Behavioural Processes, 20 (1989) 189–197
Elsevier

RESPONSE SELECTION IN OPERANT LEARNING

J E R Staddon and Y Zhang
Department of Psychology,
Duke University, Durham, NC 27706, USA

(Accepted 8 September 1989)

ABSTRACT

J E R Staddon and Y Zhang, 1989 Response selections in operant learning
Behav.Process.,20:139-197

We show that simple, contiguity-based, nonassociative response-selection process provides a qualitative account for both anomalous and nonanomalous properties of operant conditioning The process can easily be extended to permit associative effects, it may therefore represent the initial processing stage for all conditioning in higher vertebrates

Key Words reinforcement, learning model, stochastic, superstition, instinctive drift

Simple learning — operant and classical conditioning — is thought to be the key to the evolution of human intelligence, hence has been the focus of much neurobiological research with vertebrate and invertebrate animals (e g , McGaugh, Weinberger & Lynch, 1985) Most emphasis has been placed on classical conditioning; nevertheless operant conditioning — the modification of freely occurring ("emitted") behavior by rewarding or punishing consequences (*reinforcement*) — is probably the more primitive function (cf Staddon, 1983)

The way that operant reinforcement selects one activity over others — the *assignment-of-credit* problem (Minsky, 1961, Sutton, 1984) — is taken for granted in most learning models (Sternberg, 1963) which are examples of *supervised* learning, in which a "teacher" (reinforcement) is assumed to present explicit error information to the learning system in a way that has no obvious biological parallel (Crick, 1988) With a few exceptions (e g , Sutton, 1984) assignment of credit is satisfactorily solved for learning theorists by the intuitively appealing but unformalized mechanism of temporal contiguity Unfortunately, contiguity theory is violated by phenomena such as "superstitious" behavior (Skinner, 1948, Staddon & Simmelhag, 1971, Timberlake & Lucas, 1985) and instinctive drift (Breland & Breland, 1961), which are anomalous either because the activity contiguous with reinforcement is not the one actually strengthened or because activities continue to occur despite response-independent reinforcement These anomalies have never been reconciled with contiguity theory (Gardner & Gardner, 1988)

0376-6357/89/$03 50

We show that the simplest possible formal contiguity model for the assignment- of-credit problem in operant conditioning also provides a natural explanation for supposed exceptions

Standard effects. Positive reinforcement is defined by its *selective* effect when the occurrence of a reinforcer is made to depend on the occurrence of a particular response (out of two or more possibilities), the response probability should increase, and when the reinforcer is no longer presented, or is presented independently of responding, response probability should decline In addition to this property of *selection*, reinforcement is also sensitive to *delay*, and *contingency*

Our model is based on assumptions that are already well accepted, even though their effects in combination have not been fully understood arousal and adaptation, the idea that reinforcement transiently energizes a range of activities (Killeen, Hanson & Osborne, 1978), strength and competition, the idea that each activity has a certain tendency to occur and that the strongest will win, and variability, the notion of a repertoire of activities These five properties are captured in two linear discrete-time equations We define for each behavior a variable, V_1, its *strength* The competition rule is winner-take-all the activity with highest strength is the one that occurs (this is the only nonlinearity) The equations describe the changes in V values from one discrete-time instant to the next in the absence of reinforcement, or following reinforcement

After nonreinforcement $$V_1(t+1) = a_1 V_1(t) + \in (1-a_1),\ 0 \leq a_1 \leq 1, \quad (1)$$

After reinforcement $$V_1(t+1) = a_1 V_1(t) + \in (1-a_1) + b_1 V_1(t), \quad (2)$$

where $V_1(t)$ is the strength of the ith activity in discrete-time instant t, a_1 and b_1 are parameters that depend on *both* the activity *and* the reinforcer and $\in$ is a random variable sampled independently for each activity in each time instant Term $a_1 V_1(t)$ represents adaptation or short-term memory (STM) because $a_1 < 1$, this term reduces to zero with repeated iterations Term $b_1 V_1(t)$ represents the arousal effect of a hedonic event, which we assume acts on *all* activities If $b_1 > 0$, the effect is to increase V_1 (a positive reinforcer), if, $b_1 < 0$, the effect is to reduce V_1 (a punisher)

