The Behavior Analyst 1998, 21, 219–240 No. 2 (Fall)

Experimental Design: Problems in Understanding the Dynamical Behavior–Environment System

Michael Davison
Auckland University, New Zealand

In this paper, I attempt to describe the implications of dynamical approaches to science for research in the experimental study of behavior. I discuss the differences between classical and dynamical science, and focus on how dynamical science might see replication differently from classical science. Focusing on replication specifically, I present some problems that the classical approach has in dealing with dynamics and multiple causation. I ask about the status and meaning of "error" variance, and whether it may be a potent source of information. I show how a dynamical approach can handle the sort of control by past events that is hard for classical science to understand. These concerns require, I believe, an approach to variability that is quite different from the one most researchers currently employ. I suggest that some of these problems can be overcome by a notion of "behavioral state," which is a distillation of an organism's history.

Key words: experimental design, tactics of scientific research, replication, dynamical systems, multiple causation, behavioral state, data analysis

---

My purpose in this paper is to try to understand how the changing conception of science brought about by recent developments in the analysis of complex interacting time-based systems—dynamical systems—may change the way in which we look at data and do research in the experimental analysis of behavior. Because I have more worries than answers, I hope that this paper will start a discussion of these points and others, and help the evolution of a new viewpoint in the experimental analysis of behavior. I shall first discuss the differences between classical and dynamical science, and then derive some implications for our viewpoint.

In his discussion of the classical approach to science, Ruelle (1993) presents the following quotation from Laplace (1814):

> An intelligence which, at a given instance, would know all the forces by which Nature is animated, and the respective situation of all the elements of which it is composed, if furthermore it were vast enough to submit all these data to analysis, would in the same formula encompass the motions of the largest bodies of the universe, and those of the most minute atom: nothing for it would be uncertain, and the future as well as the past would be present to its eyes. The human mind, in the perfection that it has been able to give to astronomy, provides a feeble semblance of this intelligence. (p. 29)

Laplace here nicely represents the classical science paradigm. With complete knowledge of the present and complete understanding of how the world works, we would be able to explain the past and to predict the future perfectly. In such a mechanistic nirvana, there would be no probability statements, and no hedging of bets. Although the realization of this mechanistic manifesto had always clearly been technically impossible, the goal remained, in classical science, one to strive for, and in

Thanks to all of the following: The symposium on Methodology and Quantitative Analysis (Barbara Wanchisen, Chair, and contributors Henry Pennypacker, James Johnston, and Peter Killeen) at the Association for Behavior Analysis annual convention (Washington, DC, June, 1995) who contributed many thoughts to the work here; Murray Sidman for his groundbreaking *Tactics of Scientific Research*; Barbara Wanchisen, Glynn Owens, Susan Schneider, and four reviewers who commented on drafts of this work; Ivan Beale, whom I consulted on applied behavior analysis topics; my graduate students who have, over the years, been browbeaten with many of the ideas herein, and who themselves contributed greatly.

Reprints may be obtained from Michael Davison, Department of Psychology, University of Auckland, Private Bag 92019, Auckland, New Zealand (E-mail: m.davison@auckland.ac.nz).

small domains scientists worked effectively towards achieving limited goals. The approach affected psychology in its early days, too, in the development of psychophysics and in the experimental analysis of behavior. In the latter, Sidman's (1960) *Tactics of Scientific Research* consisted of the affirmation of Laplace's views within psychology.

Science, though, has changed since 1960. Over the last 35 years, it is almost as if a new persona of science has been developed. Classical science and its findings have not been overturned, usually, because these findings well represent stable-state end points of some reactions between physical things (be they chemicals or the behavior of people). What has been realized is that some reactions do not have predictable, even stable, end points, and that sometimes the interaction of completely determinate systems—ones with no random elements in them at all—can and will produce behavior that appears to be random and *chaotic*. Their behavior results from dynamic, time-based interactions between simple systems. The student affects the rat, and the rat affects the student. Their interaction in a laboratory class may result in the rat turning on the student to the experimental analysis of behavior or turning him or her off. It may result in the rat being turned on to bar pressing or turned off. The detailed futures of the rat and of the student are not predictable, though their general futures may be.

Classical science, as defined by Laplace, clearly cannot accommodate such findings, and the findings were, in a sense, cast to one side for many years and placed in the "too hard box," with, naturally, an expectation that one day they would succumb to classical science. It now appears that they will not. Although chemistry and physics, as classically conceived, could find and investigate large areas of determinism, or predictability, to mine, it seems that some more complex sciences like, I will argue, the science of behavior are likely to find few areas that are amenable to classical science. Because of the transactive nature of these more nonclassical sciences, it is likely that many of their "reactions" will not stabilize in predictable regions, and that the modern, *dynamical*, scientific approach will need to be taken much more often. Over vast tracts of these nonclassical sciences, which are clearly more complex and difficult than physics and chemistry, classical science will not work.

