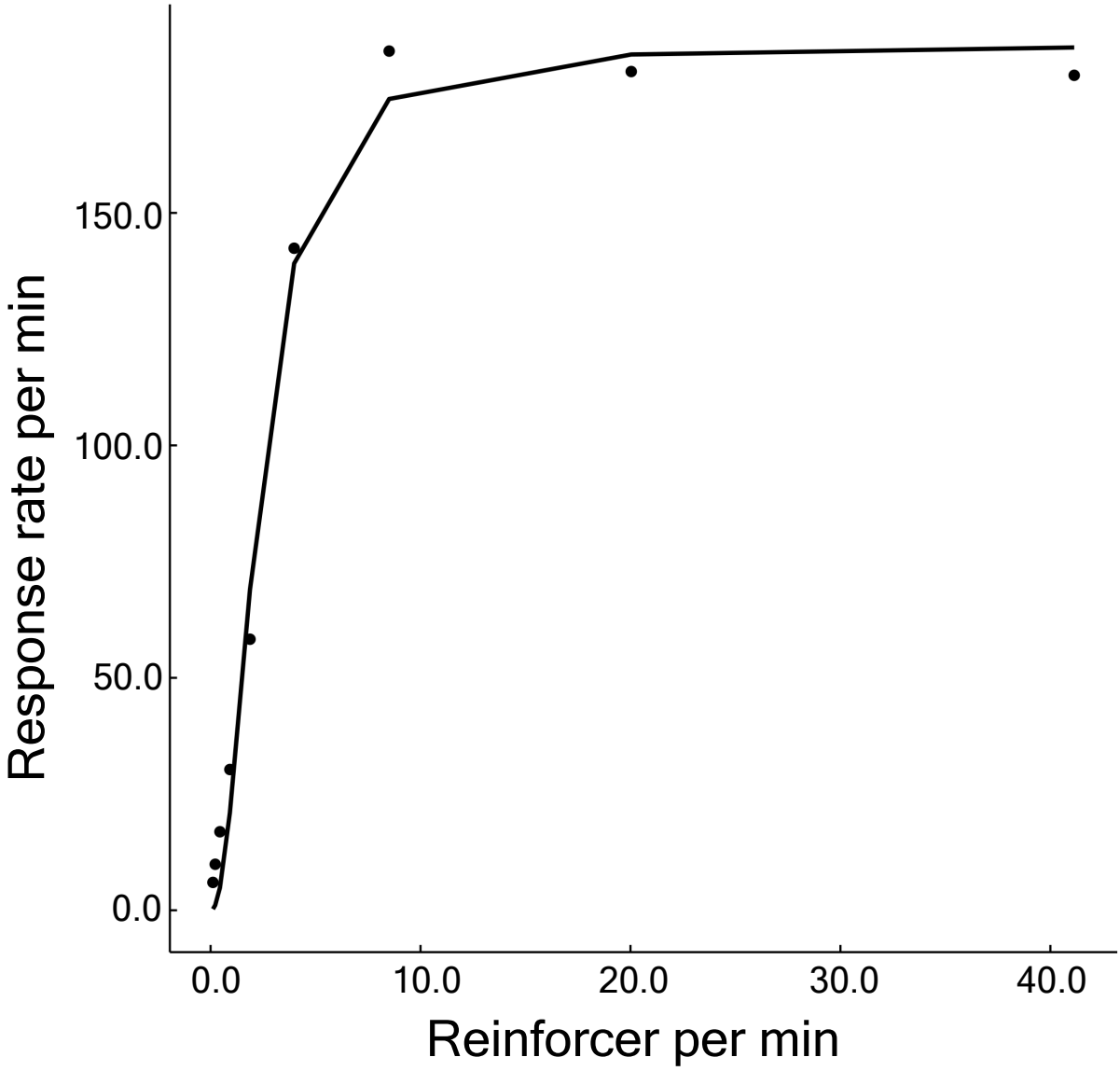

Figure 5. The response rate as a function of the reinforcement rate. The dots are from the simulation and the line is the modern version of Herrnstein's hyperbola (the generalized matching law) fitted to the data.

