HOW TO MAXIMIZE REWARD RATE ON TWO VARIABLE-INTERVAL PARADIGMS

ALASDAIR I. HOUSTON AND JOHN MCNAMARA

OXFORD UNIVERSITY AND UNIVERSITY OF BRISTOL

Without assuming any constraints on behavior, we derive the policy that maximizes overall reward rate on two variable-interval paradigms. The first paradigm is concurrent variable time-variable time with changeover delay. It is shown that for nearly all parameter values, a switch to the schedule with the longer interval should be followed immediately by a switch back to the schedule with the shorter interval. The matching law does not hold at the optimum and does not uniquely specify the obtained reward rate. The second paradigm is discrete trial concurrent variable interval-variable interval. For given schedule parameters, the optimal policy involves a cycle of a fixed number of choices of the schedule with the shorter interval followed by one choice of the schedule with the longer interval. Molecular maximization sometimes results in optimal behavior.

Key words: optimization, matching, immediate maximization, reward rate, variable-interval schedule

INTRODUCTION

An animal working on a fixed-interval schedule gets a reward for the first response it makes after a time equal to the schedule interval has elapsed since the last reward. A variable-interval schedule is based on a similar principle, but the interval that must elapse is not fixed but is a random variable. On a variable-interval schedule of mean T min (VI T min), the average interval that must elapse is T min. From this definition, it can be seen that the longer an animal faced with a VI schedule waits after receiving a reward, the more likely it is that a response will be rewarded.

In concurrent VI VI experiments (conc VI VI), the animal can make responses on one of two independent VI schedules. A changeover delay (COD) is often imposed when a switch is made from one schedule to another. There are various ways of implementing the COD (see, e.g., Catania, 1966), but the basic idea is that once the animal has switched schedules, the COD must elapse before a reward can be obtained. We discuss the COD further below.

Because the schedules are independent, time spent on one schedule counts as time elapsed on the other. Thus, the longer the animal has been on one schedule, the more likely it is that a reward has been set up on the other one, and so it pays the animal to alternate between the schedules. Herrnstein (1961) found that the pattern of alternation resulted in the following relationship between responses and rewards:

$$\frac{P_1}{P_1+P_2}=\frac{R_1}{R_1+R_2} \tag{1}$$

where

P_i is the total number of responses (pecks) made on schedule i ($i = 1{,}2$),

R_i is the total number of rewards obtained from schedule i.

This relationship, known as the matching law, (Baum, 1974; de Villiers, 1977; Herrnstein, 1970) has turned out to be a very fruitful generalization about operant behavior. It can be rearranged to yield

$$\frac{P_1}{P_2}=\frac{R_1}{R_2}$$

and this form has been generalized to

Tony Nevin and two referees had to wrestle with a very terse precursor of this paper. Their comments helped us to clarify our position and strengthen the results in the process. We thank Tim Beardsley, Sara Shettleworth, and John Staddon for their comments on various drafts. Robin McCleery gave invaluable advice on the computing. A. I. H. was supported by a SRC grant to Dr. D. J. McFarland and Dr. J. R. Krebs, and subsequently by a NSF grant to Duke University. Reprints may be obtained from Alasdair I. Houston at his current address: Department of Psychology, Duke University, Durham, North Carolina 27706.

$$\frac{P_1}{P_2} = k_1 \left(\frac{R_1}{R_2}\right)^{k_2} \qquad (2)$$

where k_1 and k_2 are constants (Baum, 1974; Staddon, 1968). Matching to relative times has also been found (Baum & Rachlin, 1969; Catania, 1966). Equation 3 expresses this relationship in a way that will be required later in our discussion:

$$\frac{T_1}{T_2} = \frac{R_1}{R_2} \qquad (3)$$

where T_i is the total time spent on schedule i.

Ethologists distinguish between causal and functional accounts of behavior (see, e.g., Hinde, 1970; Tinbergen, 1951; Toates, 1980). Causal accounts, which correspond to molecular analysis in operant terms, seek mechanisms that will produce the observed behavior. There have been several attempts to understand the matching law in this way (Herrnstein, 1979; McDowell & Kessel, 1979; Myerson & Miezin, 1980; Staddon, 1977). Functional accounts, on the other hand, try to explain behavior in terms of its contribution to an animal's ability to survive and reproduce. There have been many attempts to formulate such accounts as optimality principles (see, e.g., Maynard Smith, 1978; McFarland, 1977; Oster & Wilson, 1978). One optimality principle that is especially relevant to this paper is that of maximizing rate of energetic gain. The resulting theory, known as optimal foraging theory, is reviewed by Krebs (1978). When possible types of behavior do not differ significantly in their energetic costs, and all items of food have the same energetic value, the criterion of maximizing the rate at which energy is gained can be replaced by that of maximizing the rate at which food items are obtained. This is the approach we take to the conc VI VI paradigm. The activity that maximizes the reward rate is obtained, and it is shown that it does not involve matching in the form of Equation 3. The analysis used differs from previous models of conc VI VI performance in that the COD is included and constraints on the form of the solution are avoided. We also analyze a discrete trial VI procedure and show that maximizing the immediate probability of reward does not maximize overall reward rate. Each model is preceded by a review of the relevant theoretical issues.

CONTINUOUS-TIME CONC VI VI SCHEDULES

Previous Work and Our Approach

The first attempts to find the optimal behavior (Rachlin, 1978; Staddon & Motheral, 1979) involved similar approaches and came to the conclusion that matching holds when reward rate is maximized. Both approaches can be summarized as follows. For any VI schedule there is an associated schedule function which specifies the reward rate obtained from a given response rate. The conc VI VI paradigm is then represented by adding the two schedule functions to give an expression for the overall reward rate as a function of the response rate on each schedule. Standard maximization techniques, subject to a constraint on total response rate, or to a response cost, then yield matching. A problem with this approach, as Staddon and Motheral (1979) admit, is that it implicitly allows responses to occur on both schedules simultaneously.

Heyman (1979a) objects to the approach of Staddon and Motheral (1978), on the grounds that it ignores the two ways in which rewards may be obtained on a conc VI VI schedule-viz., by switching to the schedule from the other schedule or by working on the schedule. Although Staddon and Motheral (1979) have argued that their model implicitly represents switching, they admit that it does not give a precise account of conc VI VI schedules. (In any case, it is argued below that the random responding assumed by Staddon and Motheral is not in itself optimal).

The model presented by Heyman and Luce (1979a, b) contains explicit terms for the two ways an animal can get rewards on each schedule. It is assumed that the rate of responding on a schedule is so high that there is no delay between a reward being set up and obtained. Thus, in contrast to the models of Rachlin (1978) and Staddon and Motheral (1978) the fundamental variable is not the response rate. Instead it is the probability of leaving a schedule. To elaborate, Heyman and Luce take the constant probability of a switch from one schedule to another, as found by Heyman (1979b; see also Myerson & Miezin, 1980) as a constraint on possible strategies. The problem therefore reduces to finding the probability of leaving each schedule that maximizes the reward rate. Like Staddon and Motheral (1978), Heyman and Luce do not incorporate the COD, and so if no constraint is introduced, the optimal solution is to switch infinitely often. A constraint is provided by the parameter I, which reflects the mean time between switches. In their Figure 3, Heyman and Luce (1979 a) plot expected reward rate against proportion of time on the better schedule for various values of I. Except for the degenerate case when $I = 0$ (which corresponds to switching infinitely often), the optimal proportion of time on the better schedule does not match the proportion of programmed reward rates. Without comparing the optimal proportion of time with the proportion of obtained rates, Heyman and Luce (1979a, b) conclude that matching does not maximize reward rate. (Houston [Note 1] shows that the proportion of obtained rates is quite close to p, which supports the claim made by Rachlin [1979] that matching to obtained rates will hold in the Heyman-Luce formulation.)

Our main objection to the model of Heyman and Luce (1979a) is that the exponential pattern of switching times that they impose as a constraint is not in itself optimal (see below and Appendix 1). It is true that they also refer to a fixed-time model (p. 138), but despite the fact that we have not seen a detailed account of it, we claim that without a COD it will require a constraint like the parameter I to avoid zero stay times (i.e., switching infinitely often). We do not object to the idea of a maximum rate of switching as a constraint on behavior; what is not clear from the model of Heyman and Luce is why this maximum rate is not always attained. We also question the idea of keeping I constant in the model because it is observed to be constant in some circumstances.

In contrast to Heyman and Luce (1979a, b), we do not take any aspect of an animal's behavior as a constraint. We are therefore able to derive the maximum rate obtainable and evaluate actual performance against this value.

A complete account of optimal behavior on conc VI VI would specify both the stay times and the interresponse intervals on each schedule. To simplify this problem, we consider the VI paradigm in which the animal does not have to make responses—the rewards on the chosen schedule are delivered when set up (e.g., Brownstein & Pliskoff, 1968). This is really a concurrent variable time-variable time (VT VT) procedure. We derive the behavior that maximizes reward rate on such schedules for all possible VTs and CODs and show that it does not involve matching in the form of Equation (3), but that a form of biased matching (Baum, 1974) does hold. The performance of various matching policies is also presented.

The Model

We consider an animal faced with two concurrent VT schedules which we refer to as Schedule 1 and Schedule 2. We use $i = 1,2$ to denote these schedules and assume that Schedule i has a constant probability λ_i of setting up a reward per unit time, i.e., λ_i is the mean reward rate on schedule i. The probability that a reward has not been set up on schedule i by time t is $e^{-\lambda_i t}$, and the average time between rewards on schedule i (i.e., the schedule interval) is $1/\lambda_i$. (Note that this implies a negative exponential distribution of intervals between rewards being set up on a schedule. Baum and Rachlin (1969) used such a distribution, but Brownstein and Pliskoff (1968) used a uniform distribution of intervals.) When the first reward on the schedule that is not delivering rewards becomes due ("set up"), this stops the timer for the schedule and the reward is held until the schedule comes into operation. When the animal switches to schedule i it experiences a COD of duration τ_i before any rewards can be received. The animal has to decide when to switch from one schedule to the other. The measures we use are illustrated in Figure 1. For example, at Point A the animal switches to Schedule 2. After the COD of duration τ_2, the animal spends a time a_2 on the schedule before switching back at Point C. The time t_2 between a switch from Schedule 1 and a switch from Schedule 2 is thus $\tau_2 + a_2$. The times a_1 and a_2 will be referred to as the stay times.

To clarify the difference between a_i and t_i, the model will now be described in the context of experimental procedures. Baum and Rachlin (1969) required pigeons to choose a schedule by standing on one or the other side of a chamber. When the bird stood on the left side, a red light shone, and the associated VT schedule delivered rewards. A move to the right side resulted in a green light coming on after a COD of 4.25 sec. The other VT schedule then delivered its rewards when they became set up. A white light shone during the COD. In terms of Figure 1, a switch occurs at Point A and the white light is on for the duration τ_2. The white light is replaced by the green light at Point B, and Schedule 2 is now able to deliver rewards. A switch back to Schedule 1 is initiated at Point C, so Schedule 2 has been able to deliver rewards for a time a_2. Baum and Rachlin (1969) summed the stay times for each schedule, but the usual procedure is to start measuring the time on a schedule from the decision to switch, that is to measure t_i rather than a_i. For example, in the procedure used by Brownstein and Pliskoff (1968), a blue light shines and Schedule 1 is in operation until the animal operates the changeover key at Point A. The light immediately changes to amber, but the schedule cannot deliver any rewards until the COD τ_2 has elapsed. Thus the light is blue for time t_1 and amber for time t_2. Brownstein and Pliskoff measured the total time during the session for which the light was blue and the total time for which it was amber.

Our model applies to both procedures, but the one used by Baum and Rachlin (1969) has some conceptual advantages. One is that the COD is signaled—the model assumes that the animal knows the value of the COD, so, at least in theory, it seems reasonable to indicate its magnitude. The other is use of the times a_1 and a_2 rather than t_1 and t_2. The former are really the times *on* a schedule, in that they are the times during which a schedule can deliver rewards. The variables t_1 and t_2 are the times between switches to and from Schedules 1 and 2 respectively and so will be called interswitch times. As $t_i = a_i + \tau_i$, either measure can be used. Note that we have ignored the time taken to eat the rewards. This is equivalent to stopping the timers when a reward is obtained.

In most experiments τ_1 is equal to τ_2 (Pliskoff, 1971, is an exception), but it turns out that no extra effort is required to find the optimal behavior when the CODs are not equal.

As has been said above, it is assumed that the animal knows the parameters of the problem (i.e., λ_1, λ_2, τ_1, and τ_2) and so does not have to acquire information about them as it goes along (The problem of acquiring information on variable-ratio schedules is discussed by Krebs, Kacelnik, and Taylor [1978], see also McNamara and Houston [1980]). In addition, because the probability of a reward being set up on a schedule is constant, obtaining a reward provides the animal with no further information about future rewards, and hence the optimal stay time does not depend on whether a reward has been obtained. Furthermore, the optimal behavior requires a fixed rather than varying stay time on each schedule. In fact, it can be shown (see Appendix 1) that replacing a variable stay time by its mean value increases the reward rate. The argument also applies to the case considered by Heyman and Luce (1979), so that their optimum, which involves varying stay times, is not the best that can be attained.

The above arguments show that the optimal policy must take one of the following two forms:

I. Never Switch; i.e., always stay on the schedule with the higher reward rate.
II. Stay for times a_1 and a_2 on Schedules 1 and 2 respectively, where a_1 and a_2 are nonnegative constants.

The mean reward rate for a policy in Class II can now be derived. Considering Schedule 1 first, rewards can be obtained in two ways: (i) by returning to the schedule (Point D in Figure 1)—this will be called *collecting*. (ii) by waiting for a time a_1 on the schedule—i.e., by *staying*. Let p_1 be the probability of collecting a reward at Point D. From the figure it can be seen that a time $a_2 + \tau_1 + \tau_2$ has elapsed since the animal was last on Schedule 1. It follows from the definition of λ_1 that the probability of no reward is

$$e^{-\lambda_1(a_2+\tau_1+\tau_2)}$$

and so

$$p_1 = 1 - e^{-\lambda_1(a_2+\tau_1+\tau_2)} \tag{4}$$

The animal then stays for a time a_1 on Schedule 1, for which the expected reward is $\lambda_1 a_1$. Repeating the argument for Schedule 2 gives the following expression for the mean reward on the cycle shown in Figure 1:

$$p_1 + p_2 + \lambda_1 a_1 + \lambda_2 a_2$$

where

$$p_2 = 1 - e^{-\lambda_2(a_1+\tau_1+\tau_2)}. \tag{5}$$

The total duration of the cycle is $a_1 + a_2 + \tau_1 + \tau_2$, so that if $R(a_1,a_2)$ is the mean reward rate for stay times a_1 and a_2, then

$$(a_1 + a_2 + \tau_1 + \tau_2)R = p_1 + p_2 + \lambda_1 a_1 + \lambda_2 a_2. \tag{6}$$

The notation can be simplified as follows. First, note that the CODs enter the equation for R (Equation 6) only by way of Equations 4