Note that the relation $a_1 + b_1 < 1$ must hold if V_1 is not to rise without limit in repeated iterations of Equation 2

Consider first the case of two or more identical activities (i e , $a_1 = a_2$, $b_1 = b_2$), which permits derivation of all the standard reinforcement properties In the absence of reinforcement, because the two parameters are the same for both activities each activity will occur equally often, on average If positive reinforcement is delivered for each occurrence of Activity 1, then at that instant by the highest-wins competition rule $V_1 > V_2$, hence the increment to 1, bV_1, must be greater than the increment to 2, bV_2 If the reinforcement occurs frequently enough that the increment in V_1 does not decay to zero by the next reinforcement V_1 will be steadily incremented relative to V_2, so that Activity 1 will come to dominate This conclusion holds whichever activity is reinforced, thus the process satisfies the *selection* condition

The essential feature of the reinforcement mechanism is that reinforcement always adds *some* increment to *all* activities, but the *largest* increment goes to the highest–V activity

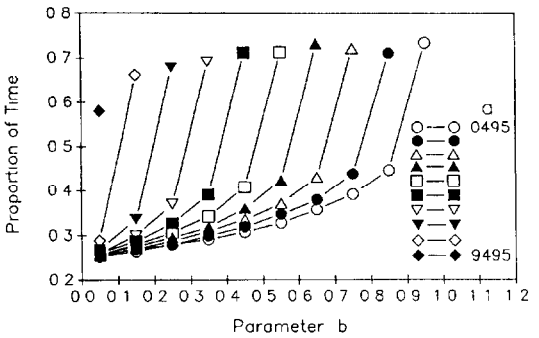

Figure 1 Parameter Space Simulation of the effects of 100% reinforcement on one of four identical (same a and b) activities for a range of a_i and b_i values (0 0495–0 9495 in increments of 0 1) Ordinate shows the proportion of time (iterations) taken up by the reinforced activity Note that in absence of any reinforcing effect each activity should take up about 25% of the time Each point in this and the next two figures is the average of $2x10^5$ iterations The random variable $\in$ in eqs 1 & 2 had a rectangular distribution over the interval 0 – 1

Figure 1 shows the asymptotic effects of reinforcing every occurrence of one behavior, in a set of four, for a wide range of parameter pairs The reinforced behavior is always facilitated, and the proportion of time taken up increases with increases in either parameter We have obtained similar results for ensembles of 2, 4 and 8 identical activities

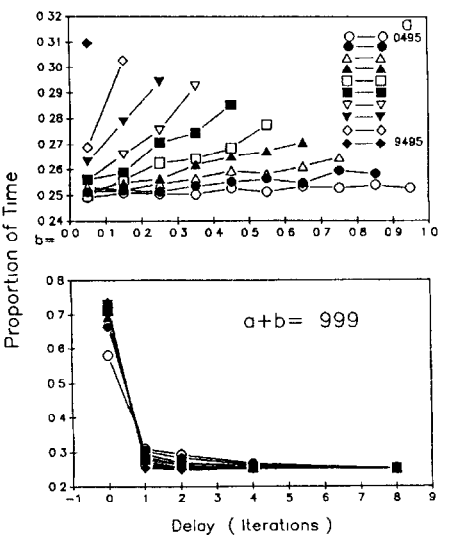

Figure 2 Effect of reinforcement delay Top panel Simulation of the effects of 100% reinforcement on one of four identical activities for a range of a and b values when the reinforcement is delayed one iteration Bottom panel Delay–of–reinforcement gradient (1–8 iterations) for a range of parameter pairs for which $a + b = 0\ 999$