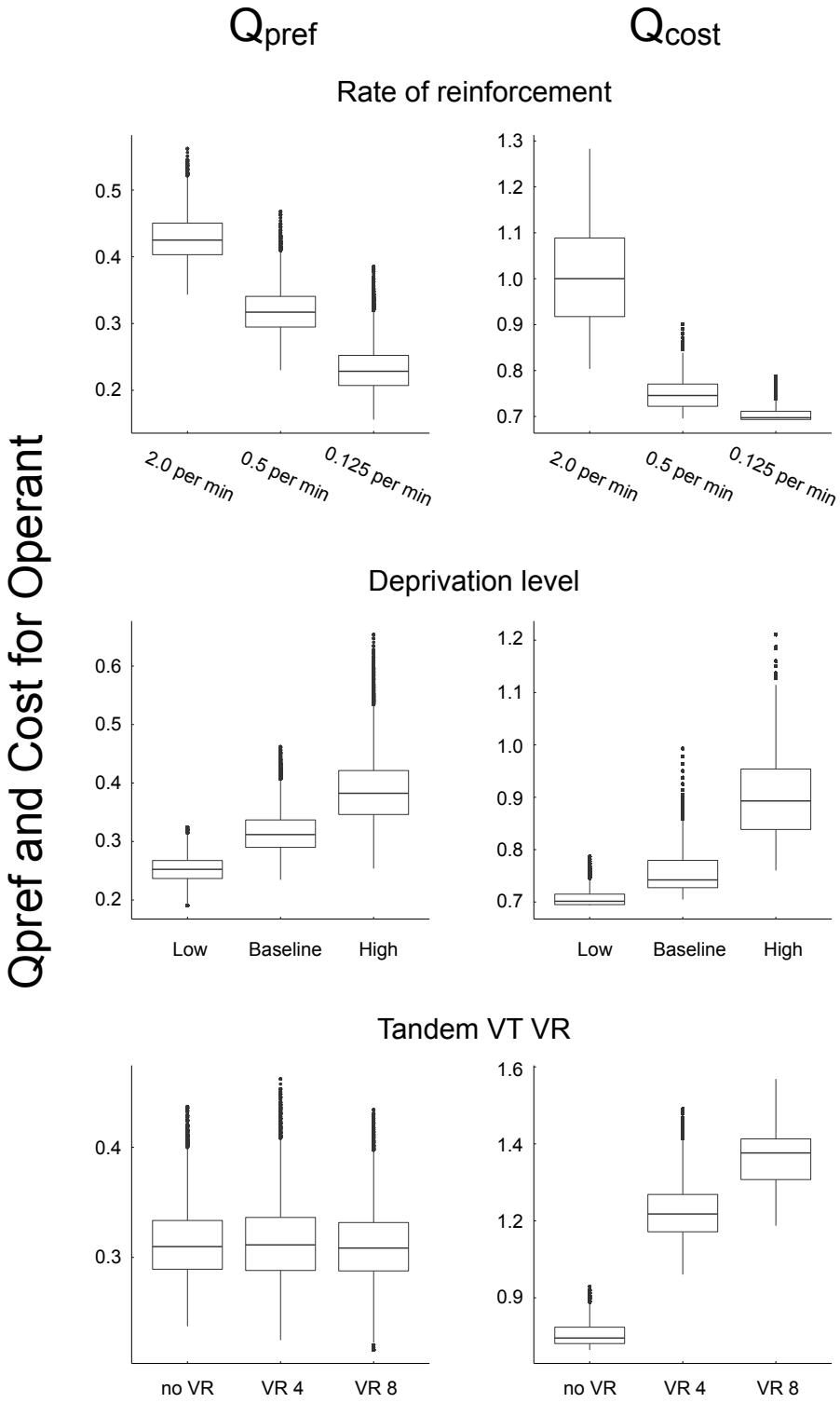

Figure 6. Boxplots of Q_{pref} and Q_{cost} in each simulation. The top, middle, and bottom rows correspond to the reinforcement rate, the deprivation level, and the tandem VT VR simulations, respectively, and the left and right columns show Q_{pref} and Q_{cost}, respectively.

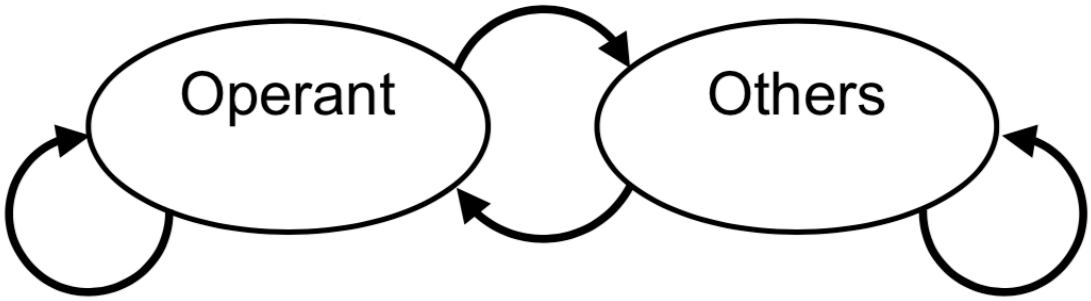

Figure 7. Two-state model.