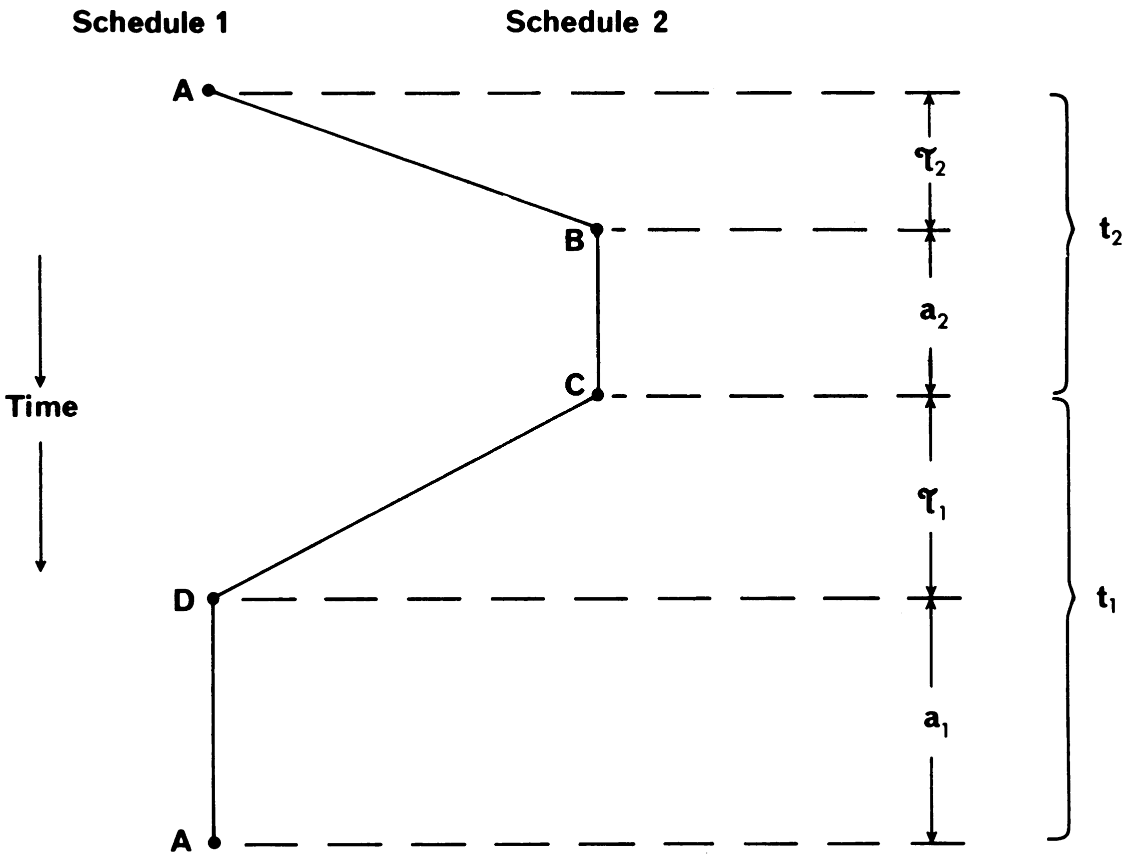

Fig. 1. Illustration of stay times and CODs for a typical cycle. The diagram refers to the steady state attained after a single switch has been made.
a_i is the stay time on schedule. i.
τ_i is the COD for a switch to schedule i.
$t_i = a_i + \tau_i$

and 5, in which they always appear as $\tau_1 + \tau_2$. This means that for given λ_1 and λ_2, the optimal behavior depends on just the sum of the CODs. Thus any problem with unequal CODs is equivalent to one with each COD having the value τ, where $2\tau = \tau_1 + \tau_2$. Secondly, it involves no loss of generality to suppose $\lambda_1 \geqq \lambda_2$ and rescale the mean reward rates by setting $\lambda_1 = 1$. The rate on Schedule 2 becomes λ_1/λ_2, which will be called λ. This procedure amounts to using the mean interval on the better schedule as the unit of time, so τ must be rescaled accordingly. The optimal behavior for all intervals and CODs can now be found, but it is convenient to think of Schedule 1 as a VI 1-min and Schedule 2 as a VI $1/\lambda$ min.

We now consider the problem of maximizing the reward rate R. If the optimal values of the stay times a_1 and a_2 are both greater than zero, then a necessary (but not sufficient) condition for them to maximize R is that $\partial R/\partial a_1$ and $\partial R/\partial a_2$ be zero. Rewriting Equation 6 in the modified notation yields

$$\begin{aligned}(a_1 + a_2 + 2\tau)R &= p_1 + p_2 + a_1 + a_2\lambda \\ &= 2 - e^{-\lambda(a_1+2\tau)} - e^{-(a_2+2\tau)} \\ &\quad + a_1 + a_2\lambda \end{aligned} \tag{7}$$

Thus,

$$(a_1 + a_2 + 2\tau)\frac{\partial R}{\partial a_1} + R = 1 + \lambda e^{-\lambda(a_1+2\tau)} \tag{8}$$

$$(a_1 + a_2 + 2\tau)\frac{\partial R}{\partial a_2} + R = \lambda + e^{-(a_2+2\tau)}. \tag{9}$$

Setting $\partial R/\partial a_1 = \partial R/\partial a_2 = 0$ in Equations 8 and 9 gives

$$R = 1 + \lambda e^{-\lambda(a_1+2\tau)} = \lambda + e^{-(a_2+2\tau)}. \tag{10}$$

Equation 10 can be rewritten as

$$R = 1 + \lambda - \lambda p_2 = 1 + \lambda - p_1 \qquad (11)$$

which implies that

$$\lambda p_2 = p_1 \qquad (12)$$
$$(\lambda_1 p_1 = \lambda_2 p_2).$$

Equation 12 is a relationship between the programmed reward rates and the probabilities of collecting a reward on returning to a schedule. If the common value of λp_2 and p_1 is denoted by A, then Equation 11 can be written as

$$R = 1 + \lambda - A. \qquad (13)$$

As $1 + \lambda$ is the maximum possible reward rate (obtainable when $\tau = 0$), A is the rewards per unit time that are lost through having to spend time switching. The optimal policy minimizes the value of A. To get some idea of the range of the reward rate, note that a rate of 1 can always be obtained just by staying on the better schedule.

Equations 12 and 13 do not guarantee a solution; i.e., they may require a_1 or a_2 to be negative. It is therefore necessary to investigate the optimal policies in more detail. To this end the policies which involve switching (i.e., policies in Class II), will be subdivided on the basis of whether they involve staying on a schedule or switching back immediately:

II Use both schedules.
IIa "Stay-Stay". Stay for a positive time on both schedules; i.e., $a_1 > 0, a_2 > 0$.
IIb "Stay-Switch." Stay for a positive time on Schedule 1, but only collect on Schedule 2—switch back to Schedule 1 without staying; i.e., $a_1 > 0, a_2 = 0$.
IIc "Switch-Stay." $a_1 = 0, a_2 > 0$.
IId "Switch-Switch." $a_1 = 0, a_2 = 0$.

It will be shown that the last two forms are never optimal.

Two equals VIs

We introduce the detailed analysis by considering the case $\lambda = 1$, i.e., the two schedules are equal. The only parameter now is τ, but it is expressed in terms of the schedule interval and must be scaled to convert it to seconds. For example $\tau = .1$ for two VIs of 1 min is a COD of 6 secs, but for two VIs of 3 min it is a COD of 18 secs.

It is obvious that when τ is greater than one, it is never worth switching, so attention is confined to $0 < \tau < 1$. The equality of the schedules means that the optimal stay times will also be equal. This common value will be called a. Using $f(a)$ to denote the reward rate as a function of a, Equation 7 implies

$$(a + \tau)f(a) = a + 1 - e^{-(a+2\tau)} \qquad (14)$$

The optimal value of a can be found by the standard procedure of differentiating Equation 14 with respect to a and equating $df(a)/da$ to zero. In Appendix 2 it is shown that this procedure results in a unique positive optimal value a^*, i.e.,

(a) the equation $\dfrac{df(a)}{da} = 0$ has a unique solution, $a = a^*$, for $a \geqq 0$.
(b) $a^* > 0$.
(c) $f(a^*) = \max\limits_{a \geqq 0} f(a)$.
(d) $f(a^*) > 1$.

This shows that a policy of the type Stay-Stay is optimal when the two VIs are equal. Appendix 2 shows that a^* is the solution to the following equation:

$$(\tau + a^* + 1)e^{-(a^*+2\tau)} = 1 - \tau \qquad (15)$$

We list some values of a^* in Table 1, from which it can be seen that a^* is very small for small CODs. In fact, for small τ, an expansion in terms of powers of τ (see Appendix 2) shows that

$$a^* \simeq \frac{2}{3}\tau^2.$$

As τ increases so does the waiting time a^* and as $\tau \uparrow 1$, a^* tends to infinity. However, the rate at which a^* tends to infinity is slow

$$a^* \sim \log(1/1 - \tau) \qquad \tau \uparrow 1.$$

Finally, for $\tau \geqq 1$, it is not worth switching at all.

Because the optimal stay times are equal on equal VIs, matching holds at the optimum. It

Table 1

The optimal value of the stay time when the schedule intervals are equal.

τ	.0124	.0248	.0492	.0967	.1425	.1867	.2702	.3472	.4828	.5952	.7606	.8640	.9254	.99935
a*	.0001	.0004	.0017	.0067	.0150	.0266	.0596	.1055	.2344	.4095	.8789	1.4720	2.1493	7.6013

Table 2

Reward rate R as a function of the stay time a on each schedule when the schedule intervals are equal.

$\tau = .1867$		$\tau = .5952$		$\tau = .864$	
a	R	a	R	a	R
.01	1.6699	.1	1.1865	.05	.9639
.0266	1.6703	.4095	1.2019	.2	.9911
.04	1.6701	.7	1.1959	.6	1.0263
.08	1.6668	1.0	1.1836	1.0	1.0338
.10	1.6642	1.3	1.1699	1.472	1.0412
.50	1.5763	2.0	1.1401	1.8	1.0400
1.00	1.4719	2.5	1.1227	2.4	1.0367
2.00	1.3293	3.0	1.1084	2.8	1.0342
5.00	1.1559	4.0	1.0869	5.0	1.0230

must be emphasised, however, that matching holds for all equal stay times on equal schedules, so the matching at the optimum is trivial. This illustrates that the matching relation does not suffice to determine the stay times, a point to which we will return later.

Matching at a suboptimal value of a does not result in a great drop in reward rate, as can be seen from Table 2.

Unequal Schedules: Division of Parameter Space

Any conc VI VI experiment is completely characterized by the parameters λ and τ, and so can be represented as a point in the parameter space shown in Figure 2. In this section it is proved that the parameter space is divided into three regions such that in each region a different form of policy is optimal.

We first show that policies of Types IIc and IId are never optimal. Appendix 3 establishes that if $0 < \lambda < 1$ and one of the Class II policies is optimal, then the optimal stay times $a_1{}^*$ and $a_2{}^*$ satisfy

$$a_1{}^* > a_2{}^*. \qquad (16)$$

Equation 16 means that policies of the form Switch-Stay and Switch-Switch cannot be optimal for $0 < \lambda < 1$. Since these forms are also not optimal when $\lambda = 1$, it follows that they cannot be optimal for any λ or $\tau > 0$. The only possible forms of optimal policy are therefore Never Switch, Stay-Stay, and Stay-Switch. We will establish that the parameter space is divided between these policies as shown in Figure 2. Above the solid line, τ is too big for switching to be profitable and so the optimal policy is to stay on the better schedule (Class I). Below the solid line one of the Class II policies, either Stay-Stay or Stay-Switch, is optimal. We show that for most of this region, including the area usually investigated in experiments, Stay-Switch is optimal, i.e., the animal should visit the worse schedule just to collect rewards that may be there, but should then switch straight back to the better schedule. This region is to the left of the broken line in Figure 2. To the right of this line Stay-Stay is optimal, i.e., the animal waits for a time greater than zero on each schedule.

The I v.II Boundary. Consider λ to be fixed. Suppose that for some τ the optimal behavior is of the form Stay-Stay. Now increase τ. As the optimal reward rate, R^*, falls the time spent on the good schedule per cycle will increase, until, as τ tends to some value $\hat{\tau}(\lambda)$,R^* will fall to 1, $a_1{}^*$ will increase to infinity, and the optimal behavior will flip from Stay-Stay to Never Switch. We know that for $\lambda = 1$ this happens when $\tau = 1$. We will find the transition point for $\lambda \leqq 1$. Full details can be found in Appendix 4.

To do this we solve the equation

$$\frac{\partial R}{\partial a_1}(a_1{}^*,a_2{}^*) = 0$$

and

$$\frac{\partial R}{\partial a_2}(a_1{}^*,a_2{}^*) = 0$$

for $a_1{}^*$ and $a_2{}^*$, and let $a_1{}^*$ tend to infinity. Eliminating $a_2{}^*$ yields the following relationship between τ and λ:

$$\hat{\tau}(\lambda) = (1/2\lambda)(1 + \lambda + (1 - \lambda)\log(1 - \lambda))\ \lambda > \lambda_c \qquad (17)$$

(We show below that this is valid for λ greater than a value λ_c which will be derived.)

The corresponding value for $a_2{}^*$ is

$$a_2{}^* = -(1/\lambda)(1 + \lambda + \log(1 - \lambda)) \qquad (18)$$

Now the analysis will hold if the behavior is of the form Stay-Stay. But this requires $a_2{}^* > 0$. Thus we require

$$1 + \lambda + \log(1 - \lambda) < 0,$$

which is true if and only if $\lambda > \lambda_c$, where

$$\lambda_c = .841405.$$

The corresponding $\hat{\tau}$ value is

$$\tau_c = .920703.$$

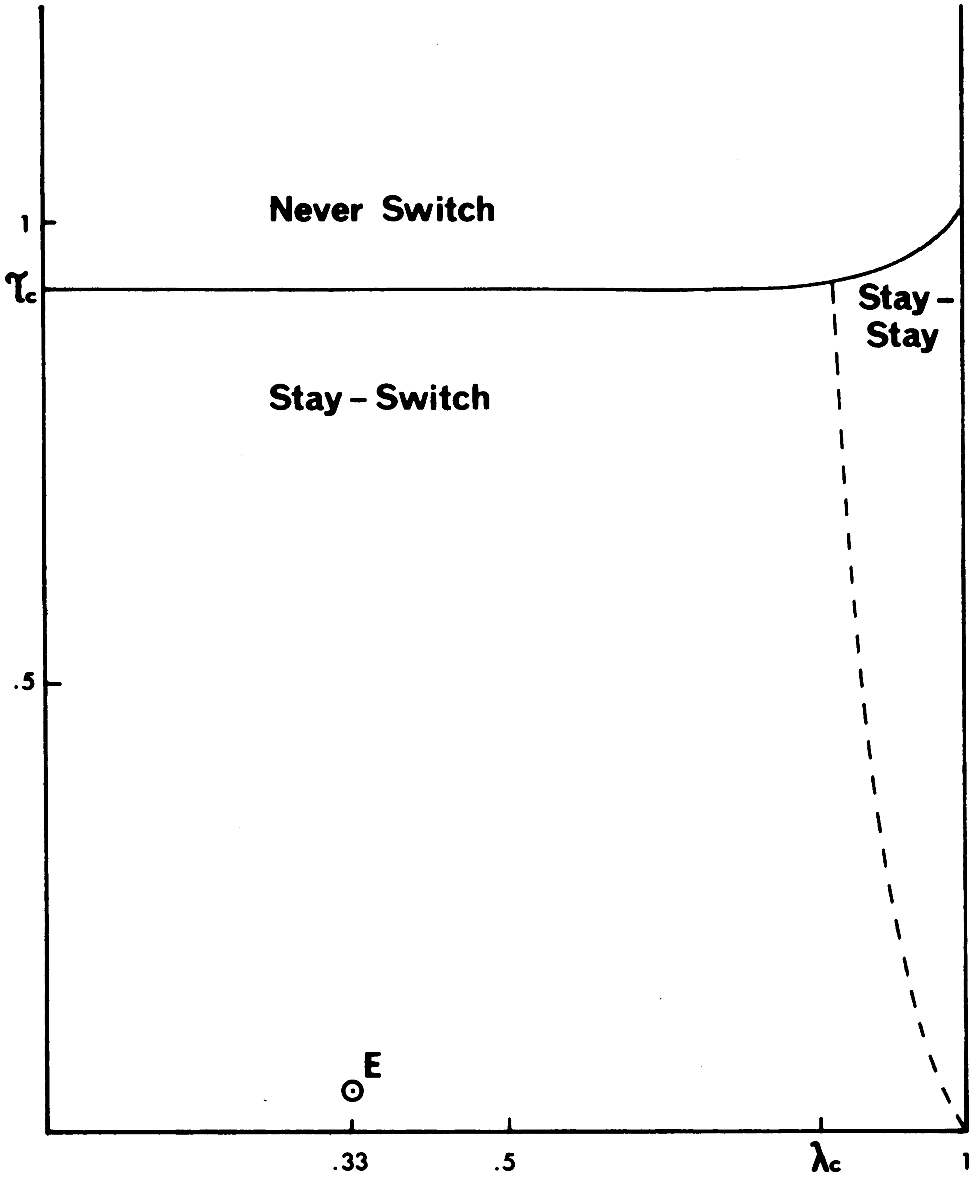

Fig. 2. The parameter space for all possible conc VT VT schedules. The space is divided into three regions. Above the solid line, it is never optimal to switch to the worse schedule. Below this line both schedules are visited, but in the Stay-Switch region rewards are collected from the worse schedule but no time is spent waiting on this schedule—the optimal policy involves switching back to the better schedule immediately. In the Stay-Stay region the animal waits on both schedules. The point E is a typical experiment. If the unit of time is 1 min, it represents a conc 1-min 3-min with a .5 sec COD. $\lambda_c \simeq .84$, $\tau_c \simeq .92$.