I will discuss why this may be, and try to show some pointers on how we might reconceive our manifesto in the experimental analysis of behavior.

The Classical Science Approach

In classical science, noise is a nuisance. It subverts the clear demonstration of the effects of different levels of independent variables, and of different independent variables. Classical approaches to science thus focus on ways of experimentally or statistically eliminating, minimizing, and controlling noise. Using one or more of these methods, classical science is able to display the underlying systematic effect (it is usually singular) present in data, and the noise, whatever it was caused by, is cast on the garbage heap of science, without any hint of recycling.

The classical science approach usually assumes a single attractor—a point of stability—in the sense that the action of an independent variable on the system will lead the system willy-nilly to the same stable point each time the independent variable is applied. It assumes that the effects of a particular independent-variable level will lead behavior to a definite and predictable stability point from wherever the system is started. For example, take the chemical model. Silver nitrate plus sodium chloride gives silver chloride (a precipitate) and sodium nitrate: $NaCl + AgNO_3 \rightarrow AgCl + NaNO_3$. There is a stable result for this reaction, and the amount of silver chloride product de-

pends on the relative amounts of the initial reagents and the temperature and pressure and so on—and maybe in some reactions, the presence of a catalyst. The results of the reaction, and the amounts of the product, are predictable and do not depend on the history of the reagents, for example, on how the sodium chloride was produced. The reagents are *historically pure*. The situation in which the experiment is done is also pure, a clean test tube, and again is *historically clean*.

There is, however, a dynamical transaction between the reagents and their products. It is not the case that all silver nitrate is converted to silver chloride, and the reverse reaction to produce silver nitrate from sodium nitrate plus silver chloride also takes place: $NaCl + AgNO_3 \leftrightarrow AgCl + NaNO_3$. The forward and reverse reactions, however, are in a dynamic balance, and the attractor—the stability point of the system—is unitary, well defined, and stable. The system at stability can be dealt with as if it were *not* a dynamic transaction for the purposes of stable-state prediction.

But, in some areas of classical science, there are multiple attractors in a system. This means that there is more than one stable condition at which the reaction may end. Thus, if some chemicals or mixtures are heated and then cooled, for instance, the cooling does not continue in a linear fashion. The temperature fall ceases or is reversed at some times as different reactions or states are reached which themselves produce heat. The rate of heat loss is attenuated at these stages, and they are at least transiently stable. Many chemical mixtures (e.g., glass) exist in a semistable state (glass is a supercooled liquid), and states can be changed very rapidly (by a stone through a windshield).

However, it remains that classical science as a process requires that the reagents are historically pure—they contain no evidence at all (or very minimal evidence) of where they came from—and that the situation in which the reagents are mixed is historically clean and contains no, or very minimal, evidence of the history of the situation or setting.

Tactics of Scientific Research (Sidman, 1960) in the experimental analysis of behavior was an attempt to provide purification and cleansing methods for the study of behavior via the classical science approach. It accepted that the raw materials of the science of behavior were dirty and contaminated, and that doing classical science with such reagents in real-life, or wild, situations was likely to fail. The introduction of the so-called Skinner box (and of many automated testing situations in other research areas, such as activity chambers, the Wisconsin General Test Apparatus, standardized pencil-and-paper tests, etc.) was a rational and quite successful attempt to make cleaner the situation in which we could study behavior. They eliminated the handling of animals and people that could affect results. Some eliminated distracting (thus, uncontrolled) visual and auditory stimuli, and the behavior of conspecifics. If, on the other hand, such stimuli were part of the focus of the research, then these stimuli could be better controlled and manipulated in such environments.

The physical test environment also promoted the mechanical, and later electronic, control of stimuli and of the relations between behavior and its consequences, thus again making the reagents more pure. Contingencies controlled by another organism, such as the human experimenter, are clearly impure in the sense that their operation depends on a set of potentially very dirty and uncontrolled variables that control the behavior of the experimenter.

At one level, these developments in test apparatuses and experimental controllers provided a environmental system for the study of behavior that was either clean (unlikely) or at least had a controlled and consistent dirtiness. A controlled dirtiness is okay for classi-

cal science if, of course, the dirtiness does not preferentially react with some reagents and not with others, or with some levels of a reagent only. The dirtiness just produces error variance.

But one of our reagents—our subjects—were never pure in an historical sense. The genetic make-up (a long-term historical effect) and the more proximate life span history of our subjects differed. When these things were critical, we instituted techniques of using littermates and even identical twins. We also developed the procedure of trying to control proximate history by providing explicitly equalizing experiences or, in the study of historical effects themselves, of providing explicitly different histories between subjects or conditions (e.g., Wanchisen, Tatham, & Mooney, 1989; Weiner, 1964). We have even used techniques in which we have attempted to control complete life span experiences, such as bringing up animals under controlled conditions, as in the procedure known in ethology as "the deprivation experiment" (Hinde, 1970).