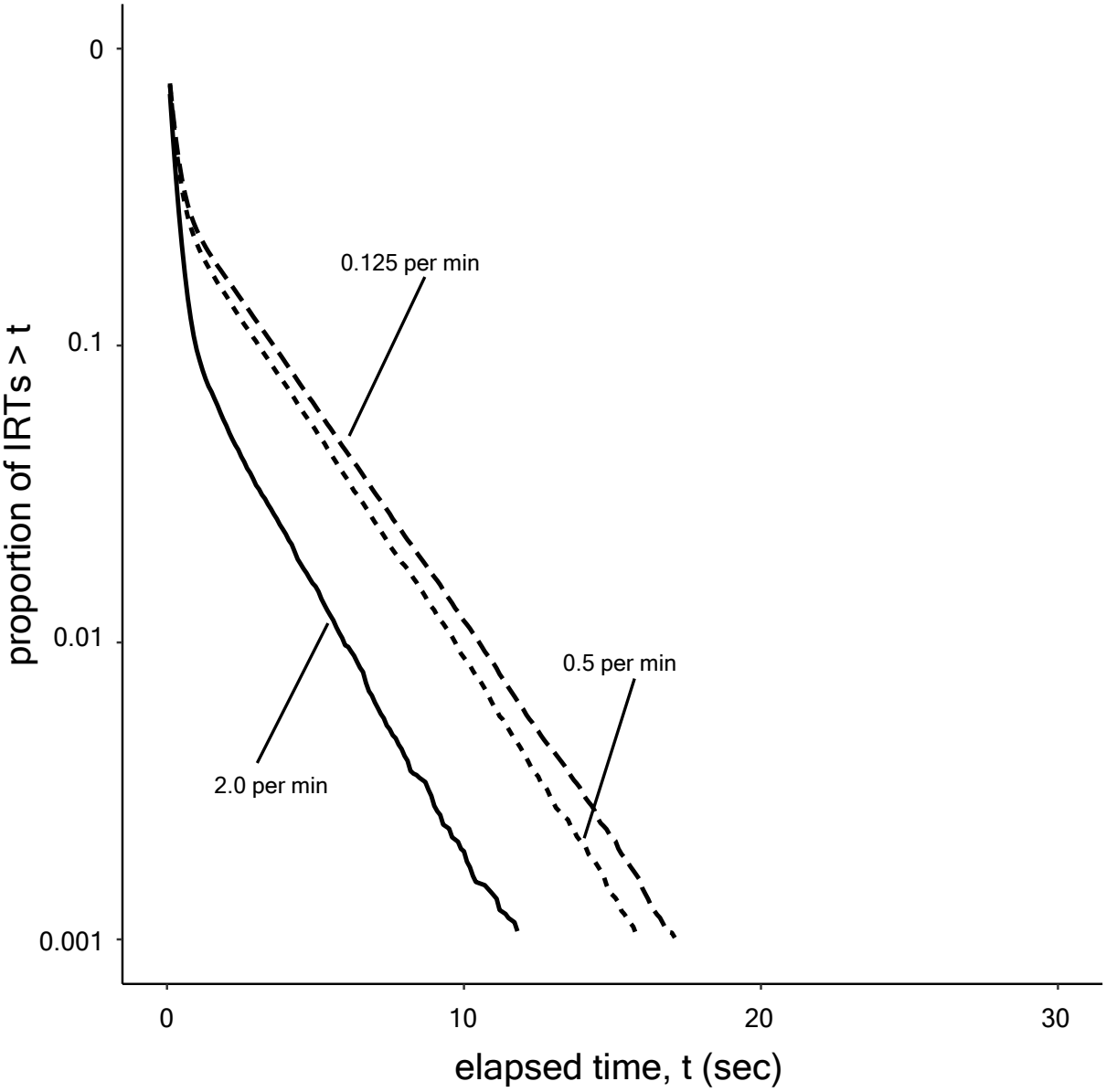

Figure 8. Log-survivor plots of IRTs generated by the two-state model under VI 30 s, VI 120 s, and VI 480 s schedules.