For $\lambda \leqq \lambda_c$ the analysis indicates that $a_2{}^* = 0$ as τ tends to $\hat{\tau}(\lambda)$, – the value at which a change to Never Switch occurs. Thus for $\lambda \leqq \lambda_c$ we should be looking for a Never Switch vs. Stay-Switch boundary rather than a Never Switch vs. Stay-Stay boundary. We find this by setting

$$\frac{\partial R}{\partial a_1}(a_1{}^*,0) = 0,$$

and letting $a_1{}^* \to \infty$. This gives $\hat{\tau}(\lambda) = \tau_c$ for all $\lambda \leqq \lambda_c$.

The Stay-Stay vs. Stay-Switch Boundary. We will now investigate the values of λ and τ for which the optimal policy undergoes a transition of form from $a_2^* = 0$ to $a_2^* > 0$. We solve the equations

$$\frac{\partial R}{\partial a_1}(a_1^*,0) = 0,$$

$$\frac{\partial R}{\partial a_2}(a_1^*,0) = 0.$$

This leads to a relationship between τ and λ which we denote by $\tau = \tilde{\tau}(\lambda)$ (See Appendix 5 for details). Let

$$S(\lambda) = \frac{\lambda + e^{-2\tilde{\tau}(\lambda)} - 1}{\lambda},$$

then

$$S(1 - \log S) = 2 - 2\tilde{\tau} - e^{-2\tilde{\tau}}.$$

Note that this equation only has a solution for $\lambda \geqq \lambda_c$ since we require $S \geqq 0$ and $2 - 2\tau - e^{-2\tau} = 0 \leftrightarrow \tau = \tau_c$. Some values of $\tilde{\tau}(\lambda)$ are given in Table 3.

Unequal Schedules: The Optimal Stay Times

In the Stay-Stay region, $\frac{\partial R}{\partial a_1} = \frac{\partial R}{\partial a_2} = 0$, so we can use Equation 10, together with Equation 7, to find a_1^* and a_2^*. The resulting equations (see Appendix 6) are transcendental, which is to say they have no algebraic solution. They were therefore solved to a specified degree of accuracy using a standard computer library package (NAG library routine CO5AAA), which was also used to find the boundary between the Stay-Stay and Stay-Switch regions and the optimal stay times in the Stay-Switch region.

Table 3

The boundary between Stay-Stay and Stay-Switch regions.

$\tilde{\tau}$	λ
.06	.9812
.08	.9744
.10	.9683
.15	.9537
.20	.9399
.30	.9147
.40	.8927
.50	.8739
.60	.8585
.70	.8478
.80	.8396
.90	.8394

Some examples of the optimal stay times a_1^* and a_2^* for various values of λ and τ are given in Table 4. It can be seen that, for any value of λ, both a_1^* and a_2^* increase as τ is increased.

Stay-Stay policies are only optimal when λ is quite close to 1, in fact probably too close for the two schedules to be readily distinguished as different by an animal. We therefore concentrate on the Stay-Switch region. Equation 10 can no longer be used because we cannot assume that $\partial R/\partial a_2 = 0$. Although we have lost this condition, we also have one less unknown, so we put $\partial R/\partial a_1 = 0$ in Equation 8 and rewrite Equation 7 with $a_2 = 0$. This yields:

$$R = 1 + \lambda e^{-\lambda(a_1+2\tau)}$$

and

$$(a_1 + 2\tau)R = 2 - e^{-\lambda(a_1+2\tau)} - e^{-2\tau} + a_1$$

Requiring both these equations to hold defines the optimal value of a_1, written a_1^*. Putting

$$y = \lambda(a_1^* + 2\tau)$$

and eliminating R from the two equations gives

$$(y + 1)e^{-y} = 2 - 2\tau - e^{-2\tau} \qquad (19)$$

Table 4

The optimal behavior for various Stay-Stay policies. The first column gives λ, with the ratio of λ_1 to λ_2 in brackets. The next three columns give the COD and the optimal stay times. The last three columns are the ratio of stay times, reward rates, and interswitch times respectively. If matching is to hold at the optimal solution, t_1^*/t_2^* must equal R_1^*/R_2^*.

λ	τ	a_1^*	a_2^*	a_1^*/a_2^*	R_1^*/R_2^*	t_1^*/t_2^*
.90	.50	.426	.051	8.227	1.399	1.679
(1.1111)	.80	1.681	.317	5.295	2.054	2.210
.925	.50	.384	.102	3.760	1.288	1.468
(1.0811)	.70	.941	.309	3.047	1.503	1.627
	.80	1.535	.472	3.241	1.743	1.835
	.90	3.119	.667	4.671	2.510	2.564
.95	.30	.115	.031	3.636	1.115	1.251
(1.0526)	.50	.341	.153	2.230	1.184	1.288
	.70	.851	.420	2.024	1.318	1.384
	.90	2.662	.954	2.792	1.902	1.436
.975	.30	.095	.054	1.770	1.056	1.117
(1.0256)	.50	.299	.205	1.461	1.089	1.134
	.70	.758	.539	1.408	1.150	1.177
	.90	2.224	1.317	1.688	1.404	1.409

If we take Equation 19 to be a functional relationship defining $y(\tau)$ in terms of τ, then the optimal stay time becomes

$$a_1{}^* = \frac{1}{\lambda}\, y(\tau) - 2\tau \qquad (20)$$

From Equation 10 it can be seen that the corresponding value of R is

$$R^* = 1 + \lambda e^{-y(\tau)} \qquad (21)$$

The probability of receiving a reward after switching to Schedule 2 is

$$1 - e^{-\lambda(a_1{}^* + 2\tau)} = 1 - e^{-y}.$$

From Equation 19, it follows that y depends only on τ, and hence on optimal Stay-Switch policies, the probability of receiving a reward from Schedule 2 is independent of λ. Because no time is spent on Schedule 2, the decision to switch to it just depends on the probability of a reward being set up and not on the reward rate λ on the schedule.

Before concentrating on behavior in the Stay-Switch region and its relation to experimental results, we illustrate the transition from Stay-Switch to Stay-Stay that occurs for $\lambda_c < \lambda < 1$ as τ increases. As Table 5 shows, when τ is below $\tilde{\hat{\tau}}(\lambda)$ ($\simeq$ .365 in this case) the optimal policy is of the form Stay-Switch, $a_2{}^*$ is therefore zero, and $a_1{}^*$ increases with increasing τ. Once τ is above $\tilde{\hat{\tau}}(\lambda)$, the optimal policy involves staying on both schedules. As τ increases, both $a_1{}^*$ and $a_2{}^*$ increase, but as τ tends to $\hat{\tau}(\lambda)$, $a_1{}^*$ tends to infinity and $a_2{}^*$ tends to a finite limit. For $\tau > \hat{\tau}(\lambda)$, the optimal policy is Never Switch; all the time is spent on Schedule 1.

Table 5

The optimal values of a_1, a_2, and R for $\lambda = .9$. When τ is less than .365, the optimal policy is Stay-Switch, i.e., $a_2 = 0$.

τ	$a_1{}^*$	$a_2{}^*$	R^*
.05	.0130	0	1.8130
.15	.0523	0	1.6554
.20	.0800	0	1.5843
.30	.1553	0	1.4561
.36	.2167	0	1.3874
.38	.2404	.0041	1.3658
.40	.2660	.0101	1.3448
.45	.3387	.0285	1.2952
.65	.8240	.1564	1.1331
.75	1.3005	.2580	1.0724
.85	2.2791	.3790	1.0251
.90	3.5714	.4335	1.0071
.92	5.1989	.4468	1.0016

The remainder of this section is devoted to VIs with $0 < \lambda < \lambda_c$ and $0 < \tau < \tau_c$. Stay-Switch is optimal for the whole of this region, so $a_2{}^* = 0$ and $t_2{}^* = \tau$. Table 6 gives some values of $a_1{}^*$ and R^*, obtained from Equations 19 to 21. The interswitch time $t_1{}^*$ is given by $a_1{}^* + \tau$. Figure 3 shows that $t_1{}^*$ increases with increasing τ for a given value of λ, and decreases with increasing λ for a given value of τ. As we said earlier, experiments measure the sum of all values of t, over a session, so it is not obvious how close animals come to our predictions. It is clear that animals are suboptimal in adopting varying stay times, but apart from Heyman (1979b), no one has given detailed accounts of this variability. It is therefore hard to evaluate this aspect of behavior. One way to compare the model and the data would be to estimate the mean value of t_1 from the total time on a schedule and the number of switches, when such information is given. It seems more straightforward, however, to compare the switching rates directly. Another measure of interest is the reward rate on each schedule. To facilitate the discussion the following notation will be used:

S = the switching rate (changeover per min)
R_i = the reward rate on schedule i

The model makes 2 switches in a time $a_1{}^* + 2\tau$, so the optimal switching rate, S^*, is given by the following equation:

Table 6

Optimal stay times $a_1{}^*$ (upper half) and optimal rewards per unit time $R_1{}^*$ (lower half) for various values of τ and λ. If the unit of time is a minute, τ ranges from .5 sec. to 24 secs.

			λ		
τ	.125	.25	.3333	.5	.6666
.0083	.115	.049	.033	.016	.008
.0166	.233	.099	.067	.033	.017
.0333	.471	.202	.135	.068	.034
.0833	1.206	.520	.348	.177	.091
.2	3.056	1.328	.896	.464	.248
.4	6.876	3.038	2.078	1.119	.639
.0083	1.123	1.246	1.328	1.492	1.656
.0166	1.121	1.242	1.322	1.484	1.645
.0333	1.117	1.234	1.312	1.467	1.623
.0833	1.105	1.211	1.281	1.421	1.562
.2	1.081	1.162	1.216	1.325	1.433
.4	1.048	1.096	1.128	1.192	1.255

$$S^* = \frac{2}{a_1^* + 2\tau} \qquad (22)$$

The optimal values of R_1 and R_2 are given by the number of rewards from a schedule per cycle divided by the duration of a cycle i.e.,

$$R_1^* = \frac{a_1^* + 1 - e^{-2\tau}}{a_1^* + 2\tau} \qquad (23)$$

and

$$R_2^* = \frac{1 - e^{-\lambda(a_1^* + 2\tau)}}{a_1^* + 2\tau} \qquad (24)$$

If, for the moment, we follow Heyman and Luce (1979) in assuming that responses cost nothing, we can draw on data from VI as well as VT schedules. De Villiers (1977) says that two generalizations can be made about switching rates: (i) for fixed schedule rates, animals switch less frequently as the COD is increased, and (ii) for fixed COD animals switch less frequently as R_1/R_2 departs from unity. Both these features are seen in the behavior of S^*.

(i) S^* *vs.* τ *for constant* λ. It can be seen from Table 6 and Figure 3 that for a given value of λ, a_1^* increases as τ increases. It therefore follows from Equation 22 that S^* decreases as τ increases. The dependence of S^* on τ for $\lambda = \frac{1}{3}$ is shown in Figure 4. We have taken the time unit to be a minute, so that the predictions can be compared with data from the conc 1-min 3-min experiments of Brownstein and Pliskoff (1968), and Todorov (1971). (The latter experiment involved VI rather than VT schedules, and all the lights were extinguished when a switch was made). The experiments show the required decrease in switching rate, but in general the observed rates are less than the optimal rate.

The optimal stay times, and hence the optimal switching rate, depends on the sum of the CODs, so if the sum is kept constant, S^* will not change. None of the conditions used by Pliskoff (1971) can be used to test this point, because the COD sum for each condition is different. It should be the case, however, that S^* decreases with $\tau_1 + \tau_2$, and the data show this trend. Pliskoff used two equal VIs throughout his experiment. The optimal behavior on equal VIs requires equal stay times, regardless of whether τ_1 is equal to τ_2. Pliskoff found, however that when τ_1 and τ_2 were different, the switching depended on the magnitude of τ_1 to be experienced on the switch, i.e., behavior seems to be influenced by short-term factors.

(ii) S^* vs. R_1^*/R_2^* for constant τ. Table 7 shows as R_1^*/R_2^* increases, S^* decreases. This sort of relationship has been found by Herrnstein (1961) and Brownstein and Pliskoff (1968), but to be precise they kept the absolute value of the COD fixed, so τ varies with the value of the better schedule. For example, Herrnstein (1961) used a COD of 1.5 sec with VIs of 3-min 3-min, 2.25-min 4.5-min, and 1.8-min 9-min, for which τ takes the value .0083, .0111, and .0139 respectively. The variation in τ is so small that it is reasonable to take it to be constant in such cases.

Table 7 also shows that R_1^*/R_2^* is virtually the same as $1/\lambda$, which means that the ratio of obtained reward rates equals the ratio of programmed rates. This equality only holds for small τ, as can be seen from Table 8, which gives a detailed account of optimal performance on conc VT 1-min VT 3-min schedules. It can be seen from the table that as the COD increases so does the ratio of obtained reinforcements. This trend has been found by Allison and Lloyd (1971) and Shull and Pliskoff (1967).

The above examples show that some of the general trends found in conc VI experiments are also found if the optimal policy is followed. But before any summary of the relationship between data and the model can be made, the tendency to match the ratio of reward rates to the ratio of total times on each schedule requires analysis. This deserves a section to itself.

The matching law. The standard form of time-based matching is given by Equation 3. One problem with this equation is that the total times T_1 and T_2 do not uniquely specify the absolute value of the stay times, but it is these absolute times that determine the reward

Table 7

The dependence of S^* (optimal number of switches per min) on R_1^*/R_2^* (the ratio of obtained reward rates on the optimal policy) for a COD of .3 sec. Note that for a COD as small as this, the ratio of obtained rates is very close to the ratio of programmed rates (i.e., $1/\lambda$).