In summary, we are working with impure reagents in a dirty environment, both of which we have tried to cleanse in various ways. Using a classical science approach, we assume generally that any combination of independent variables that we care to investigate will produce a single, stable, replicable result, though it may be muddied by random error variance.

Replication: Psychology As a Type I Error

If it was not the case that our independent variables produced single, stable, replicable results, would we ever have seen this, either empirically or from the purview of our published science?

In the classical science approach of Tactics of Scientific Research, replication is the major criterion. In this, Sidman (1960) follows both other sciences and the approaches in other parts of psychology. A published successful replication of a result has a special place in the classical approach to science in validating (direct replication) and, often, extending (systematic replication) the applicable domain of an effect. A failure to replicate also has a special status, at least in theory. Such a failure, if and when published, should imply that the original finding was not "real" in some sense, and depended perhaps on special but unspecified conditions of the original experiment. However, as we all know, failures to replicate are difficult, even impossible, to publish. Johnston and Pennypacker (1993) are quite clear that the onus of a failure to replicate falls on those who have failed—it is up to them to discover the "reason" for their failure. This may not be at all easy to do.

In my own experience, work in my laboratory has failed to replicate accepted results twice. First, Lesle Charman (1983) in her PhD dissertation failed to replicate systematically an effect that had been both initially reported and subsequently directly replicated. The effect was the short-component effect in multiple variable-interval (VI) VI schedules (Shimp & Wheatley, 1971; Todorov, 1972). These researchers found that when multiple-schedule component durations were short (5 to 10 s), behavior allocation to the components became very much more sensitive to the rates of reinforcers in the components, and that subjects came to distribute their behavior between components with the sort of sensitivity shown in concurrent schedules. Charman commenced her experiment with one component short and the other long, and found typical long-component multiple-schedule performance. As a result, she made both components short, and still failed to find the reported effect. For the next couple of years, we desperately tried to discover the reason for this failure. In the end, she directly replicated the original experiments. The effect, it turned out, depended on keeping the component schedules constant and changing the

component durations, whereas we had, through most of our experiments, kept the component durations constant and changed the component schedules. We were able to publish this very extensive work (Charman & Davison, 1982), albeit with some difficulty and only marginal effect. The short-component effect depended on the sequential history of the organism in the experiment; if you like, thinking of the experimental space as a piece of terrain, entering the terrain from one direction leads to a different hill being climbed compared to entering the terrain from another direction.

My second example is our work on magnetic discrimination in the pigeon (Alsop, 1987). Bookman (1977) reported that pigeons could discriminate the presence from the absence of a magnetic field when they flew down a flyway towards their mates. Alsop and I reasoned that, with our expertise in signal-detection procedures, we should be able to obtain the same result in a yes-no signal-detection procedure in a standard experimental chamber. However, the initial replication failed, and all of the training and testing procedures that we instituted (which work well with discriminations in other modalities) also failed. I took these data with me on a visit to the United States and presented the results at various universities. Almost without fail, I found researchers at each institution who had tried to replicate the original result and had failed. None could get their work published. In the end, Alsop and I approached the journal *Animal Learning & Behavior* and appraised the editor of the situation; to his eternal credit, he decided that an issue of the journal should be prepared that documented all these failures.

Let us first follow the logic of replication in the statistical model. But as we do so, we should wonder how much of the argument also applies to approaches that try to avoid statistical analyses.

At any one time, a number of researchers are likely, because of the zeitgeist, to be conducting research on a similar subject. Let's say we have 20 such researchers. On the average, at the 5% level of significance, one may obtain a positive finding by chance, and will publish his or her results. The others, having no statistically significant result, cannot publish. (What is, indeed, similar research is actually an interesting and nontrivial question, and the answer to this depends on and convolves with systematic replication! There is a sense in which "similarity" in research cannot be judged until after the domain, and related domains, have been thoroughly researched.)

Next, 20 researchers read the initial result and decide a replication is called for. Nineteen obtain nonsignificant results, and one obtains a significant result. The 19 cannot publish their research, and one publishes a successful replication. The effect is now discovered and replicated and may even start to appear in the introductory texts within 10 years.

In psychology, how many of our "effects" are Type I errors? Without knowing the number of failures, it is hard to say. Certainly, some are. Without carrying out a signal-detection analysis of all successes and failures, we cannot be sure whether an effect really exists. If we are investigating a difference in behavioral results between two levels of an independent variable, as predicted by a theory, we may have 20 papers that showed an effect with Level 1 of the independent variable (hits, in signal-detection parlance) and 20 more that showed no effect with Level 2 (correct rejections). If the unpublished misses and false alarms totaled only two cases, we can be pretty sure of the effect and maybe even the theory. If, however, they totaled 200, we ought to be pretty damn sure that the effect is not there, and the theory is wrong. But, under the current practices of science, we never know the number of misses and false alarms.