Figure 2 shows the effects of reinforcement *delay* The same increasing pattern is seen in the parameter–space plot (Top), but now STM–parameter a has the greatest effect (compare the highest point on each curve the profile is decreasing in Fig 2, but increasing in Fig 1) The higher the a value, the less likely that the delayed reinforcer, when it actually occurs, will strengthen a behavior other than the target behavior The Bottom panel of Fig 2 shows selected data from the Top panel plotted in standard delay–of– reinforcement–gradient form The declining pattern is relatively independent of the values of a and b so long as they sum to a constant

Note that if two identical activities with high a values are both reinforced, but with a small difference in delay, then the more contiguous activity will eventually predominate The system is potentially very sensitive to small contiguity differences Animals are also (Killeen, 1978), which is paradoxical given the existence of so–called "superstitious" behavior, which has sometimes been interpreted as a failure to discriminate much larger contiguity differences Our model is consistent with both "superstition" (as we show in a moment) and high sensitivity to contiguity, hence can resolve this paradox

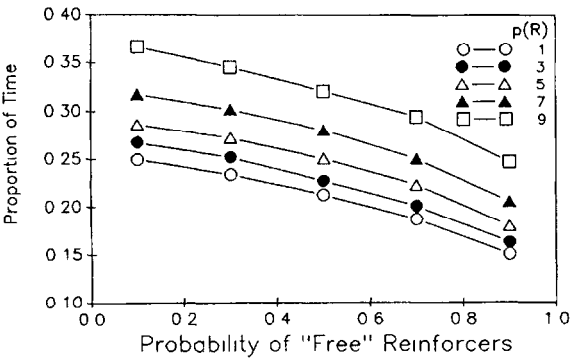

Figure 3 Effects of contingency Each curve shows the effect on the level of the reinforced behavior of variation in the probability of response– independent reinforcement, for a given probability of response–dependent reinforcement Parameter values $a = b = 0\ 4495$

Contingency is the fact that the strengthening effect of reinforcement depends on its *correlation* with the reinforced behavior, not just on contiguity Thus, if the target behavior is reinforced intermittently, or if it is reinforced every time but reinforcement also occurs at other times, the behavior will be strengthened less than if it is reinforced exclusively Figure 3 shows that our model has both these properties Vertical comparison shows the effect of the probability of response–dependent reinforcement the higher the probability, the larger the proportion of time taken up by the reinforced activity Horizontal comparison shows the effect of "free" (response–independent) reinforcers the higher this probability, the lower the level of the explicitly reinforced act

Anomalous effects. Most anomalies arise when activities have *different* parameter values We discuss three superstitious behavior, differential conditionability, represented by typologies, such as the distinction between emitted and elicited behavior, and instinctive drift

"Superstitious" behavior is typically produced when an animal such as a pigeon is given periodic response–independent food reinforcement If food delivery is frequent enough, the animal develops various kinds of vigorous stereotypies, despite the fact that food delivery is independent of its behavior Our model is too simple to account for the associative and temporal properties of this phenomenon, but the model can under restricted conditions produce apparently stereotyped behavior, even with a set of identical activities An example is shown as the four cumulative response records on the Left in Fig 4 The curves show the levels of four identical activities, with high–a and low–b values, under continuous reinforcement: Over any given epoch on the order of 100–1000 iterations the distribution of activities is highly skewed, with one activity tending to predominate Examples are indicated by the vertical lines over the epoch labeled "A", for example, Activity 1 predominates, whereas during epoch B, Activity 2 is the dominant one Eventually, every activity will have its place in the sun, but over a limited epoch, one activity will seem to have been preferentially strengthened