$1/\lambda$	t_1^*/t_2^*	R_1^*/R_2^*	S^*
1.5	1.84	1.50	140.72
2.0	2.79	2.00	105.54
3.0	4.69	3.01	85.38
4.0	6.58	4.01	60.79
8.0	14.16	8.03	26.38

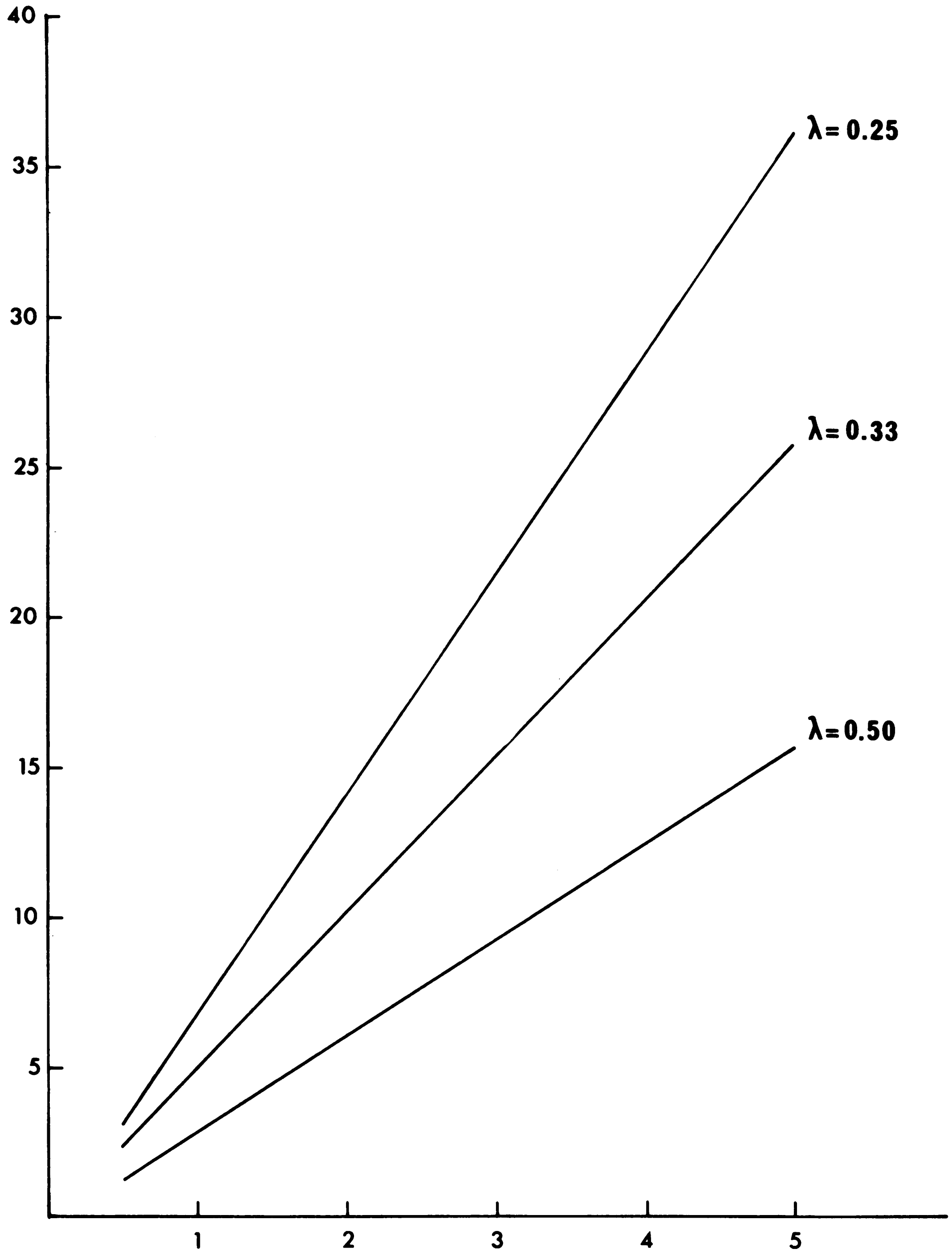

Fig. 3. The optimal interswitch time on the better schedule, t_1^*, in secs as a function of the COD in secs for three values of λ. The lines are actually very slightly curved but have been drawn as straight for convenience.

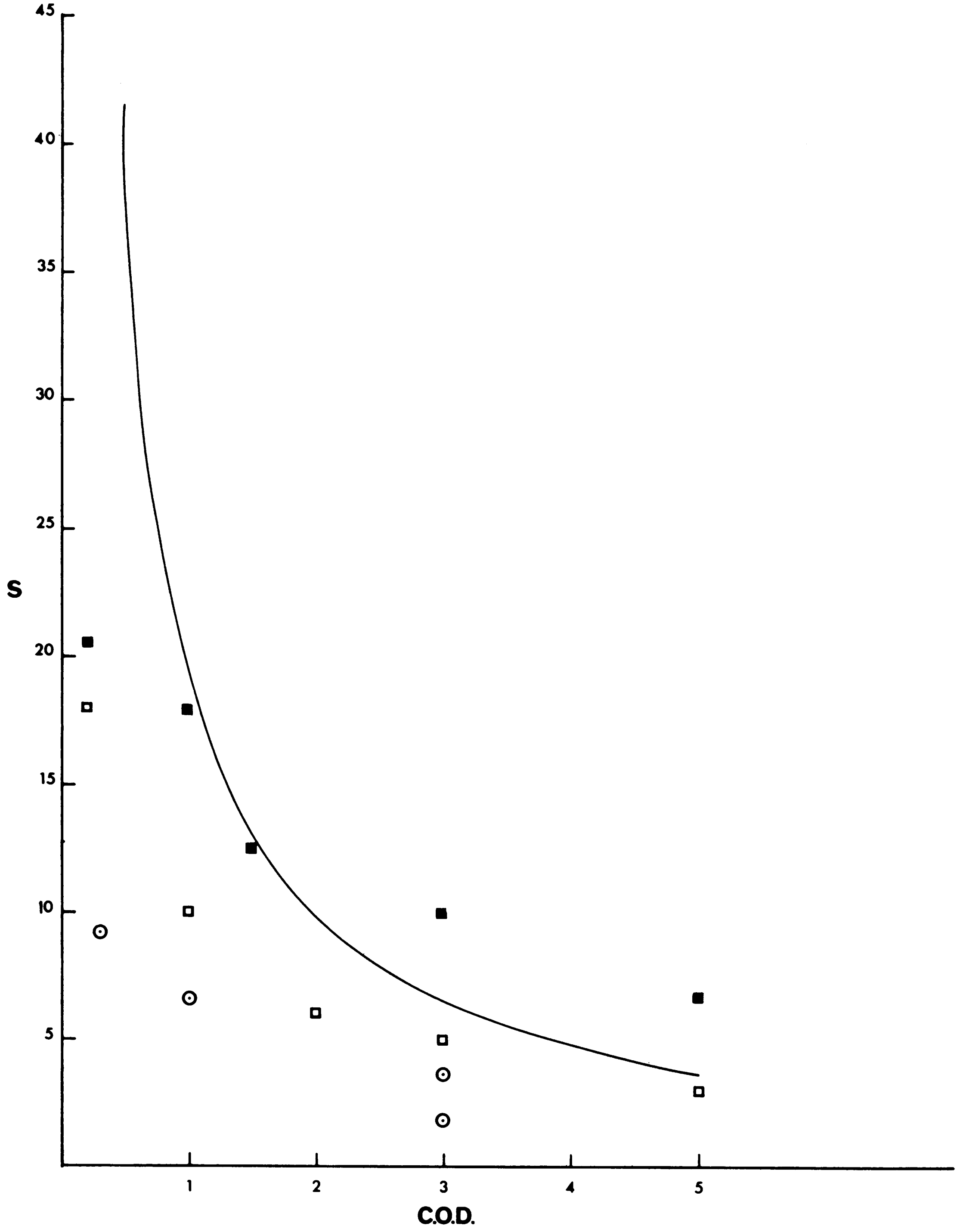

Fig. 4. The switching rate, *S*, in switches per min as a function of the COD in secs on a VT 1-min VT 3-min schedule. The line shows the optimal value of *S;* the symbols are data from the following experiments:
Filled squares—Brownstein and Pliskoff (1968) Bird 787
Unfilled squares—Brownstein and Pliskoff (1968) Bird 6494
Circles—Todorov (1971) Bird P13

Table 8

The optimal value of the stay time, reward rate, ratio of rewards, ratio of interswitch times, and the switching rate as a function of COD on a VT 1-min VT 3-min.

COD *(sec.)*	$a_1^*(sec)$	R^*	R_1^*/R_2^*	t_1^*/t_2^*	S^* (min^{-1})
.5	1.966	1.328	3.016	4.948	40.51
1.0	4.000	1.322	3.034	5.016	20.03
1.5	6.074	1.317	3.051	5.049	13.23
2.0	8.104	1.312	3.069	5.056	9.92
2.5	10.192	1.306	3.087	5.084	7.90
3.0	12.311	1.301	3.105	5.104	6.55
5.0	20.889	1.281	3.181	5.179	3.88
9.0	39.083	1.243	3.351	5.343	2.50
18.0	86.362	1.169	3.833	5.798	1.99
30.0	171.313	1.092	4.821	6.710	.60
45.0	369.290	1.026	7.518	9.206	.29

rate (see Equation 10). For any policy with fixed stay times,

$$T_i = n_i(a_i + \tau) = n_i t_i$$

Where n_i is the number of times the animal visits Schedule *i*. For experiments of reasonable length, it can be assumed that $n_1/n_2 = 1$, and hence

$$\frac{T_1}{T_2} = \frac{t_1}{t_2}$$

For the policies we consider, Equation 3 therefore becomes

$$\frac{t_1}{t_2} = \frac{R_1}{R_2} \tag{25}$$

In contrast to the optimality analysis, which specifies t_1 and t_2 for any value of λ and τ, Equation 25 links these times to the obtained reward rates. As a result, the matching law does not specify the behavior that will be observed for given values of the schedule parameters.

Before presenting the main results of this section, which concern matching in the Stay-Switch region, we claim on the basis of the results given in Table 4 that matching does not hold for optimal policies in the Stay-Stay region. Matching requires that the last two columns of this table be equal, which is not the case. Furthermore, R_1^*/R_2^* does not match a_1^*/a_2^* or $1/\lambda$.

In the Stay-Switch region it is possible to prove that matching cannot be optimal. If matching is always optimal, then Equations 23 to 25 imply that

$$\frac{a_1^* + \tau}{\tau} = \frac{a_1^* + 1 - e^{-2\tau}}{1 - e^{-y}} \tag{26}$$

It is shown in Appendix 7 that Equation 26 implies that y depends on λ, which contradicts Equation 19. Therefore Equation 26 cannot hold, and matching cannot be optimal.

This result raises two further questions about matching in the Stay-Switch region: (i) Does any more general form of matching hold? (ii) How much does it cost to match?

(i) Is a general form of matching optimal?

The first question has a simple answer. In the Stay-Switch region, the optimal behavior results in a generalized form of matching (Baum, 1974) as long as τ is kept constant. The exact form is given by the following equation:

$$\frac{a_1^* + \tau}{\tau} = k(\tau)\frac{R_1^*}{R_2^*} + c(\tau) \tag{27}$$

where k(τ) and $c(\tau)$ are constant for given values of τ (see Appendix 7 for details). Equation 27 follows immediately from Equations 23 and 24 and the fact that the quantity $y(\tau) = \lambda(a_1^* + 2\tau)$ depends only on τ and not on λ. Varying λ then results in straight lines when $\frac{a_1^* + \tau}{\tau}$ is plotted against the ratio of obtained rewards, as Figure 5 shows.

Equation 27 differs from the generalized matching equation proposed by Baum (1974) (Equation 2 above) in having a constant term $c(\tau)$. The term $k(\tau)$ corresponds to Baum's bias parameter (k_1 in Equation 2). This has an interesting theoretical implication. Baum (1974) claims that bias reflects an unaccounted factor that alters an animal's preference between the schedules. His reason is that when $R_1 = R_2$, matching (Equation 3) requires that $t_1 = t_2$, so that *k* specifies how much the time allocation ratio deviates from unity. Baum (1974) interprets *k* as depending on uncontrolled differences in the two schedules, but our analysis challenges this view. In effect, we have shown that an animal should choose to allocate its time in such a way that a generalized matching equation holds. An implication is that bias can occur without there being any hidden asymmetry in the schedules—note that R_1 and R_2 are obtained reward rates. This is not to deny that animals do show preferences, but these are revealed by cases when the *programmed* rates are meant to be equal on the two schedules. Such preferences may be very marked (one bird studied by Baum and Rachlin (1969) spent five times as much time on one schedule as on the other when both were VI 2-min) but are out-

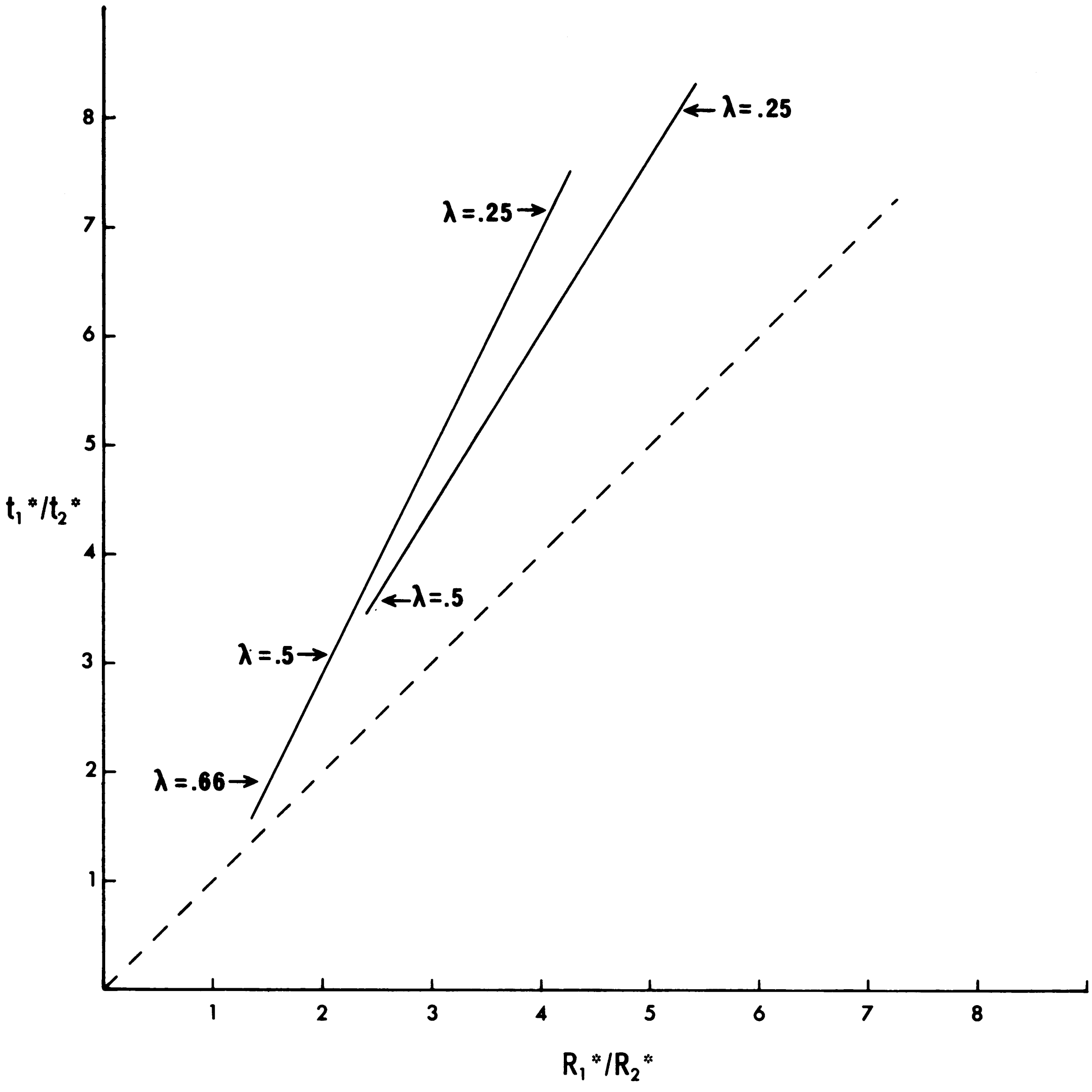

Fig. 5. The form of biased matching given by Equation 27. The upper line is for $\tau = .05$, the lower line is for $\tau = .3$. The lines are traced out by varying λ, some values of which are given in the figure. The lines cannot be continued indefinitely to the left because the optimal policy changes to Stay-Stay.

side the scope of our discussion. What we wish to draw attention to is that it is only the matching law that requires k to be seen as a preference. Outside the framework of matching, equality of the obtained rates on unequal schedules does not require equality of times spent.