Perhaps we could argue that the experimental analysis of behavior, which eschews the statistical model, is rela-

tively immune to such errors? No, it is not immune. For a start, it is hard to publish results in which some subjects did something quite different from others (an implicit statistical criterion). But even without the statistical model, there is still the inability of researchers to publish (i.e., of editors to accept) null or negative results without an incredible power in the design, a power usually much greater than was available in the original experiment.

I think it could well be argued that, in our science, a failure to replicate has an equal status, perhaps even a greater status, than the original finding. The failure to replicate provides information that is missing in the original finding. It says, clearly, either that (a) the original finding was a chance finding; or (b) there are subtle effects of either history or of current variables that are active here; or, perhaps, (c) either the original researcher or the replicator used poor techniques (though this is not really discriminable from the second point). But the meaning of a null or negative result in the classical science approach is not necessarily the same as the meaning of the same result in dynamical science.

The behavioral situation is a swamp. There are too many things entering and leaving and interacting for it to be easily understood. Technically, behavior "suffers" from ("enjoys" might be a better term!) multiple causation. Even, I suggest, our most refined situations (say psychophysics and laboratory research in the experimental analysis of behavior) are sufficiently swamp-like to be difficult to comprehend. Other areas (social psychology, applied behavior analysis) are orders of magnitude more complex. We are not a classical science, nor can we ever be, and we must face that challenge. We have been led astray, I believe, by our finding some replicable and quantitative relations between behavior and environment, such as the matching law and its successors. For instance, it seems likely that measures of choice have prospered simply because the measures they use are ratios of responses and time as measures, and as a result much variation is canceled out. Absolute approaches to behavior have fared much less well (e.g., Herrnstein, 1970). There may be very few other simple, classical relations like the matching law.

Multiple causation can be handled by classical science with no real problem, save the problems engendered by the complexity of the systems, as long as the systems we are investigating have static single end points. But we are not dealing with such systems.

More Problems: Dynamic Transactions

It was Skinner (1950) who pointed out that the interaction of the organism and its environment was dynamic. An animal's behavior changes the environment, and the change in the environment changes the animal's behavior, and so on. This is illustrated nicely in the Columbia *Jester* cartoon mentioned by Skinner (1959). In this, one rat in a Skinner box tells another, "Boy, have I got this guy conditioned! Every time I press the bar down he drops in a piece of food." (There was a subsequent cartoon in which the second rat replied "Yeah, but how do we know it applies to rats?") Baum (1973) developed this general idea, and introduced the notion of E rules (environment rules, or how the environment or the experiment works) and O rules (organism rules, or how the organism works). An experiment is a set of E rules designed to elucidate one or more O rules. This notion is relatively simple when taken at the level of a single environmental controlling variable that affects a single behavior, even though the application of a single level of an independent variable may affect behavior in such a way as subsequently to change the level of the independent variable from its initially applied level to another level. For instance, the reinforcer rates or durations that the experimenter arranges are unlikely to be

those that a subject actually receives. As long as there is a negative feedback relation in the system, the dependent variable and independent variable can come to what is known as a detailed (or semi) stability—a particular response rate and a particular reinforcer rate that serve as coattractors. If we measure both the dependent and independent variables when the system has stabilized, we can quantitatively relate the results of this obtained level of the independent variable to other measured independent-variable levels resulting from other experimenter-applied independent-variable levels and gain some indication of the function or equation relating dependent-variable level to independent-variable level. Without the negative feedback, by the way, all we may get is the amplification of the dependent and independent variables to some minimal or maximal end point under each of the independent-variable manipulations.

This approach seems relatively straightforward, apart from the need to relate dependent variables to *measured* independent-variable levels, rather than to the applied, experimenter-controlled or initiated independent-variable levels. This is generally not required in classical science because the independent-variable level is not often affected dynamically by the dependent-variable level that results from the initial independent-variable level. In classical sciences, the feedback of dependent variable to independent variable is broken. In chemistry, I can weigh out a known amount of silver nitrate, and add it to a known amount of sodium chloride, and predict the result. But if I arrange a VI 30-s schedule for bar pressing, the rat will always get fewer than two reinforcers a minute in a session, and indeed (if it fails to respond) may get none. I cannot predict what the obtained reinforcer rate will be from the arranged rate, because this also depends on the response rate.