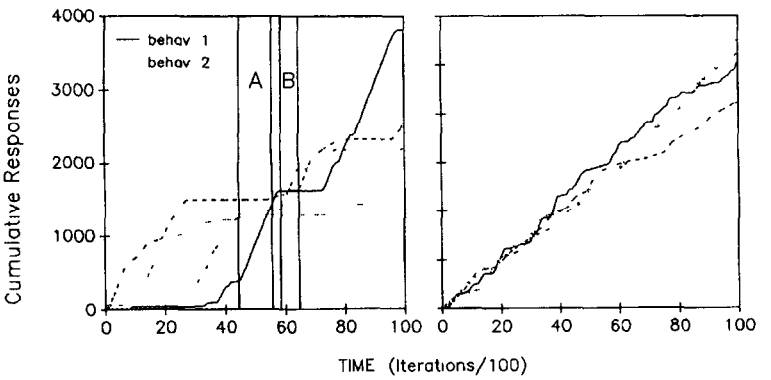

Figure 4 Effects of absolute probability of "free" reinforcement ("superstitious" behavior) Left panel Simulated cumulative records of four identical activities (a = 0 9495, b = 0 0495) when reinforcement occurs during each iteration Right panel Cumulative records when reinforcement occurs with p(R) = 0 5

The Right panel in Fig 4 shows the effect of reducing reinforcement probability the epochs over which behavior appears stereotyped are now much shorter

The stereotypy illustrated in the Left panel is the outcome of a positive–feedback process that resembles a suggestion of Skinner (1948) Skinner's explanation for all superstitious behavior is that when a (response– independent) reinforcer is delivered some behavior will be occurring and will be automatically strengthened, if the next reinforcer follows soon enough, the same behav-

ior will still be occurring and will be further strengthened, and so on, until the behavior appears to dominate

Two ingredients are lacking in this account an explicit statement about the temporal properties of behavior that are necessary for the process to work (parameters a and b in the present model), and recognition of the importance of the time window over which observations are carried out Our analysis shows that (i) the STM parameter, a, must be very high for the process to work at all, even with reinforcers occurring at the maximum rate (ii) The time window is critical if observations are continued for long enough, no activity will predominate Thus, the effect of "free" reinforcers is simply to increase the autocorrelation of activities, rather than to single out one over others permanently (Staddon & Horner, 1989) And (iii) even a modest reduction in reinforcement probability causes a sharp reduction in stereotypy Since reinforcement is in fact intermittent in all published demonstrations of superstition, Skinner's process, at least in our version of it, is unlikely to be responsible We suggest an alternative in a moment

Nevertheless, this analysis of superstition shows how our model accounts for the reliable finding of *hysteresis* in operant conditioning the fact that once a response has been strengthened via response–contingent reinforcement, it may persist even when reinforcers are delivered independently of responding (Herrnstein, 1966) Given a high enough a value, even intermittent "free" reinforcers may be enough to maintain an already–strengthened activity, at least for a while, and for longer than if reinforcement were withdrawn entirely

The distinction between *emitted* and *elicited* behavior parallels the procedural difference between operant and classical conditioning (Skinner, 1948) elicited behavior (salivation is an example) is behavior elicited by a reinforcer but not modifiable by operant conditioning; emitted behavior (lever pressing is an example) is not usually elicited by the reinforcer, but is modifiable by operant conditioning (the rat learns to press the lever to get food)

The existence of a typology is usually a sign that we lack understanding of the underlying process It is interesting, therefore, that the distinction between these two types of behavior emerges in a natural way from our model Recall that parameters a and b must add to less than one if V is to be bounded Moreover, as Fig 1 illustrates, the STM parameter a strongly determines the effect of operant (consequential) reinforcement when a is small, even large values for the arousal parameter, b, have little effect The importance of a is exaggerated by delay or intermittence of reinforcement when a is low, even the highest b value is insufficient to permit much strengthening of the behavior by operant reinforcement (cf Fig 2) Moreover, since $a + b < 1$, a highly elicitable (high–b) activity must have a low a value, hence will be only weakly susceptible to strengthening by operant reinforcement Conversely, activities with high a values will be readily conditionable, but will not generally be elicitable, because they must also have low b values Thus, the dichotomy between emitted and elicited behavior may be a consequence of the complementary relation between the arousal and STM parameters that is forced by stability considerations

Instinctive drift is perhaps the most striking exception to a contiguity account of reinforcement Breland and Breland (1961) reported several instances that conform to the following pattern Behavior A (e g , a raccoon putting a wooden egg onto a chute) is successfully "shaped" by response–contingent food reinforcement, but after a while Behavior A is supplanted by Behavior