(*ii*) *How much does it cost to match?* There is no simple answer to this question, because matching does not specify the stay times. This is obviously true in the trivial case of equal schedules, when all equal stay times result in matching, but it also holds for unequal schedules. We justify this claim by rewriting the matching law in terms of the stay times:

$$\frac{a_1+\tau}{a_2+\tau}=\frac{a_1+1-e^{-(a_2+2\tau)}}{\lambda a_2+1-e^{-\lambda(a_1+2\tau)}} \qquad (28)$$

Equation 28 has two unknowns, a_1 and a_2, so the stay times for a given λ and τ are not uniquely determined by the matching law. There is, in fact, an infinite set of stay times

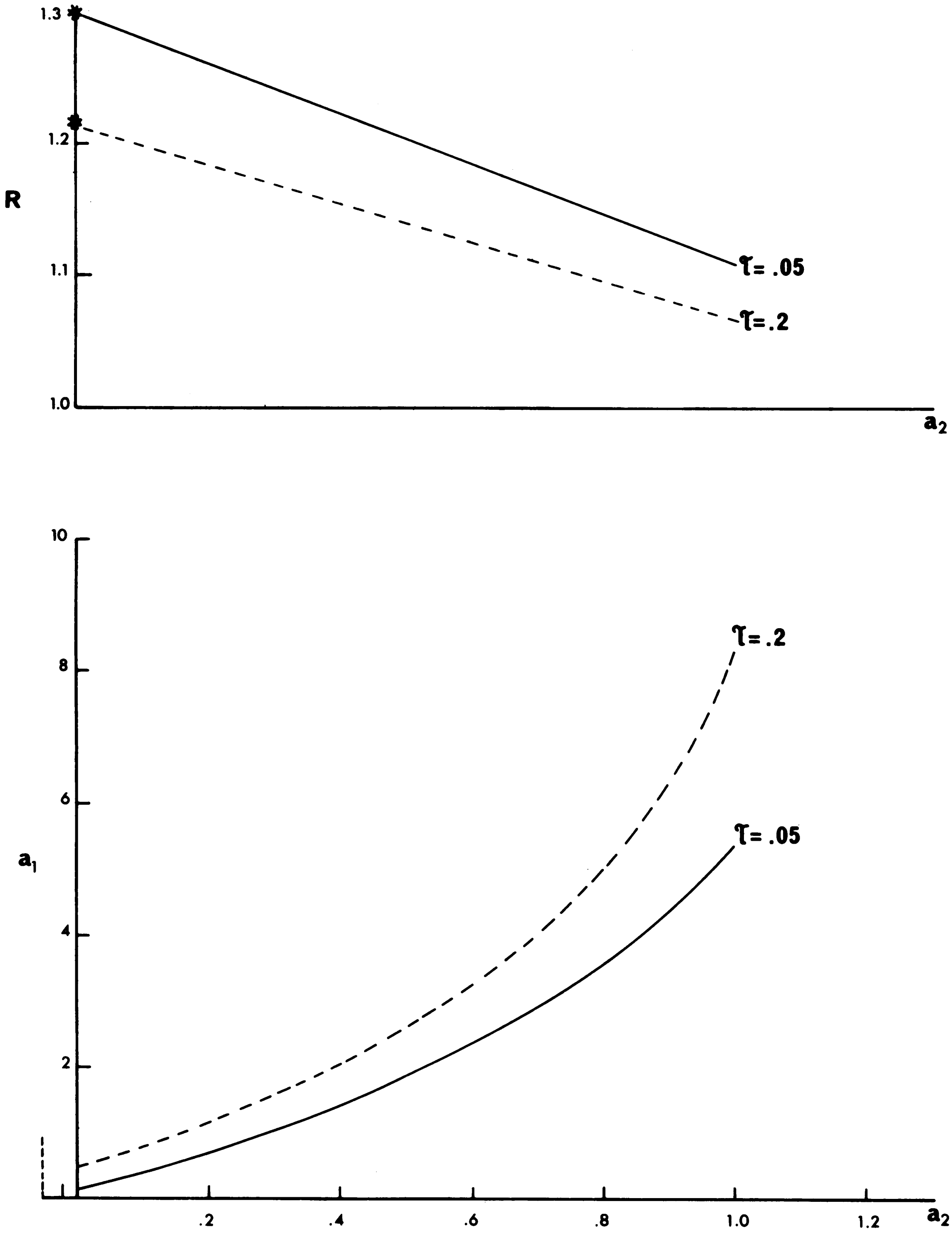

Fig. 6. Matching when $\lambda = .3333$ for two values of τ. The lower half shows the matching value of a_1 (from Equation 28) as a function of a_2. The optimal value of a_1 for each value of τ is shown offset at the extreme left. The upper half shows the reward resulting from matching to a given value of a_2. The optimal reward rate for each value of τ is shown by the stars. (The upper lines are drawn as straight but are actually very slightly curved.)

that comply with the matching law. We have investigated some of these stay times in the region of the optimal solution. A set of values of a_2 was chosen, and Equation 28 was used to obtain the value of a_1 that gave matching. Some results can be seen in Figure 6, which shows that when a_2 is at its optimal value of zero, matching performs quite well. As a_2 increases, however, the reward rate obtained from matching decreases. As an extreme example, on a conc VI 1-min VI 2-min with a 3-sec COD, matching holds when $a_1 \simeq 19$ min and $a_2 = 1.75$ min, but the reward rate (1.042 rewards/min) is not much above what could be obtained from staying on the better schedule. The optimal stay times would give an extra 24.6 rewards per hour.

Generality and Summary

Our model provides an accurate representation of the reward rate resulting from fixed stay times on conc VT VT schedules which have a negative exponential distribution of intervals, as described by Fleshler and Hoffman (1962). It was suggested above that the model can also be applied to VI schedules if it is assumed that the animal's rate of responding is high enough for it to obtain all the rewards as soon as they are set up. This assumption, made by Heyman and Luce (1979a), is tempting, but does not take any account of the possible energetic costs of making responses. Heyman and Luce (1979a) p. 138 present arguments for ignoring response costs, but it can be claimed that they oversimplify the issue. For example, McSweeney (1977) shows that there is a significant (albeit small) change in overall response rate with overall reward and discusses some drawbacks with previous investigations. Furthermore, response rates on each schedule are not equal, and vary with time since the changeover (Silberberg & Fantino, 1970).

There is a sense in which the evidence we have cited is beside the point, in that the argument used by Heyman and Luce (1979) can be opposed on purely theoretical grounds. To say that there are no response costs because animals behave as if there were none is to assume that animals are behaving optimally. This is implicitly an inverse optimality procedure (McFarland, 1977); i.e., it infers the optimality criterion from the observed behavior. But if behavior is not optimal, then the inference will not yield correct results (see Houston, 1980 for a general discussion).

The results of this section can now be summarized against the background of previous work. Our model is the first to incorporate the COD and to avoid any assumptions about the form of behavior. We derive the behavior (i.e., stay time on each schedule) that maximizes reward rate for all possible values of the two VI s and the COD.

The way the form of the optimal policy depends on λ and τ is shown in Figure 2. The results can be summarized as follows:

1) If $0 < \lambda \leqq \lambda_c$, then for $0 < \tau < \tau_c$, Stay-Switch is optimal; i.e., the animal switches to the worse schedule to collect rewards, but spends no time there. The time spent on Schedule 1 increases as τ increases. As τ tends to τ_c from below, the time spent on the better schedule tends to infinity, until when τ reaches τ_c, the optimal policy changes its form to remaining on the better arm forever.

Note that the point of transition from Stay-Switch to Never Switch does not depend on λ provided $\lambda \leqq \lambda_c$. The poorer schedule is chosen because, having waited sufficiently long on Schedule 1, a reward will almost certainly be received after a switch to Schedule 2. But λ is so small that it is not worth waiting on this schedule.

2) When $\lambda_c < \lambda < 1$, then for τ small $(0 < \tau < \tilde{\tau}(\lambda))$ the optimal policy is Stay-Switch. But as τ increases past $\tilde{\tau}(\lambda)$, although nothing dramatic happens to a_1^*, it becomes worth waiting on Schedule 2. As τ increases further, both a_1^* and a_2^* increase until as $\tau \uparrow \hat{\tau}$ (λ), we have $a_1^* \to \infty$ and a_2^* tends to some positive but finite limit. For $\tau > \hat{\tau}$ (λ) the optimal behavior is Never Switch.

3) For $\lambda = 1, \tilde{\tau}(\lambda) = 0$, and hence for $0 < \tau < \hat{\tau}$ $(1) = 1$ one waits on both schedules for a positive time.

All experiments of which we are aware fall in the region $0 < \lambda < \lambda_c$, $0 < \tau < \tau_c$. In this region matching in the form of Equation 3 is not in general optimal, but matching is close to the optimum if the stay time on the poorer schedule is close to zero. Provided τ is kept constant, a form of biased matching holds when behavior is optimal.

DISCRETE TRIAL CONC VI VI

Immediate and Overall Maximization

All the optimality analysis discussed so far is concerned with the behavior over the whole of the experimental session. In the literature on operant behavior, such a view is referred to as "molar", to distinguish it from "molecular" analysis, which is concerned with moment-to-moment decisions. Shimp (1969) claimed that momentary maximization of the probability of reward on VI schedules would result in matching. We believe that Shimp's use of the term 'optimal' in this context has led to an unwarranted opposition between molar (i.e., global) and molecular (i.e., immediate or momentary) maximization. We have pointed out elsewhere (McNamara & Houston, 1980) that immediate maximization fails to produce the global optimal policy for exploiting two unknown variable ratio schedules. (This is equivalent to what is known in decision theory as the two-arm bandit problem.) The relationship between immediate and global maximizing cannot be investigated in the continuous time model of the previous section because the COD means that the immediate probability of reward is zero when a switch is made. This section, therefore, considers a discrete trial variant of the conc VI VI paradigm and shows that immediate maximizing does not always maximize reward rate. The results obtained agree with those of Staddon, Hinson, and Kram (1981), who investigate a wide range of discrete choice paradigms.

The Form of the Optimal Policy

We now consider a completely different form of VI problem in which the animal is presented with a simultaneous choice between two keys at regular intervals in time. Associated with each key is a VI schedule that runs during trials and intertrial intervals. The schedule associated with a given key stops running when a reward is set up and the reward is held until a response is made on that key. In practice (e.g., Nevin, 1969) the trial is indicated by illuminating both keylights. A response to either key turns off both lights for the intertrial interval and a reward is delivered if one had been set up for the chosen key. In the terminology of Staddon et al. (1981) this is dual assignment with hold.

To represent this procedure we take the time between successive trials to be the unit of time, so trials occur at times $t = 0, 1, 2, \ldots\ldots$ Let T_i be the time between the collection of a reward from Key i and the next reward being set up. T_i is exponentially distributed with mean $1/\lambda_i$. Without loss of generality, we assume $\lambda_1 \geqq \lambda_2$. The earliest time at which a reward can be collected from Key i is N_i, where N_i is the smallest integer that is greater than or equal to T_i. It can be seen that

$$\text{Prob } (N_i > k) = e^{-\lambda_i k} = q_i^k, \text{ say,}$$
$$\text{where } q_i = e^{-\lambda_i} \text{ and } p_i = 1 - q_i.$$

(p_i is the probability of a reward being set up on Key i between successive trials on this key). Thus N_i has a geometric distribution with parameter p_i, i.e.,

$$Prob\ (N_i = n) = p_i q_i^{n-1}, \qquad n = 1, 2, 3, \ldots$$

Because the delay between trials occurs whether or not the animal switches between keys, the policy of always choosing the same key (which is analogous to the policy Never Switch in Section 2) will not be optimal unless either

$$p_1 = 1$$

or

$$p_2 = 0$$

These cases will be ignored as trivial. We assume that the optimal policy has the form:

(A) Choose Key 1 for n_1 trials ($n_1 \geqq 1$).
(B) Choose Key 2 for n_2 trials ($n_2 \geqq 1$).
(C) Return to (A).

An expression for the mean reward rate $R(n_1,n_2)$ under such a policy can be obtained by considering the two ways in which a reward can be set up on a key: when the animal returns to the key after choosing the other key or during the choices on the key. For Key 1 the animal returns to it after n_2 responses to Key 2. The probability of no reward for the first response on Key 1 is therefore $q_1^{n_2+1}$, so the probability of a reward is $1 - q_1^{n_2+1}$. The expected reward for the remaining $(n_1 - 1)$ responses on Key 1 is $p_1(n_1 - 1)$. Applying the same argument to Key 2 gives the following equation:

$$(n_1 + n_2)\, R\, (n_1,n_2) = p_1\, (n_1 - 1) + (1 - q_1^{n_2+1}) + p_2(n_2 - 1) + (1 - q_2^{n_1+1}),$$

which can be rearranged to give:

$$R\,(n_1,n_2) = 1 - \frac{q_1(n_1-1) + q_2\,(n_2-1) + q_1^{n_2+1} + q_2^{n_1+1}}{n_1+n_2}. \tag{29}$$

The optimal policy for this problem is to choose n_1 and n_2 such that $R\ (n_1,\ n_2)$, given by Equation 29, is maximized. Staddon et al. (1981) conjecture on the basis of their simulations that the optimal policy requires making only one trial on the poorer schedule before switching back to the better schedule. We prove this conjecture in Appendix 8; i.e., we establish that, for $0 < p_2 \leqq p_1 < 1$, the optimal policy $(n_1{}^*, n_2{}^*)$ always has $n_2{}^* = 1$. It is first shown that any policy with $n_2 \geqq 2$ can be improved upon. To do so we consider two cycles of a policy (n_1,n_2), i.e., a total of $2n_1 + 2n_2$ choices. It is then demonstrated that a series of choices of length $n_1 + 2n_2$ can be replaced by a series of the same length that ends in the same state but has a higher expected reward rate than the original series. Having proved that $n_2{}^*$ must be less than 2, we show that $n_2{}^*$ cannot be equal to zero, from which it follows that $n_2{}^* = 1$.

The fact that $n_2{}^* = 1$ simplifies the problem, in that $n_1{}^*$ is the value of n_1 that maximizes $R\ (n_1,\ 1)$. Equivalently, $n_1{}^*$ is the value of n_1 that minimizes

$$1 - R\,(n_1,\,1) = \frac{q_1\,(n_1-1) + q_1^2 + q_2^{n_1+1}}{n_1+1}$$

Some values of $n_1{}^*$ are given in Table 9. The dependence of R on n_1 is illustrated in Figure 8 of Staddon et al. (1981).

Immediate Maximization

Shimp (1969) suggested that organisms should behave optimally in the sense of always making the choice that momentarily has the greatest expected value. This principle is known as immediate or molecular maximization. In the case of the discrete VI problem we are considering here, immediate maximization predicts a fixed sequence of responses which involves approximate matching to the programmed reinforcement rates (Shimp, 1969; Staddon, 1980). We derive the matching relationship as follows. At the start, the choice is between p_1 and p_2 and, by definition $p_1 > p_2$ so Key 1 is chosen. The animal will continue to choose Key 1 until the probability of a reward from Key 2 is at least as great as p_1. Now after n_1 trials have been made on Key 1, the probability of a reward on Key 2 is $1 - q_2^{n_1+1}$. The condition for a switch to Key 2 is

$$p_1 \leqq 1 - q_2^{n_1+1}$$

but

$$p_1 = 1 - q_1$$
$$\therefore\ q_1 \geqq q_2^{n_1+1} \tag{30}$$

After one trial has been made on Key 2, the probabilities of reward are $1 - q_1^2$ and p_2 for Keys 1 and 2, respectively. As $1 - q_1^2 > 1 - q_1 = p_1 > p_2$, the next trial will be made on Key 1. In other words, immediate maximization requires that $n_2 = 1$. An approximation for n_1 can be obtained by considering equality to hold in Equation 30. By definition, $q_i = e^{-\lambda_i}$. Therefore

$$e^{-\lambda_1} = e^{-\lambda 2(n_1+1)}$$

and hence

$$n_1 + 1 = \lambda_1/\lambda_2$$

When equality holds in Equation 30, immediate maximizing predicts indifference between the two keys. If, at this point, Key 1 is chosen again instead of switching to Key 2, then

$$n_1 = \lambda_1/\lambda_2 \tag{31}$$

and, as $n_2 = 1$, Equation 31 can be written as

$$n_1/n_2 = \lambda_1/\lambda_2$$

which is a form of matching to the programmed rates of reinforcement. This relationship only holds when λ_1/λ_2 is an integer and Key 2 is chosen as soon as its probability of a reward is equal to (instead of greater than) the probability of Key 1.