However, a problem arises because the stability that is obtained between response and reinforcer rates on VI schedules (for example) is a *detailed stability,* which means that the behavior–environment transaction is stable at only one of many possible stability points. The attractors in the present situation may be weak, and there may be a number of such weak attractors. Anger (1956) introduced this notion to the experimental analysis of behavior in relation to performance on VI schedules. He suggested that, at stability, the frequency distribution of interresponse times matched the frequency distribution of reinforced interresponse times, but that this equalization could occur for many "semistable" pairs of distributions. Thus, one could obtain a number of functions relating response rates to reinforcer rates, including no effect of reinforcer rate on response rate (*locked rate*; Ferster & Skinner, 1957). We now know from subsequent research (Catania & Reynolds, 1968; Herrnstein, 1970) that such functions are probably not as variable as might have been expected from Anger's analysis. However, significant variability does remain. It may be that the location at which stability is achieved is dependent on the path taken—the history of behavioral contingencies—prior to the exposure to the condition, as has been shown quite consistently in research on behavioral history (see review by Wanchisen, 1990).

Indeed, it has been shown quite clearly that performance on concurrent VI VI schedules can be manipulated, apparently permanently, by historical exposures to contingencies of reinforcement. Davison, Sheldon, and Lobb (1980) trained pigeons on equal concurrent VI VI schedules in which, in occasional discriminated parts of sessions, an extra contingency was added to one of the alternatives. This contingency naturally changed relative performance during the discriminated period. However, when the contingency was removed (i.e., the reinforcer was delivered noncontingently), the performance differential was maintained over a large number of sessions. The interesting result here is that the

performance differential gave the subject no more reinforcers. In a sense, it was maintained as a *superstition*, and indeed both Type I and Type II superstitions (see Herrnstein, 1966) are examples of detailed, semistable relations between responding and reinforcement.

Dynamical transactions between behavior and environment do not necessarily lead to single stable end points in the relation between behavior and environment. There may be many end points, the end points may be unstable, and they may consist of a constantly changing system.

Feedback Functions and Semistability

The ability of a behavior–environment system to stabilize at more than one pair of dependent- and independent-variable values can occur for either of two reasons. First, if there are multiple attractors, such as two or more ways in which the subject can gain reinforcers, different subjects may behave differently, probably as a result of differing personal histories. Even if the two alternative ways of gaining reinforcers are not equal, different historical trajectories (technically, "world lines") may result in different subjects climbing different peaks and stabilizing at different points if there is a chasm between the peaks.

An example here is the research of Vaughan (1981). Using concurrent schedules, Vaughan provided two areas of performance in which a subject's relative time allocation could match (equal) its received relative reinforcer frequency. However, between these two matching areas, there was an area in which performance, according to melioration theory, would be directionally pushed. Vaughan reversed this directional push in a second experimental condition. All but 1 subject moved from one matching area to the other. The 1 that did not move was, presumably, not adequately accessing the directional-effect area, and it was given therapy to help it move into this area.

Second, even if there is no chasm between the different sides from which a hill can be climbed, an organism may stabilize at a suboptimal level if its horizon in either time or distance is very limited, or if whatever mechanism is available for hill climbing is sensitive only to large absolute gradients or gradient changes. The organism may be said to satisfice (Simon, 1957; Staddon, 1980) rather than to optimize overall.

An example here is the report by Azrin and Hake (1969) on positive conditioned suppression. Azrin and Hake investigated the effects on VI performance in rats of a brief signal that was followed by the delivery of a reinforcer overlaid on a VI baseline (a positive conditioned-suppression procedure). They found that 15 of 18 subjects showed a deceleration of responding during the signal, and 3 showed accelerated responding. For reasons not entirely clear to me, the authors decided to use "therapy" for 2 rats that showed conditioned acceleration to bring their performance into line with the larger number of subjects. It is more likely that there existed a relatively flat gradient between these two nominal performances, and that the "therapy" simply provided a history that led the recalcitrant subjects across this flat to nominal suppression. It is thus also likely that had the 15 subjects that showed suppression been provided with other "therapy," their performances could have been changed to acceleration. In either case, a bald statement that the positive conditioned suppression procedure *produces* deceleration (or, for that matter, acceleration) seems untenable.

The peaks and chasms, and the gradients between them, in the terrain describing behavior–environment relations are described by feedback functions (Baum, 1973, 1992; Nevin & Baum, 1980) as well as by the O rules. A feedback function is a description of how the environment works, such as how high is a reinforcer rate, or how large are reinforcers, or how delayed are reinforcers, and so forth, as a func-

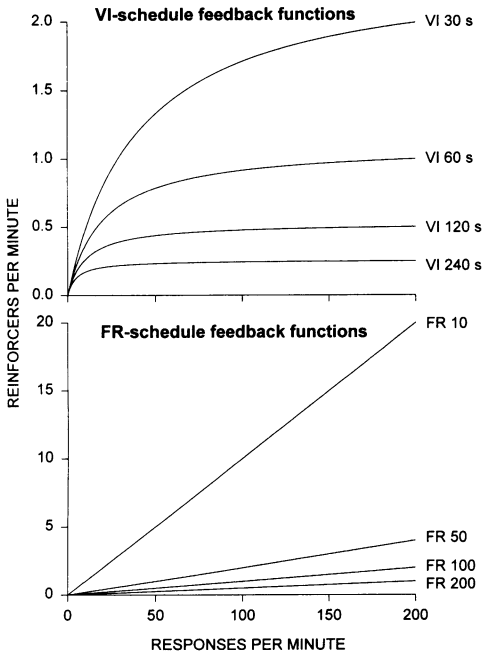

Figure 1. Feedback functions for various FR schedules (upper panel) and VI schedules (lower panel) showing reinforcer rates obtained as a function of response rates emitted.