B ("washing" the egg between the animal's forepaws), which is part of the animal's natural foraging repertoire Behavior B is inappropriate in this context because it competes with Behavior A on which food delivery actually depends Behavior A is contiguous with the reinforcer, but Behavior B is ultimately the one strengthened

A variety of naturalistic interpretations have been offered for these effects (and for the related phenomenon of "superstition"), but they follow from our model on the assumption that "instinctive" behaviors are characterized by very high a values (STM persistence) Given a set of activities with moderate a and b values, and one activity with a very high a value (and nonzero b), reinforcement of one of the moderate–a activities will cause it to predominate initially (because it has a higher b value than the "instinctive" activity) But, because increments to the high–a activity cumulate more effectively than the (larger) increments to the reinforced activity, it may predominate eventually, even if it is never contiguous with reinforcement

These effects are illustrated in the cumulative records in the top panel of Fig 5 Each set of four records shows four activities, three with moderate a and b values, one with a high a value and low b value The Left records show the free–operant (unreinforced) levels of the four activities (the low frequency of the high–a activity is a consequence of statistical properties of the model not relevant to the present point) The Center and Right panels show the effect of reinforcing one of the low–a activities with a probability of 0 25 or 1 00 The increasing reinforcement probability has two effects it causes the level of the reinforced activity ("R") to increase above the level of the other two low–a activities, but it causes a disproportionate increase in level of the high–a activity, which is predominant when p(R) = 1 (Right) The two bottom panels in Fig 5 show a magnified picture of initial acquisition in the p(R) = 1 condition, illustrating the transition from dominance by the reinforced, low–b, activity to predominance of the "instinctive", high–a, activity If a low a value activity is reinforced, it may predominate initially, but will be supplanted by the high–a value activity, even if it has a lower "arousability" (b value)

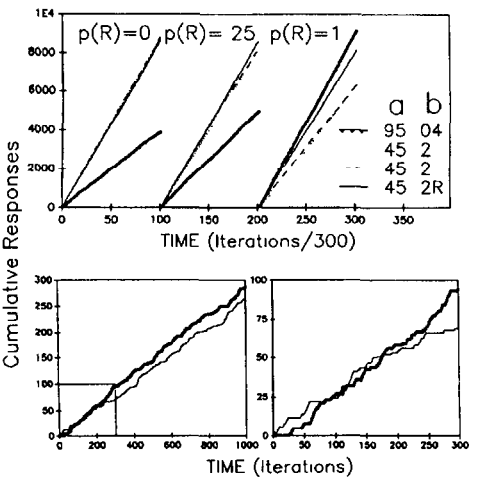

Figure 5 Instinctive drift Top panel, Left Cumulative records of four activities in the absence of reinforcement (operant levels) high–a, low–b activity is the least frequent (see parameter values in the table) Center The effect of reinforcing one of the three intermediate–a activities (a = 0 45, b = 0 2R) with probability 0 25 Right Effect of reinforcing every occurrence of the intermediate activity Bottom panel Magnified pictures of the beginning of the record in the Right panel at the top ("instinctive" and reinforced activities only)

Unsignaled shock-avoidance. Avoidance behavior in the absence of a "safety" signal has always posed a problem for the contiguity view of reinforcement because there is nothing for the avoidance response to be contiguous with An explicit reinforcing event is lacking in what is termed a shock- rate-reduction schedule, for example (Herrnstein & Hineline, 1966) In such a schedule brief shocks are delivered with probability p(Sh) = x_s each second or so If the animal makes the designated avoidance response, p(Sh) is reduced to a value $x_R < x_s$ until the next shock occurs, when it returns to x_s Rats learn (with some difficulty) to make the avoidance response on such procedures, despite the fact that no tangible event occurs contiguous with the avoidance response Our model can accommodate this behavior if it is assumed that electric shock and other aversive events reduce, rather than increase, V values (i e , $b < 0$) In such a case electric shocks reduce all V values, but the V value of the avoidance response will be reduced less than others' because on average it is followed by shock after a longer delay Hence, it will be favored in the competition and, in an ensemble of identical activities, will come to predominate