It might be thought that when the immediate probability of reward is the same on each

Table 9

n^* (the optimal value of n_1, first number) and $n(im)$ (the value of n_1 predicted by immediate maximization, second number) for various values of q_1 and q_2.

	.9	28	21	20	15	15	11	11	8	8	6	6	4	4	3	2	2
	.8	13	10	9	7	6	5	5	4	3	3	2	2	1	1		
	.7	8	6	5	4	4	3	3	2	2	1	1	1				
	.6	5	4	3	3	2	2	2	1	1	1						
q2	.5	3	3	2	2	1	1	1	1								
	.4	2	2	1	1	1	1										
	.3	2	1	1	1												
	.2	1	1														
q1			.1		.2		.3		.4		.5		.6		.7		.8

key, either choice of key will give the same reward rate. The following example shows that this is not always the case. Nevin (1969) presented pigeons with a VI 1-min on Key 1 and a VI 3-min on Key 2. He used an intertrial interval of 6 sec, and as this interval defines our unit of time, we have

$$\lambda_1 = .1$$
$$\lambda_2 = .033$$
$$q_1 = e^{-.1} = .9048, \text{ and}$$
$$q_2 = e^{-.033} = .9673.$$

(Nevin actually used a uniform distribution of eleven intervals ranging from 10 sec to 110 sec for the VI 1-min and nine intervals ranging from 10 sec to 350 sec for the VI 3-min.) Immediate maximization predicts that the first and second responses should be made on Key 1. After these two choices, the probability of a reward on Key 2 is $1 - q_2^3 = p_1$, so the probabilities are equal. But, from Equation 29

$$R(2,1) = 1.23915, \text{ and}$$
$$R(3,1) = 1.24146.$$

Thus the reward rates differ, even though the immediate probabilities are the same. (In this case the difference is too small to be noticable to the animal, but it suffices to make the theoretical point.) For these values of q_1 and q_2, the policy of matching to programmed rates (i.e., $n_1 = 3$, $n_2 = 1$) is optimal. $R(4,1) = 1.24098$, so the reward rate does not vary much with n_1 around the optimum.

It is also possible to find cases in which immediate maximizing is ambiguous but neither sequence is optimal. For example, if $q_2 = .9$ and $q_1 = q_2^5 = .59049$, then the two immediate maximizing sequences, (4,1) and (5,1), do worse than both (6,1) and (7,1):

$$R(4,1) = .45787$$
$$R(5,1) = .45965$$
$$R(6,1) = .46008$$
$$R(7,1) = .45974$$

Once again all the policies are similar, but the reward rate obtained from immediate maximization can be improved on by staying for longer on the better arm. In fact it can be shown (see Appendix 9) that n_1^* is equal to or greater than $n(im)$, where $n(im)$ is the value of n_1 resulting from immediate maximization. Table 9 shows that n^* and $n(im)$ are identical when q_2 is only slightly bigger than q_1, but $n^* > n(im)$ when q_2 is much bigger than q_1. (See also Staddon et al. (1981), Figure 8.)

As Staddon et al. (1981) show, there are problems for which immediate maximization always results in the optimal reward rate, but the discrete VI paradigm is not one of them. It is possible, however, that immediate maximization could be used as a decision rule. For many values of p_1 and p_2, this results in optimal behavior, and even when it does not the reward rate is not much less than the optimum. Immediate maximization is thus a rule of thumb that approximates the optimal policy.

It must be stressed that both the optimal policy and immediate maximization involve a fixed sequence of choices, and that the sequence is always of the form: make n trials on the better schedule followed by one trial on the worse schedule then n on the better schedule, etc. For some values of p_1 and p_2, (see Table 9) the resulting value of n is the same, i.e., $n^* = n(im)$, so the existence of a fixed sequence, or (more realistically) sequential dependencies, will not distinguish between immediate maximization and optimization.

There is no consensus on how animals perform on the discrete VI paradigm we have discussed. Silberberg, Hamilton, Ziriax, and Casey (1978) found some correspondence between the predictions of immediate maximization and the sequential probabilities they observed, but some of the sequences that the birds produced could not have occurred on the basis of immediate maximization. In contrast to these results, Nevin (1979) reanalysed the data from Nevin (1969) and found that choice often ran counter to the predictions of immediate maximization.

Some differences between the procedures used by Nevin (1969) and Silberberg et al. (1978) are considered by Nevin (1979). From a theoretical point of view, the fact that different sorts of schedules were used is crucial. Our analysis applies only to the constant probability schedules based on Fleshler and Hoffman (1962), which were used by Silberberg et al. (1978). As Nevin (1969) points out, on the schedule he used, a reward becomes more likely as the number of consecutive unrewarded trials on the schedule increases. The choices predicted by immediate maximization on such schedules are, therefore, not those given by Silberberg et al. (1978) and used by Nevin (1979). A consideration of the intervals used by Nevin

(1969) shows that it is possible for immediate maximization to produce both a long run of consecutive choices to the better schedule and two consecutive choices to the worse schedule. Both cases involve sequences of unrewarded trials; immediate maximization requires a switch of schedule following a reward. This prediction is not borne out by Nevin's data—see Nevin (1969) Table 2. Furthermore, there seems to be no way for immediate maximization to predict three or more consecutive choices of the worse schedule, yet Nevin (1979) found that such sequences occurred.

Matching

It has been shown that immediate maximization results in approximate matching to the programmed rates. We now consider matching to obtained rates. If $n_2 = 1$, as is required by both optimality and immediate maximization, then matching means that the following equation holds

$$\frac{1 - q_1^2 + p_1(n-1)}{1 - q_2^{n+1}} = n \tag{32}$$

For many values of p_1 and p_2, Equation 32 will not have an integer solution, so a continuous approximation is adopted. If m is defined as the value of n that solves Equation 32, it can be shown that $m > n(im)$. Introducing m and rearranging Equation 32 yields

$$m(p_1 + q_2^{m+1} - 1) = p_1 + q_1^2 - 1$$

which means that

$$m = \frac{q_1 - q_1^2}{q_1 - q_2^{m+1}} \tag{33}$$

If the value of m given by Equation 33 is to be finite, then the following inequality must hold

$$q_1 > q_2^{m+1}$$

But the continuous approximation for immediate maximization defines $n(im)$ by the equation

$$q_1 = q_2^{n(im)+1}$$

It therefore follows that

$$q_2^{n(im)+1} > q_2^{m+1}$$

and hence

$$n(im) < m$$

Because m and $n(im)$ have not been constrained to take only integer values, this result must not be applied indiscriminately.

There may be no simple relationship between immediate maximization and the ratio of obtained rates since, as Staddon et al. (1981) point out, the ratio of the obtained rates depends on the absolute probabilities of reward, whereas Equation 24 shows that $n(im)$ is given in terms of the ratio of λ_1 to λ_2.

We have not obtained any analytic results concerning matching and n^*. Staddon et al. (1981) show by simulation that when $p_1 = .75$ and $p_2 = .25$ the optimal policy gives approximate matching over a wide range of n.

DISCUSSION

We have derived the behavior that maximizes reward rate on two VI paradigms, and have shown that matching does not specify the optimal policy. Now that the maximum possible reward rate for a given set of parameters is known, the performance of animals under these circumstances can be assessed. If VIs have some resemblance to the food distributions that animals encounter in the wild, then we should expect animals to perform well on them. Although the conc VI VI procedure was developed as a convenient laboratory technique for analyzing an animal's choices, it does indeed have a natural analogue. Davies and Houston (in press) suggest that the renewal of food items on the territory of a pied wagtail (*Motacilla alba yarrellii* Gould) can be represented by the following equation

$$N(t) = K(1 - e^{-bt}) \tag{34}$$

where

$N(t)$ = number of items at a point on the territory
t = time since point was last visited
b, L = constant for a given day and territory.

Equation 34 is identical to the expected reward on returning to a VI schedule after a time t away from it, given that the value of a reward is K. The VI paradigm can thus be identified with the problem of exploiting two renewing patches. The COD then becomes the time it takes to travel between the patches. In the VT paradigm, the animal can detect items as they appear in the patch, whereas the standard VI procedure requires attempts at prey capture to be made without knowledge of whether a prey item is present.

The reason for switching from one patch (or schedule) to the other is that the probability of a food item being present in the other patch increases with the time since the patch was last visited. The reward rate in a patch does not change with the time spent in it. In contrast, the problem analyzed by Charnov (1976) involves a reduction in capture rate as time in the patch increases. Thus circumstances that resemble VI schedules require switching because of renewal, while the circumstances considered by Charnov (1976) require switching because of depletion. Further discussion of the relationship between schedules and optimal foraging theory can be found in Staddon (1980).

The optimal policies we have obtained maximize the reward rate rather than the net rate of energy intake. If the rate at which energy is expended during switching is the same as the rate while on a schedule, maximizing reward rate is equivalent to maximizing net rate of energy intake. This equivalence is likely to hold for performance in a Skinner box, but not for traveling between patches in the wild. In general, if switching is energetically more costly than staying, the stay times that maximize net rate of energy intake will be longer than those that maximize reward rate.

Regardless of whether the empirical basis of matching is molar or molecular, it is clear from our analysis that the relative molar measures in Equations 1 to 3 are insufficient for a characterization of performance on VI schedules. It is the absolute value of the stay times that determines the reward rate, not their ratio. (This point can also be seen in the model of Heyman and Luce (1979), where the reward rate depends not just on the proportion of time on the better schedule, but on this proportion and the switching rate, as indicated by I.) A striking feature of the results is that for a given concurrent procedure, the optimal value of these absolute measures does not vary. To be specific, optimal performance on a conc VT VT schedule requires always waiting for a time a_1* on the better schedule and a time a_2* on the worse schedule. Adopting a range of stay times with mean a_1* and a_2* reduces the reward rate—see Appendix 1. Similarly, on the discrete VI paradigm the optimal policy involves repeating the sequence of n* trials on the better schedule followed by one trial on the worse schedule. It need hardly be said that animals do not show such regularity of behavior. Similar discrepancies between theory and data occur when optimal choice of prey items is investigated. Theory predicts all-or-nothing preferences, whereas animals sometimes take items they "ought" not to take (e.g., Krebs, Erichsen, Webber, & Charnov, 1977; Lea, 1979). One way to reconcile such results with optimality theory is to suggest that animals are sampling to take account of possible changes in their environment. We have discussed this idea and its implications at length elsewhere (McNamara & Houston, 1980). It amounts to saying that although the experimenter knows that the parameters of the schedule will remain constant during an experimental run, the animal does not "know" this. We do not wish to claim that this is the whole story. There will be limits to the accuracy of performance that an animal is capable of achieving. It is nevertheless important not to overlook the fact that the animal may not be "built" (by evolution) to perform the task set it by a laboratory experiment. We prefer to suggest that animals have relatively simple decision rules that perform quite well on problems similar to those the animal encounters in the wild (see Houston, 1980, for some examples). Herrnstein and Vaughan (1980) take the same view and discuss a rule ("melioration") which performs well on conc VI VI but badly on conc VI VR. It is well worth comparing various schedules in this way, because on VI schedules the reward rate is not very sensitive to changes in behavior. It would be interesting to know if behavior is less variable on schedules with sharper optima (c.f. the range of variation hypotheses put forward by Staddon, 1976).

Because of the insensitivity of reward rate on VIs to changes in behavior, it is possible for a policy which is quite different from the optimal policy to perform well. For example, Figure 8 of Staddon et al. (1981) shows that in the discrete VI procedure, reward rate is sometimes virtually independent of n. The same effect can be seen in the model of performance on interdependent conc VIs proposed by Heyman and Luce (1979). Although the optimal value of the proportion of time on the better schedule is often far from the matching value, their Figure 4 suggests that the loss in reward rate as a consequence of matching would be small.

We wish to emphasize that maximizing the probability of reward for each response does not necessarily maximize overall reward rate. Staddon (1968) pointed out the distinction, but its importance seems to have been overlooked. Like Staddon et al. (1981) we view immediate maximizing as a decision rule rather than an optimality principle in its own right. We have shown that in the discrete VI procedure, immediate maximizing is quite a good rule, and it will always be good if the optimality criterion is to maximize some discounted future reward rate rather than the overall reward rate. Staddon et al. (1981) show that immediate maximization can be used as a rule in continuous time procedures without a COD, and it is possible that some modified momentary principle could be used on schedules with a COD. In contrast, matching does not specify behavior—it is possible to find a range of types of behavior that result in matching for given schedule parameters (see Figure 6). This suggests that matching by itself cannot be determining behavior.

REFERENCE NOTE

1. Houston, A. I. Matching to obtained rates in the model proposed by Heyman and Luce. Unpublished manuscript.

REFERENCES

Allison, T. S., & Lloyd, K. E. Concurrent schedules of reinforcement: Effects of gradual and abrupt increases in changeover delay. *Journal of the Experimental Analysis of Behavior,* 1971, **16,** 67-73.

Baum, W. M. On two types of deviation from the matching law: Bias and undermatching. *Journal of the Experimental Analysis of Behavior,* 1974, **22,** 231-242.

Baum, W. M., & Rachlin, H. C. Choice as time allocation. *Journal of the Experimental Analysis of Behavior,* 1969, **12,** 861-874.

Brownstein, A. J., & Pliskoff, S. S. Some effects of relative reinforcement rate and changeover delay in response-independent concurrent schedules of reinforcement. *Journal of the Experimental Analysis of Behavior,* 1968, **11,** 683-688.

Catania, A. C. Concurrent operants. In W. K. Honig (Ed.) *Operant behavior: Areas of research and application.* New York: Appleton-Century-Crofts, 1966.

Charnov, E. L. Optimal foraging: The marginal value theorem. *Theoretical Population Biology,* 1976, **9,** 129-136.

Davies, N. B., & Houston, A. I. The costs and benefits of satellites on pied wagtail's territories. *Journal of Animal Ecology,* in press.

de Villiers, P. A. Choice in concurrent schedules and a quantitative formulation of the law of effect. In W. K. Honig & J. E. R. Staddon (Eds.) *Handbook of operant behavior.* Englewood Cliffs, N.J.: Prentice-Hall, 1977.

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

Herrnstein, R. J. Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Behavior,* 1961, **4,** 267-272.

Herrnstein, R. J. On the law of effect. *Journal of the Experimental Analysis of Behavior,* 1970, **13,** 243-266.

Herrnstein, R. J. Derivatives of matching. *Psychological Review,* 1979, **86,** 486-495.

Herrnstein, R. J., & Vaughan, W. Melioration and behavioral allocation. In J. E. R. Staddon (Ed.) *Limits to action: The allocation of individual behavior.* New York: Academic Press, 1980.

Heyman, G. M. Matching and maximising in concurrent schedules. *Psychological Review,* 1979, **86,** 496-500. (a)

Heyman, G. M. A Markov model description of changeover probabilities on concurrent variable-interval schedules. *Journal of the Experimental Analysis of Behavior,* 1979, **31,** 41-51. (b)

Heyman, G. M., & Luce, R. D. Operant matching is not a logical consequence of maximizing reinforcement rate. *Animal Learning and Behavior,* 1979, **7,** 133-140. (a)

Heyman, G. M., & Luce, R. D. Reply to Rachlin's comment. *Animal Learning and Behavior,* 1979, **7,** 269-270. (b)

Hinde, R. A. *Animal Behaviour.* New York: McGraw-Hill, 1970.

Houston, A. I. Godzilla v. the creature from the black lagoon. In F. M. Toates & T. R. Halliday. *The analysis of motivational processes.* London: Academic Press, 1980.

Johns, M., & Miller, R. G. Average renewal loss rate. *Annals of Mathematical Statistics,* 1963, **34,** 396-401.