tion of some aspect of the performance of the subject. They are, in Baum's terms, E rules. Feedback functions can sometimes be calculated for schedules of reinforcement. For instance, the feedback function of reinforcer rate for response rate for a fixed-ratio (FR) schedule requiring N responses per reinforcer is simply $R = B/N$, where R is the reinforcer rate obtained and B is the overall response rate emitted. This equation is graphed in the lower panel of Figure 1. All this graph and equation show is that obtained reinforcer rate is proportional to the subject's response rate (and the proportionality is the ratio requirement). The faster the subject responds, the higher its obtained reinforcer rate. Exactly the same feedback-function equation is true for variable-ratio (VR) N. However, as we well know, performance on FR schedules is quite different from performance on VR schedules, in terms of both overall response rates and the temporal distribution of responding, so there must be other controlling variables affecting detailed performance. Feedback functions could be written for these other controlling variables and aspects of performance also. Remember always, though, that feedback functions have no implication about the control of behavior by an independent variable—even a steep feedback function for a variable that is *irrelevant* to behavior does not imply control by that environmental independent variable. All that a feedback function says is that behavior will affect that aspect of the environment. An appropriate example here is Davison and Kerr's (1989) experiment in which they asked whether concurrent-schedule performance was controlled by the overall rate of obtained reinforcers. They did this by arranging that the degree of choice on concurrent VI VI schedules changed the overall rate of reinforcement available. They arranged a series of conditions in which more extreme preference increased reinforcer rates and another series of conditions in which more extreme preference decreased reinforcer rates. Their results showed absolutely no control over preference by absolute reinforcer rate. The feedback function existed, both theoretically and in fact, but it did not control behavior.

Reinforcer-rate feedback functions for schedules other than FR schedules are more difficult to calculate. An example is those for VI schedules (Baum, 1973, 1992; Nevin & Baum, 1980). Detailed aspects of the subject's performance, such as the pattern of bursts and breaks in responding and the interresponse time distribution generally, will affect the output of reinforcers by the environment. An approximation to a VI feedback function is shown in the upper panel of Figure 1. The shape is an increasing hyperbola, in which response- rate changes at low response rates affect reinforcer rates substantially (because the schedule will act more like an FR 1 schedule at these points). But when response rates are high, changes in response rates have hardly

any effect on reinforcer rates. The feedback function at higher rates of reinforcement is flat.

Flat areas of feedback functions for variables that *do* control behavior are areas in which differential control is weak. Let us assume that overall reinforcer rate does control response rate on VI schedules. For the VI feedback functions in the lower panel of Figure 1 at high response rates, changes in high response rates produce vanishingly small changes in reinforcer rates. With a wide, flat attractor of this sort, the location of the behavior of a subject within this area would be weakly specified, and subjects' performances may differ in their locations in such areas because of inherent subject differences (e.g., the weight of a pigeon's head could partly determine the response rate via interpeck times). The differing impure histories of the different subjects may also determine location on the feedback function. For instance, response rate on VI 60 s might be lower after exposure to VI 180 s than after exposure to VI 10 s. Finally, within subjects, different local histories (say, previous exposure to differential-reinforcement-of-low-rate [DRL] or FR schedules) may well determine location. It is no accident that the study of historical schedule effects on current behavior (Wanchisen, 1990; Weiner, 1964) has used, as its assessment device, FI schedule performances that have much the same feedback function as VI schedules, with a long flat portion to the feedback function. What is being used here as an assessment device for historical effects is what Davison et al. (1980) termed *weak contingencies*—contingencies of reinforcement that do not strongly push behavior in any direction. For their history conditions, such studies use strong contingencies such as FR or DRL schedules, in which the reinforcer rate (the imputed controlling variable) is strongly and steeply related to response rate.

Overall, the point being made here is that if a controlling variable has a flat, or relatively flat, feedback-function relation with the behavior it controls, then the point at which the dependent and independent variables come to semistability may be quite unpredictable on the basis of knowledge of the *current* value of the controlling variable. Rather, the point at which the system achieves semistability may be dependent on historical factors such as the direction (e.g., from a low response rate or from a high response rate) from which the new values were approached. Such flat feedback systems are, as Davison et al. (1980) argued, very appropriate for assessing historical influences and are quite inappropriate for assessing proximate influences. An example they gave was behavior on personality tests. Given that there are no differential reinforcers (apart from those arising from the test instructions, or demand characteristics of the situation) for differential behavior, instruments such as personality tests are likely to be good at measuring historical effects, but they may be poor, still, in helping to predict behavior under strong proximate conditions in which history is overwhelmed by current contingencies. It seems doubtful, for example, that the results of personality tests would be helpful in predicting behavior-analytic interventions in which, usually, strong proximate controlling variables are applied. On the other hand, everyday life seems to consist largely of relatively weak proximate controlling variables with rather flat feedback functions, and in such conditions historical variables probably have a rather large effect on current behavior.