Appendix

Brief Introduction to Reinforcement Learning

This section summarizes key concepts of reinforcement learning that are used in this paper. For details, readers are referred to textbooks such as Sutton and Barto (2018).

Reinforcement learning is a framework where an agent learns an optimal policy (a mapping from a state to an action) that maximizes cumulative rewards through trial and error in a given environment. The agent interacts with the environment by performing actions, receiving rewards, and observing states. More precisely, in each discrete time step $(0, 1, 2 \ldots, t)$, the agent performs an action and the environment gives a new state and a reward to the agent.

We explain key concepts of reinforcement learning; state, action, and reward. A set $\mathcal{S} = \{s_1, s_2, \ldots\}$ defines the states of the environment. The agent at state s_i can take an action from available actions (behaviors) at that state, $\mathcal{A}(s_i) = \{a_1, a_2, \ldots\}$. If the agent at state s_i takes action a_j, the environment gives a feed back to the agent as a reward $R(s_i, a_j) \in \mathbb{R}$, where $\mathbb{R}$ is the set of real numbers.

Let us show an example of modeling a concurrent chain schedule experiment with reinforcement learning. In this experiment, a pigeon is in an operant chamber, where a touch panel is mounted on a wall of the chamber and the panel shows two circles, 2.5 cm in diameter, side by side. At the start of trials, a discriminative stimulus is presented to the pigeon; here, each circle shows either a horizontal or vertical line. Then, the pigeon is required to respond to one circle, and as soon as the pigeon pecks to the panel, the circles are illuminated by red and green or yellow and blue light depending on the pigeon's response. The trial comes to its end and a reward is presented to the pigeon.

In this environment, the number of state is three and expressed as $\mathcal{S} = \{s_1, s_2, s_3\}$ and they correspond to situation where the pigeon is required to respond to keys. The number of action is two in each state, corresponding to respond to the left key and right key, and it expressed as $\mathcal{A}(s_i) = \{a_{\text{left}}, a_{\text{right}}\}$ for $i = 1, 2, 3$. Reward is correspond to reinforcer and, in state s_1, $R(s_1, a_{\text{left}}) = R(s_1, a_{\text{right}}) = 0$ for both responses, in state s_2, $R(s_2, a_{\text{left}}) = 4$ and $R(s_2, a_{\text{right}}) = 0$, and in state s_3, $R(s_3, a_{\text{left}}) = R(s_3, a_{\text{right}}) = 2$ for

both responses.

The agent decides its action based on the value function $Q(s_i, a_j) \in \mathbb{R}$. More specifically, at state s_i, the probability of choosing action a_j is defined by a softmax function to be $p_j = \exp(\beta Q_t(s_i, a_j)) \big/ \sum_{a \in \mathcal{A}(s_i)} \exp(\beta Q_t(s_i, a))$, where $\beta > 0$ is a hyperparameter called the inverse temperature that controls the balance between exploration and exploitation.

What the agent learns is the value function. Suppose the agent interacts with the environment at discrete time steps $t = 0, 1, 2, \ldots$. A popular learning algorithm is Q-learning, which updates the value function by $Q^{(t+1)}(s^{(t)}, a^{(t)}) = Q^{(t)}(s^{(t)}, a^{(t)}) + \alpha(r^{(t)} - Q^{(t)}(s^{(t)}, a^{(t)}))$, where α is a hyperparameter called the learning rate. $Q^{(t)}(s^{(t)}, a^{(t)})$ is the value to take action $a^{(t)}$ at state $s^{(t)}$ in time step t.

Here we simulate the interaction between the pigeon and the environment using the concepts defined so far. A trial starts from time $t = 0$, and the environment gives the initial state, $s^{(0)} = s_1$, to the agent. Then, the agent chooses an action from $\mathcal{A}(s_1) = \{a_{\text{left}}, a_{\text{right}}\}$ according to p_j, $j \in \{\text{left}, \text{right}\}$. At the next time step, $t = 1$, the environment gives a new state, $s^{(1)} = s_2$, to the agent but no reward is presented because $R(s_1, a_j) = 0$. The agent choose their action again by softmax functios and we let $a^{(1)} = a_{\text{left}}$. Then the environment gives the reward $R(s_2, a_{\text{left}}) = 4$ to the agent and the agent update $Q(s_2, a_{\text{left}})$, by Q-learning algorithm. There are no state after the $s^{(1)}$, The trial comes to its end and next trial starts.