Krebs, J. R. Optimal foraging: Decision rules for predators. In J. R. Krebs, & N. B. Davies, (Eds.) *Behavioural ecology.* Oxford: Blackwell Scientific Publications, 1978.

Krebs, J. R., Erichsen, J. T., Webber, M. I., & Charnov, E. L. Optimal prey selection in the great tit, *Parus major. Animal Behaviour,* 1977, **25,** 30-38.

Krebs, J. R., Kacelnik, A., & Taylor, P. Test of optimal sampling by foraging great tits. *Nature,* 1978, **275,** 27-31.

Lea, S. E. G. Foraging and reinforcement schedules in the pigeon: optimal and non-optimal aspects of choice. *Animal Behaviour,* 1979, **27,** 875-886.

Maynard Smith, J. Optimization theory in evolution. *Annual Review of Ecology and Systematics,* 1978, **9,** 31-56.

McDowell, J. J., & Kessel, R. A multivariate rate equation for variable-interval performance. *Journal of the Experimental Analysis of Behavior,* 1979, **31,** 267-283.

McFarland, D. J. Decision making in animals. *Nature,* 1977, **269,** 15-21.

McNamara, J., & Houston, A. I. The application of statistical decision theory to animal behaviour. *Journal of Theoretical Biology,* 1980, **85,** 673-690.

McSweeney, F. K. Sum of responding as a function of reinforcement on two-key concurrent schedules. *Animal Learning and Behavior,* 1977, **5,** 110-114.

Myerson, J., & Miezin, F. M. The kinetics of choice:

An operant systems analysis. *Psychological Review,* 1980, **87,** 160-174.

Nevin, J. A. Interval reinforcement of choice behaviour in discrete trials. *Journal of the Experimental Analysis of Behavior,* 1969, **12,** 875-885.

Nevin, J. A. Overall matching versus momentary maximizing: Nevin (1969) revisited. *Journal of Experimental Psychology: Animal Behavior Processes* 1979, **5,** 300-306.

Oster, G. F., & Wilson, E. O. *Caste and ecology in the social insects.* Princeton: Princeton University Press, 1978.

Pliskoff, S. S. Effects of symmetrical and asymmetrical changeover delays on concurrent performances. *Journal of the Experimental Analysis of Behavior,* 1971, **16,** 249-256.

Rachlin, H. A molar theory of reinforcement schedules. *Journal of the Experimental Analysis of Behavior,* 1978, **30,** 345-360.

Rachlin, H. Comment on Heyman and Luce: "Operant matching is not a logical consequence of maximizing reinforcement rate". *Animal Learning and Behavior,* 1979, **7,** 267-268.

Shimp, C. P. Optimal behavior in free-operant experiments. *Psychological Review,* 1969, **76,** 97-112.

Shull, R. L., & Pliskoff, S. S. Changeover delay and concurrent schedules: Some effects on relative performance measures. *Journal of the Experimental Analysis of Behavior,* 1967, **10,** 517-527.

Silberberg, A., & Fantino, E. Choice, rate of reinforcement, and the changeover delay. *Journal of the Experimental Analysis of Behavior,* 1970, **13,** 187-197.

Silberberg, A., Hamilton, B., Ziriax, J. M., & Casey, J. The structure of choice. *Journal of Experimental Psychology: Animal Behavior Processes,* 1978, **4,** 368-398.

Staddon, J. E. R. Spaced responding and choice: A preliminary analysis. *Journal of the Experimental Analysis of Behavior,* 1968, **11,** 669-682.

Staddon, J. E. R. Learning as adaptation. In W. K. Estes (Ed.) *Handbook of Learning and Cognitive Processes* (Vol. 2). New York: Erlbaum Associates, 1976.

Staddon, J. E. R. On Herrnstein's equation and related forms. *Journal of the Experimental Analysis of Behavior,* 1977, **28,** 163-170.

Staddon, J. E. R. Optimality analyses of operant behavior and their relation to optimal foraging. In J. E. R. Staddon (Ed.) *Limits to action: The allocation of individual behavior.* New York: Academic Press, 1980.

Staddon, J. E. R., Hinson, J. M., & Kram, R. Optimal choice. *Journal of the Experimental Analysis of Behavior,* 1981, **35,** 393-408.

Staddon, J. E. R., & Motheral, S. On matching and maximizing in operant choice experiments. *Psychological Review,* 1978, **85,** 436-444.

Staddon, J. E. R., & Motheral, S. Response independence, matching and maximising: a reply to Heyman. *Psychological Review,* 1979, **86,** 501-505.

Tinbergen, N. The study of instinct. Oxford: Clarendon Press, 1951.

Toates, F. M. *Animal behaviour: A systems apprach.* Chichester, Eng.: John Wiley, 1980.

Todorov, J. C. Concurrent performances: Effect of punishment contingent on the switching response. *Journal of the Experimental Analysis of Behavior,* 1971, **16,** 51-62.

Received June 5, 1980
Final acceptance January 7, 1981

APPENDIX 1

Variable Stay Times

In this appendix we allow the stay times to be nonnegative random variables which we denote by A_1 and A_2. Let $R(A_1,A_2)$ denote the corresponding mean reward rate. We will show that the policy of using the random stay times A_1 and A_2 can be improved by replacing these times by their mean values; i.e., we will show

$$R(A_1,A_2) < R(a_1,a_2), \qquad \text{(A1.1)}$$

where

$$a_i = E(A_i). \qquad \text{(A1.2)}$$

The mean reward per cycle using times A_1 and A_2 is given by

$$\begin{aligned} &E(\text{Reward/cycle}) \\ &= E[\lambda_1 A_1 + \lambda_2 A_2 + 1 - e^{-\lambda_1(A_2+\tau_1+\tau_2)} \\ &\quad + 1 - e^{-\lambda_2(A_1+\tau_1+\tau_2)}] \\ &= \lambda_1 a_1 + \lambda_2 a_2 + 2 - E[e^{-\lambda_1(A_2+\tau_1+\tau_2)} \\ &\quad + e^{-\lambda_2(A_1+\tau_1+\tau_2)}], \end{aligned} \qquad \text{(A1.3)}$$

and the mean time per cycle is given by

$$\begin{aligned} E[\text{Time/cycle}] &= E[A_1 + A_2 + \tau_1 + \tau_2] \\ &= a_1 + a_2 + \tau_1 + \tau_2. \end{aligned} \qquad \text{(A1.4)}$$

By a standard result in renewal theory (Johns & Miller, 1963)

$$R(A_1,A_2) = \frac{E[\text{Reward/cycle}]}{E[\text{Time/cycle}]}\,.$$

Thus, by Equations A1.3 and A1.4

$$\begin{aligned} (a_1 + a_2 + \tau_1 + \tau_2)R(A_1,A_2) &= \lambda_1 a_1 + \lambda_2 a_2 + 2 \\ &\quad - E[e^{-\lambda_1(A_2+\tau_1+\tau_2)} \\ &\quad + e^{-\lambda_2(A_1+\tau_1+\tau_2)}]. \end{aligned} \qquad \text{(A1.5)}$$

Now by Jensen's inequality.

$$\begin{aligned} &E[e^{-\lambda_1(A_2+\tau_1+\tau_2)} + e^{-\lambda_2(A_1+\tau_1+\tau_2)}] \\ &\qquad > e^{-\lambda_1(a_2+\tau_1+\tau_2)} + e^{-\lambda_2(a_1+\tau_1+\tau_2)} \end{aligned}$$

(The inequality is strict provided A_1 and A_2 are not both constant.) Thus by Equation A1.5

$$\begin{aligned} &(a_1 + a_2 + \tau_1 + \tau_2)R(A_1,A_2) \\ &\qquad < \lambda_1 a_1 + \lambda_2 a_2 + 2 - e^{-\lambda_1(a_2+\tau_1+\tau_2)} \\ &\qquad\quad - e^{-\lambda_2(a_1+\tau_1+\tau_2)}; \end{aligned}$$

but the right hand side of this equation equals $(a_1 + a_2 + \tau_1 + \tau_2)R(a_1,a_2)$, therefore $R(A_1,A_2) < R(a_1,a_2)$ as required.

APPENDIX 2

The Case $\lambda = 1, \tau < 1$

Without loss of generality we restrict attention to the case when $a_1 = a_2$. Let $f(a) = R(a,a)$, then by Equation 14

$$(a + \tau)f(a) = a + 1 - e^{-(a+2\tau)} \qquad \text{(A2.1)}$$

and thus

$$(a + \tau)f'(a) + f(a) = 1 + e^{-(a+2\tau)} \qquad \text{(A2.2)}$$

Equations A2.1 and A2.2 give $f'(a) = 0$ if and only if

$$(\tau + a + 1)e^{-(a+2\tau)} = 1 - \tau. \qquad \text{(A2.3)}$$

Now let

$$h(a) = (\tau + a + 1)e^{-(a+2\tau)}.$$

Then

$$h'(a) = -(\tau + a)e^{-(a+2\tau)} < 0$$

and $h(a) \to 0$ as $a \to \infty$; thus Equation A2.3 has a solution for $a > 0$ if and only if $h(0) > 1 - \tau$. Furthermore Equation A2.3 can have at most one solution for $a \geqq 0$. We set

$$g(\tau) = h(0) - \tau + 1 = (\tau + 1)e^{-2\tau} + \tau - 1. \qquad \text{(A2.4)}$$

Then $g(0) = 0$ and

$$g'(\tau) = 1 - (1 + 2\tau)e^{-2\tau} > 0$$

since $e^{2\tau} > 1 + 2\tau$ for $\tau > 0$. Thus $g(\tau) > 0$ for $\tau > 0$, and hence $h(0) > 1 - \tau$ by Equation A2.4. This shows that the Equation $f'(a) = 0$ has a unique solution a^* for $a \geqq 0$ and that furthermore $a^* > 0$. Now by Equation A2.2

$$(a + \tau)f''(a) + 2f'(a) = -e^{-(a+2\tau)}.$$

Thus, $f''(a^*) < 0$, and f has a local maximum at a^*. Since f has no other turning value for $a \geqq 0$, we have

$$f(a^*) = \max_{a \geqq 0} f(a).$$

Finally, note that by Equation A2.2

$$f(a^*) = 1 + e^{-(a^*+2\tau)} > 1,$$

and hence it is always worth switching.

We have seen that a^* satisfies Equation A2.3, i.e.,

$$(\tau + a^* + 1)e^{-(a^*+2\tau)} = 1 - \tau. \quad \text{(A2.5)}$$

We use this expression to find an approximation for a^* for small τ. Let $x = a^* + 2\tau$, then Equation A2.5 can be re-expressed as

$$1 - \tau = \frac{x}{e^x - 1}. \quad \text{(A2.6)}$$

This equation shows that x is small for small τ. Expanding in powers of x gives (to second order in x)

$$1 - \tau \simeq x\left(x + \frac{x^2}{2} + \frac{x^3}{6}\right)^{-1}$$

$$= \left(1 + \frac{x}{2} + \frac{x^2}{6}\right)^{-1}$$

$$\simeq 1 - \left(\frac{x}{2} + \frac{x^2}{6}\right) + \left(\frac{x}{2} + \frac{x^2}{6}\right)^2$$

$$\simeq 1 - \frac{x}{2} + \frac{x^2}{12}.$$

Thus

$$a^* \equiv x - 2\tau \simeq \frac{x^2}{6}. \quad \text{(A2.7)}$$

It follows that a^* is of order x^2 and hence $x = 2\tau$ to first order of x. Putting this back into Equation A2.7 gives

$$a^* \simeq \frac{2\tau^2}{3}$$

for small τ.

APPENDIX 3

The Optimal Policy Involves Staying Longer on the Better Arm

We will show $a_1^* > a_2^*$ in Region II provided $0 < \lambda < 1$.

We first established that $\frac{\partial R}{\partial a_1}(0,0) > 0$. This will show that if $a_2^* = 0$, then $a_1^* > 0$. By Equations 7 and 8 we have

$$4\tau^2 \frac{\partial R}{\partial a_1}(0,0)$$
$$= 2\tau + 2\lambda\tau\, e^{-2\lambda\tau} - 2 + e^{-2\tau} + e^{-2\lambda\tau}.$$

Let

$$f(\lambda,\tau) = 2\tau + 2\lambda\tau\, e^{-2\lambda\tau} - 2 + e^{-2\tau} + e^{-2\lambda\tau}.$$

Then $f(\lambda,0) = 0$. Also

$$\frac{\partial f}{\partial \tau}(\lambda,\tau) = 2(1 - 2\lambda^2\tau e^{-2\lambda\tau} - e^{-2\tau}).$$

But $e^{2\lambda\tau} > 1 + 2\lambda\tau$ and $e^{2\tau} > 1 + 2\tau$, thus

$$\frac{\partial f}{\partial \tau}(\lambda,\tau) > \frac{4\tau(1-\lambda)(1+\lambda+2\lambda\tau)}{(1+2\lambda\tau)(1+2\tau)} > 0,$$

since $\lambda < 1$. Hence $\frac{\partial R}{\partial a_1}(0,0) \equiv f(\lambda,\tau) > 0$ for $\tau > 0$ and $0 < \lambda < 1$.

It remains to consider the case when $a_2^* > 0$. Now by Equations 8 and 9 we have

$$(a_1 + a_2 + 2\tau)\left(\frac{\partial R}{\partial a_1} - \frac{\partial R}{\partial a_2}\right)$$
$$= 1 - e^{-(a_2+2\tau)} - \lambda(1 - e^{-\lambda(a_1+2\tau)}).$$

Thus if $\lambda(a_1 + 2\tau) \leqq a_2 + 2\tau$, then

$$\frac{\partial R}{\partial a_1}(a_1,a_2) > \frac{\partial R}{\partial a_2}(a_1,a_2).$$

In particular, we see that if $a_2^* > 0$ and $\lambda(a_1^* + 2\tau) \leqq a_2^* + 2\tau$, then

$$\frac{\partial R}{\partial a_1}(a_1^*,a_2^*) > 0.$$

This contradicts the equation

$$\frac{\partial R}{\partial a_1}(a_1^*,a_2^*) \leqq 0.$$

Thus,

$$\lambda(a_1^* + 2\tau) > a_2^* + 2\tau;$$

and hence $a_1^* > a_2^*$, since $\lambda < 1$.

APPENDIX 4

I vs. II Boundary

Consider fixed λ with $0 < \lambda < 1$. Suppose the optimal policy is of the form Stay-Stay for some τ. As τ is increased we expect the stay times to increase until, for some critical value of τ which we denote by $\hat{\tau}$ (λ), the optimal policy involves staying an infinite time on the better schedule: i.e., Never Switch. We will find the critical value $\hat{\tau}(\lambda)$.