The example, par excellence, of the flat feedback-function situation is the superstition experiment, in which there are no contingencies at all between the presence or absence of a response and the delivery of reinforcers, nor between rate of any response and rate of reinforcers (apart from the need to be in an appropriate spatial position to receive the reinforcer) (see Herrnstein, 1966, for a review). Historical contingencies come to the fore in this situation. If there is no explicit prior training, the behaviors that develop in pi-

geons are the sorts of behaviors that genetics (i.e., long-term history) define (Skinner, 1948; Staddon & Simmelhag, 1971; Wagner & Morris, 1987). If there is a clear set of historical contingencies (e.g., Herrnstein's, 1966, prior training to peck a key), then this history will be evident under adventitious reinforcement. In what is probably a more transient effect, and thus probably not a flat feedback-function situation, the relative amount of training for pecking versus eating freely available food determines the preference between these alternative responses in choice tests (Mitchell & White, 1977; White & Mitchell, 1977).

History and Flat Feedback Functions

Flat feedback-function situations, then, can generate performance that is inexplicable in terms of proximate causes and current independent variables, but is explicable in terms of historical causes. The problem, of course, is that we often do not know, and cannot guess at, the historically inaccessible previous training (though, as above, various tests may be available to help to summarize history). Further, we simply cannot scientifically allow guessing at the relevant history, because such "explanations" are too cheap and easy to provide. Moreover, for rather obvious reasons related to the drive to publish and maintain scientific output, long-term experiments that manipulate history have been generally eschewed by scientists. The result of these processes has had an interesting, and I might argue disastrous, result. In behavior analysis, we have been led to focus our basic research and, more important, our technology on very strong proximate contingencies: Major reinforcers that are highly differential (i.e., have a steep feedback function) are consistently used. There is nothing subtle at all about our technology. Now, this is fine in terms of the strength of our technology and the ease with which we can produce an effect and replicate it: The strength of the contingencies we use does not allow, most often, a failure of replication. We can account for 95% of the variance in our data. However, we may have overdetermined the system.

I guess we could argue that this is fine for the technology of our science. We need to be able to cause strong effects, and large changes in behavior, if we are to have this technology accepted. On the other hand, we could argue that such techniques were rather akin to the use of chemical or physical constraints on people who are sick, and that because the contingencies are so much stronger than those existing in everyday life, they do not fit our clients particularly well for a return to everyday life. We might also argue, on the contrary, that sustained exposure to such strong contingencies would constitute some serious historical contingencies for the future life of the client and thus be particularly helpful in the face of subsequent weak contingencies. Any conclusion here depends on questions about the degree of stimulus control by therapeutic interventions versus historical contingencies and the differential reinforcement that has been provided in each. It is well known that intervention within the problem setting is much more effective than intervention in discriminably different settings (see, e.g., Cooper, Heron, & Heward, 1987, Part 10).

It is interesting to ask whether the effects of an historical contingency are ever *eliminated* by current contingencies, or whether they are just made relatively small in *current* effect by the current contingencies themselves. This is a question of stimulus control. Presumably, in stimulus terms, if the current situation is identical to the previous situation, we are likely to see behavior similar to the historical behavior initially, and it may be maintained in the longer term if current contingencies are weak. (I am assuming no intervening strong-contingency training in the identical or a similar environment.) In this case, historical contingencies are not affected, remain active, and will

control present behavior. However, if present contingencies are strong and differ from the historical contingencies, and the present situation or setting is similar or identical to that in which the training originally occurred, history will be rewritten to an extent that depends on the duration and strength of the new contingencies. History thus may be rewritten under some circumstances, and not under others. Of course, if current contingencies are the same as historical contingencies in the same stimulus situation, history can be reinforced. The suggestion here is that history is rewritten (or reinforced) only if both (a) strong-contingency training in (b) similar stimulus conditions subsequently occurs. Otherwise, it remains intact. In this way, even very distant historical contingencies can have major present effects in identical current situations with weak and nondirective or nondifferential contingencies.