CONCLUSION

The standard steady-state, context-free properties of operant conditioning — selection, delay-of-reward, contingency and unsignaled avoidance — are all consistent with a simple assignment-of-credit mechanism that assumes each activity can be represented by a "strength" variable, V, that corresponds to filtered noise Each V value is transiently augmented or decremented by pleasant or unpleasant events in an amount proportional to its value The activity with the highest V value is the one that actually occurs These qualitative predictions are parameter-free

By allowing different activities to have different time constants and arousal effects, the major anomalies can also be explained

The present model is nonassociative hence independent of stimulus context Since the operant behavior of vertebrates is always context dependent, our model is obviously only part of the whole story, hence it is hard to know what to make of a number of partial exceptions (e g , the pecking response is both elicited and operantly conditionable, intermittent reinforcement sometimes sustains more behavior than continuous, etc) — since they may reflect later stages in the process What we have shown is that there is a surprising correspondence between the qualitative properties of this very simple response-selection process and the steady-state properties of operant behavior — suggesting that such a process may form the "front end" for the complex neural processing involved in the operant and classical conditioning of higher animals

ACKNOWLEDGEMENTS

We thank Jennifer Higa, Peter Holland, Nancy Innis, Peter Killeen and Allister Reid for comments on an earlier version Research supported by grants from the NSF

REFERENCES

Breland, K & Breland, M (1961) The misbehavior of organisms *American Psychologist*, **16**, 681–664

Bush, R R , & Mosteller, F (1961) A mathematical model for simple learning. *Psychological Review*, **68**, 313–23

Crick, F (1989) The recent excitement about neural networks *Nature*, **337**, 129–32

Gardner R A , & Gardner, B T (1988) Feedforward versus feedbackward An ethological alternative to the law of effect *Behavioral and Brain Sciences* , **11**, 3, 429–493

Herrnstein, R J , & Hineline, P N (1966) Negative reinforcement as shock-frequency reduction *Journal of the Experimental Analysis of Behavior*, **9**, 421–30

Herrnstein, R J (1966) Superstition A corollary of the principles of operant conditioning *In Operant behavior* (W K. Honig ed) New York Appleton–Century–Crofts (pp 33–51)

Killeen, P R , Hanson, S J , & Osborne, S R (1978) Arousal its genesis and manifestation as response rate *Psychological Review*, **85**, 571–81

Killeen, P R (1978) Superstition a matter of bias, not detectability *Science*, **199**, 88–90

Minsky, M (1961) Steps towards artificial intelligence *Proceedings of the Institutes of Radio Engineers*, **49**, 10–30

Skinner, B F (1938) *The behavior of organisms* New York Appleton–Century–Crofts

Skinner, B F (1948) "Superstition" in the pigeon *Journal of Experimental Psychology*, **38**, 168–72

Staddon, J E R., & Horner, J M (1989) Stochastic choice models A comparison between Bush–Mosteller and a source–independent reward–following model *Journal of the Experimental Analysis of Behavior*, **52** , 57–64

Staddon, J E R , & Simmelhag, V L (1971) The "superstition" experiment, A reexamination of its implications for the principles of adaptive behavior *Psychological Review*, **78**, 3–43

Staddon, J E. R (1983) *Adaptive behavior and learning*, New York Cambridge University Press

Sternberg, S (1963) Stochastic learning theory In *Handbook of mathematical psychology* (R. D Luce, R R Bush, & E Galanter eds) New York Wiley, 1963, Vol 2, pp 1–120

Sutton, R S (1984) Temporal credit assignment in reinforcement learning Unpublished PhD dissertation, Dept of Computer and Information Science, U of Massachusetts

Timberlake, W, & Lucas, G A (1985) The basis of superstitious behavior· Chance contingency, stimulus substitution, or appetitive behavior? *Journal of the Experimental Analysis of Behavior*, **44**, 279–99

Weinberger, N M , McGaugh, J L , & Lynch, G (1985) *Memory systems of the brain* New York Guilford Press