In the Stay-Stay region we have $\frac{\partial R}{\partial a_2}(a_1{}^*,a_2{}^*) = \frac{\partial R}{\partial a_2}(a_1{}^*,a_2{}^*) = 0$. This condition gives

$$R(a_1{}^*,a_2{}^*) = 1 + \lambda e^{-\lambda(a_1{}^*+2\tau)} \qquad \text{(A4.1)}$$

and

$$R(a_1{}^*,a_2{}^*) = \lambda + e^{-(a_2{}^*+2\tau)}. \qquad \text{(A4.2)}$$

We also have

$$\begin{aligned}(a_1{}^* + a_2{}^* + 2\tau)&R(a_1{}^*,a_2{}^*)\\ &= a_1{}^* + \lambda a_2{}^* + 2 - e^{-(a_2{}^*+2\tau)} - e^{-\lambda(a_1{}^*+2\tau)}\end{aligned} \qquad \text{(A4.3)}$$

Now as $\tau \uparrow \hat{\tau}(\lambda)$ we expect $a_1{}^* \to \infty$, thus by Equation A4.1

$$R(a_1{}^*,a_2{}^*) = 1 + o\left(\frac{1}{a_1{}^*}\right). \qquad \text{(A4.4)}$$

Then by Equations A4.2 and A4.4

$$e^{-(a_2{}^*+2\tau)} = 1 - \lambda + o\left(\frac{1}{a_1{}^*}\right). \qquad \text{(A4.5)}$$

Thus, as $a_1{}^* \to \infty$, $a_2{}^*$ tends to a finite limit which we denote by $\hat{a}_2$. By Equation A4.5

$$\hat{a}_2(\lambda) + 2\hat{\tau}(\lambda) + \log(1 - \lambda) = 0. \qquad \text{(A4.6)}$$

However, by Equations A4.4, A4.5, and A4.3 we have

$$a_2{}^*(1 - \lambda) = 1 + \lambda - 2\tau + o\left(\frac{1}{a_1{}^*}\right),$$

and hence

$$\hat{a}_2(\lambda)\,(1 - \lambda) = 1 + \lambda - 2\hat{\tau}(\lambda). \qquad \text{(A4.7)}$$

Solving Equations A4.6 and A4.7 gives

$$\hat{\tau}(\lambda) = \frac{1}{2\lambda}\,(1 + \lambda + (1 - \lambda)\log(1 - \lambda)) \qquad \text{(A4.8)}$$

and

$$\hat{a}_2(\lambda) = -\frac{1}{\lambda}\,[1 + \lambda + \log(1 - \lambda)]. \qquad \text{(A4.9)}$$

These equations hold provided the optimal policy is of the form Stay-Stay for τ just less than $\hat{\tau}(\lambda)$, and this holds provided $\hat{a}_2(\lambda) > 0$. By Equation A4.9 this holds if and only if

$$1 + \lambda + \log(1 - \lambda) < 0;$$

which in turn hold provided $\lambda > \lambda_c$, where λ_c is the unique solution (for $0 < \lambda < 1$) of the equation

$$1 + \lambda + \log(1 - \lambda) = 0. \qquad \text{(A4.10)}$$

A calculation shows

$$\lambda_c = .841405.$$

The corresponding value for $\tau_c \equiv \hat{\tau}(\lambda_c)$ is

$$\tau_c = .920703.$$

For $\lambda \leqq \lambda_c$ we have $a_2 \leqq 0$. We interpret this to mean that there is a transition from Stay-Switch to Never Switch as τ increases. We again denote the transition point by $\hat{\tau}(\lambda)$.

In the Stay-Switch region we set $\frac{\partial R}{\partial a_1}(a_1{}^*,0) = 0$; this gives

$$R(a_1{}^*,0) = 1 + e^{-\lambda(a_1{}^*+2\tau)} \qquad \text{(A4.11)}$$

and

$$\begin{aligned}(a_1{}^* + 2\tau)&R(a_1{}^*,0)\\ &= a_1{}^* + 2 - e^{-2\tau} - e^{-\lambda(a_1{}^*+2\tau)}.\end{aligned} \qquad \text{(A4.12)}$$

By Equation A4.11

$$R(a_1{}^*,0) = 1 + o\left(\frac{1}{a_1{}^*}\right),$$

and hence, by Equation A4.12

$$2\tau = 2 - e^{-2\tau} + o\left(\frac{1}{a_1{}^*}\right).$$

Taking the limit as $\tau \uparrow \hat{\tau}(\lambda)$ gives

$$2 - 2\hat{\tau}(\lambda) - e^{-2\hat{\tau}(\lambda)} = 0. \qquad \text{(A4.13)}$$

It can be verified, either by numerical calculation or by using Equations A4.8 and A4.10, that Equation A4.13 has the solution $\hat{\tau}(\lambda) = \tau_c$. Thus, $\hat{\tau}(\lambda) = \tau_c$ for all λ such that $0 < \lambda \leqq \lambda_c$.

APPENDIX 5

The Stay-Stay vs. Stay-Switch Boundary

We seek the point of transition from an optimal solution with $a_2^* = 0$ to one with $a_2^* > 0$. For given λ let $\tilde{\tau}(\lambda)$ denote the value of τ at which this transition occurs. To find this value we solve the equations $\frac{\partial R}{\partial a_1}(a_1^*,a_2^*) = \frac{\partial R}{\partial a_2}(a_1^*, a_2^*) = 0$, together with the equation $a_2^* = 0$. These equations give

$$R(a_1^*,0) = 1 + \lambda e^{-\lambda(a_1^*+2\tilde{\tau})} \qquad \text{(A5.1)}$$

and

$$R(a_1^*,0) = \lambda + e^{-2\tilde{\tau}}. \qquad \text{(A5.2)}$$

Define $S = S(\lambda)$ by

$$S(\lambda) = \frac{\lambda + e^{-2\tilde{\tau}(\lambda)} - 1}{\lambda},$$

then by Equations A5.1 and A5.2

$$S = e^{-\lambda(a_1^*+2\tilde{\tau})}. \qquad \text{(A5.3)}$$

Equation A5.1 then reduces to

$$R(a_1^*,0) = 1 + \lambda S. \qquad \text{(A5.4)}$$

However, we also know that the rate $R(a_1^*,0)$ is given by

$$(a_1^* + 2\tilde{\tau})R(a_1^*,0) = a_1^* + 2 - e^{-2\tilde{\tau}} - e^{-\lambda(a_1^*+2\tilde{\tau})}.$$

Substituting for R and $e^{-\lambda(a_1^*+2\tilde{\tau})}$ from Equations A5.4 and A5.3 then gives

$$\lambda(a_1^* + 2\tilde{\tau})S + S = 2 - 2\tilde{\tau} - e^{-2\tilde{\tau}}.$$

Finally, by Equation A5.3, $\lambda(a_1^* + 2\tilde{\tau}) = -\log S$. Therefore

$$S(1 - \log S) = 2 - 2\tilde{\tau} - e^{-2\tilde{\tau}}.$$

APPENDIX 6

Optimal Stay Times in the Stay-Stay Region

In the Stay-Stay region the optimal stay times are found by setting $\frac{\partial R}{\partial a_1}(a_1^*,a_2^*) = \frac{\partial R}{\partial a_2}(a_1^*,a_2^*) = 0$. These equations give

$$R(a_1^*,a_2^*) = 1 + \lambda e^{-\lambda(a_1^*+2\tau)} \qquad \text{(A6.1)}$$

and

$$R(a_1^*,a_2^*) = \lambda + e^{-(a_2^*+2\tau)}. \qquad \text{(A6.2)}$$

It is convenient to use the variables p_i where

$$p_1 = 1 - e^{-(a_2+2\tau)}$$

and

$$p_2 = 1 - e^{-\lambda(a_1^*+2\tau)}.$$

Then Equations A6.1 and A6.2 give

$$R(a_1^*,a_2^*) = 1 + \lambda - p_1 \qquad \text{(A6.3)}$$

and

$$p_1 = \lambda p_2 \qquad \text{(A6.4)}$$

Now the rate R is also given by

$$(a_1^* + a_2^* + 2\tau)R(a_1^*,a_2^*) = a_1^* + a_2^*\lambda + p_1 + p_2.$$

Thus, by Equations A6.3 and A6.4

$$(1 - p_2)\lambda(a^*_1 + 2\tau) + (1 - p_1)(a_2^* + 2\tau) = 2\tau p_1 - p_1 - p_2$$

But $\lambda(a_1^* + 2\tau) = \log(1 - p_2)$ and $(a_2^* + 2\tau) = \log(1 - p_1)$, therefore

$$(1 - p_1)\log(1 - p_1) + (1 - p_2)\log(1 - p_2) = 2\tau p_1 - p_1 - p_2 \qquad \text{(A6.7)}$$

If we set $p_1 = \lambda p_2$ in this equation we obtain an equation for p_2 alone. This can be solved numerically by computer and hence p_1, a_1^*, and a_2^* can be found.

APPENDIX 7

Matching in the Stay-Switch Region

We show that (a) matching is not in general optimal in this region; and (b) that a form of biased matching holds.

(a) If matching were optimal in the Stay-Switch region, then we would have

$$\frac{a^* + \tau}{\tau} = \frac{a_1{}^* + 1 - e^{-2\tau}}{1 - e^{-\lambda(a_1{}^* + 2\tau)}} \qquad \text{(A7.1)}$$

Recall that $y = (a_1{}^* + 2\tau)$ is a function of τ alone and does not depend on λ (see Equation 19). We can rewrite Equation A7.1 in terms of y to give

$$\frac{y/\lambda}{\tau} = \frac{y/\lambda - 2\tau + 1 - e^{-2\tau}}{1 - e^{-\nu}},$$

which can be rearranged as

$$\frac{y(1 - e^{-\nu}) - y\tau}{\tau(e^{-2\tau} + 2\tau - 1)} = \lambda. \qquad \text{(A7.2)}$$

However, the left hand side of this equation depends on τ but not λ and the right hand side depends on λ but not τ. Thus Equation A7.2 cannot hold in general, and hence matching is not in general optimal in the Stay-Switch region.

(b) In the Stay-Switch region we have

$$\frac{R_1{}^*}{R_2{}^*} = \frac{a_1{}^* + 1 - e^{-2\tau}}{1 - e^{-\lambda(a_1{}^* + 2\tau)}}:$$

which can be expressed as

$$(1 - e^{-y})\frac{R_1{}^*}{R_2{}^*} = (a_1{}^* + \tau) + (1 - \tau - e^{-2\tau}),$$

and hence as

$$\frac{a_1{}^* + \tau}{\tau} = k(\tau)\frac{R_1{}^*}{R^{2*}} + c(\tau) \qquad \text{(A7.3)}$$

where

$$k(\tau) = \frac{1 - e^{-\nu}}{\tau}$$

and

$$c(\tau) = \frac{e^{-2\tau} + \tau - 1}{\tau}.$$

Equation A7.3 can produce graphs that are indistinguishable from straight lines when log $(a_1{}^* + \tau)/\tau$ is plotted against log $R_1{}^*/R_2{}^*$. For example, with $\tau = .0166$ and a range of $R_1{}^*/R_2{}^*$ from 3 to 8, a linear regression gave a slope of 1.12 and an intercept of .16 with $r^2 = .99$.

APPENDIX 8

Form of the Optimal Policy for the Discrete Model

We show that $n_2 = 1$ for the optimal policy.

Consider a policy with parameters n_1 and n_2 where $n_2 \geqq 2$. We will show that this policy can be improved upon. Suppose a switch has just been made to the second key. On this policy the next $2n_2 + n_1$ trials involve $n_2 - 1$ further trials on Key 2, followed by n_1 trials on Key 1, followed by n_2 on Key 2, followed by a trial on Key 1. The total expected reward M for these $2n_2 + n_1$ trials is

$$M = n_1 + 2n_2 - [(n_1 - 1)q_1 + 2(n_2 - 1)q_2 + 2q_1{}^{n_2+1} + q_2{}^{n_1+1}]$$

Now consider replacing this block of $2n_2 + n_1$ trials by the following sequence. The first n_1 trials are made on Key 1; thereafter trials alternate between the two keys starting on Key 2 at the $n_1 + 1$-th trial and ending on Key 1 at the $2n_2 + n_1$-th trial. Let $\tilde{M}$ denote the total expected reward for this block of trials. We have

$$\tilde{M} = n_1 + 2n_2 - [(n_1 - 1)q_1 + (n_2 + 1)q_1{}^2 + (n_2 - 1)q_2{}^2 + q_2{}^{n_1+1}].$$

Then

$$\tilde{M} - M = 2(n_2 - 1)q_2 - (n_2 - 1)q_2{}^2 + 2q_1{}^{n_2+1} - (n_2 + 1)q_1{}^2.$$

We show that $\tilde{M} - M > 0$. To this end let

$$f(x) = 2(n_2 - 1)q_2 - (n_2 - 1)q_2{}^2 + 2x^{n_2+1} - (n_2 + 1)x^2.$$

Since $0 < q_1 \leqq q_2$ it is sufficient to establish that $f(x) > 0$ for $0 < x \leqq q_2$. Now

$$f'(x) = 2(n_2 + 1)(x^{n_2} - x) \leqq 0 \text{ for } 0 < x < 1.$$

Hence $f(x) \geqq f(q_2)$ for $0 < x < q_2$ and thus it is sufficient to establish that $f(q_2) > 0$; i.e.,

$$2q_2(q_2^{n_2} - n_2q_2 + n_2 - 1) > 0.$$

Let $g(z) = z^{n_2} - n_2z + n_2 - 1$.
Then

$$g'(z) = n_2z - n_2 < 0 \quad \text{for } 0 < z < 1$$

and hence $g(z) > g(1) = 0$ for $0 < z < 1$. Since $0 < q_2 < 1$ it follows that $f(q_2) = 2q_2g(q_2) > 0$. This established that $\tilde{M} > M$.

Note that the final state at the end of these two alternative blocks of $2n_2 + n_1$ trials is the same. Thus any policy with $n_2 \geqq 2$ can be improved upon and cannot be optimal. It is easy to see that the policy with $n_2 = 0$ cannot be optimal either. The reward rate under this policy is $1 - q_1$; but if we choose n_1 so that

$$q_2^{n_1+1} < q_1,$$

then

$$R(n_1,1) = \frac{1 - q_1^2 + (n_1 - 1)(1 - q_1) + 1 - q_2^{n_1+1}}{n_1 + 1}$$

$$> \frac{(n_1 + 1)(1 - q_1)}{n_1 + 1} = 1 - q_1.$$

It follows that the optimal policy must have $n_2 = 1$.

APPENDIX 9

Immediate Maximization and Optimality

In this appendix it is proved that the optimal policy involves making at least as many trials on the better schedule before switching as are made under immediate maximization.

Let n satisfy

$$1 - q_2^{n+1} < 1 - q_1. \qquad \text{(A9.1)}$$

It will be shown that this implies

$$R(n + 1,1) > R(n,1). \qquad \text{(A9.2)}$$

This can be used to show that $n(im) \leqq n^*$. To see this suppose $n < n(im)$. Since $n(im)$ is the first integer k for which $1 - q_2^{k+1} \geqq 1 - q_1$ we must have $1 - q_2^{n+1} < 1 - q_1$. Then if Equation A9.1 implies A9.2 we can deduce that $R(n + 1,1) > R(n,1)$, and hence $n \neq n^*$. This will have shown that $n \neq n^*$ for all $n < n(im)$, which implies $n^* \geqq n(im)$.

To establish that Equation A9.1 implies Equation A9.2 we proceed as follows. From Equation 29 we have

$$\begin{aligned}(n + 1)(n + 2)[R(n + 1,1) - R(n,1)] \\ = q_1^2 - 2q_1 + q_2^{n+1}(n + 2) - q_2^{n+2}(n + 1) \\ = q_2^{n+1}((n + 1)p_2 + 1) - q_1 (1+ p_1)\end{aligned}$$

The right hand side of this equation can be rearranged to give

$$\begin{aligned}D \equiv (q_2^{n+1} - q_1)((n + 1)p_2 + 1) \\ + q_1((n + 1)p_2 - p_1).\end{aligned} \qquad \text{(A9.3)}$$

We will show that $D > 0$. First consider the function f given by

$$f(x) = (1 - x)^{n+1} - (1 - (n + 1)x),$$

where $n \geqq 0$. Then $f(0) = 0$ and

$$f'(x) = (n + 1)[1 - (1 - x)^n].$$

Thus $f'(x) > 0$ for $1 > x > 0$ and hence $f(x) > 0$ for $1 > x > 0$. Setting $x = p_2$ gives

$$q_2^{n+1} > 1 - (n + 1)p_2.$$

Rearranging and taking p_1 from each side gives

$$(n + 1)p_2 - p_1 > q_1 - q_2^{n+1}.$$

Substituting this expression for $(n + 1)p_2 - p_1$ in Equation A9.3 gives

$$\begin{aligned}D > (q_2^{n+1} - q_1)((n + 1)p_2 + 1) \\ + q_1(q_1 - q_2^{n+1}),\end{aligned}$$

which gives

$$D > (q_2^{n+1} - q_1)[(n + 1)p_2 + p_1].$$

Now, by Equation A9.1, we have $q_2^{n+1} - q_1 > 0$. Thus $D > 0$. This shows that Equation A9.1 implies Equation A9.2.