In behavioral research, I guess that by the time research commences, we have hoped to have overwritten historical differences to a large extent, although we all know, having seen continuing problems in animal and human training that extend into the experiment proper, that we have not, in fact, rewritten history. Maybe we have done it sufficiently? Well, that will depend, as argued above, on the strength of the contingencies applied. If they are weak, we may never manage to eliminate historical effects (as in Skinner, 1950). If current contingencies are strong (as in the "therapy" part of the conditioned suppression experiment by Azrin & Hake, 1969, and by Vaughan, 1981), we may succeed. However, we cannot approach these questions in a categorical way, and there is no absolute criterion of success or failure in our attempts to rewrite history. We may succeed in relation to some conditions of the experiment and fail in relation to others, potentially confounding results considerably. In Herrnstein's (1966) experiment, he rewrote the dominant response in the situation and made it "pecking the key." But superstition is subject to drift, and had he continued his fixed-time 11-s conditions longer, it is likely that pecking would have ceased, and the pigeon-like behaviors noted by Skinner (1948) would have been reinstated as historically (genetically) the stronger influence.

In summary, two points have been made: First, in the presence of weak proximate contingencies, present behavior can be strongly influenced by behavioral history; and second, to avoid this, the experimental analysis of behavior tends to use, especially in its technology, strong proximate contingencies. Thus under some conditions, principally those providing weak differential reinforcement, current behavior can be strongly influenced by historical training. This effect can also lead to failures of replication, and such results may be difficult to discriminate from failures caused by dynamical behavior–environment interactions.

The Quantitative Removal of Variability and the Concept of State

Sidman (1960) spent some time discussing how variability in behavioral baselines could and should be minimized so that the effects of relevant independent variables could be clearly and unambiguously seen. I believe this now requires further consideration.

Variability in baselines can arise from both dynamical interactions causing chaotic behavior and by the operation of uncontrolled and, by definition, *effective* independent variables. Some of the latter may be inherent to the organism in some sense (e.g., diurnal rhythms, age changes, etc.), and they may be only marginally controllable by the external reengineering of the experimental situation. Some may be genuine external environmental controlling variables, such as handling and deprivation levels, and so on. These perhaps may be controllable. Given that these uncontrolled variables *do* affect behavior, they are obviously pertinent

to the control of the behavior in question.

Indeed, one might argue this way: If your baseline variability is so large as to hide the effect you are investigating, you should never reengineer the situation to eliminate the variability. Such a result simply says to the researcher that other independent variables in the environment have relatively more massive effects on the behavior in question than does the manipulation; hence, the effective control of the behavior *requires* the experimenter to investigate these other sources of control. If we reengineer the situation to decrease the baseline variability, we are simply blinkering ourselves to these other potent controlling variables and focusing on an independent variable that is both minor and footling. You might argue that following up the disruptive effects of the jackhammer on the adjoining building might be equally footling. However, the fact that extraneous stimuli do disrupt some performances and not others may be an important finding.

Let us now have a more technical look at variability from a quantitative standpoint. The reference experiment here was published by Hunter and Davison (1985) (see also Schofield & Davison, 1997), but the analysis we shall follow here is a simplified analysis. The essence of Hunter and Davison's experiment was to arrange concurrent schedules that were either concurrent VI 60 s VI 240 s or concurrent VI 240 s VI 60 s. Which of these was presented in a session varied pseudorandomly, with nothing, except the schedules themselves, to indicate which was in effect. Naturally, this procedure produces very considerable session-to-session variability in measures of choice between the alternatives, and if you did not know the procedure, you would say that the data were a terrible mess. They look like those that we would get with inadequate control of extraneous controlling variables.

We will take this as an exercise in identifying and removing variability, but we will do this mathematically rather than experimentally.

Theory—the generalized matching law—says that stable-state choice between concurrent-schedule alternatives is controlled by the relative frequency of reinforcers obtained at the two alternatives. The generalized matching law is written:

$$\log\frac{B_1}{B_2} = a \log\frac{R_1}{R_2} + \log c.$$

Assuming that this theory is correct, we can take out the effects of reinforcers and be left with pure error variance. To do this, we linearly regress the log response ratio in a single session against the log reinforcer ratio in that session (Figure 2). We obtain quite a poor fit (31% variance accounted for: Figure 2), and find that after taking out the systematic variance there remains a considerable amount of "error" variance, the residuals of the fit. Is this really uncontrolled error variance? Not necessarily. Hunter and I wondered whether any of this residual "error" variance could be accounted for by what had happened the session before today, so we then regressed all of today's *residual* performances (i.e., after taking out the predicted effect of today's reinforcers) against yesterday's obtained log reinforcer ratios. As Figure 2 shows, we again account for some (64%) of what had appeared to be error variance but is now shown to be systematic variance. In Figures 2 and 3, we continue this process, progressively using the log reinforcer ratio inputs in sessions more distant from the current session, and progressively accounting for less of the remaining error variance.

As we progressively fit the linear model between today's data and the reinforcer inputs from previous days, several effects occur: First, generally, the value of a (the sensitivity parameter, the slope between the log reinforcer ratio and log response ratio for a particular previous session) decreases. Second, the variance accounted for first

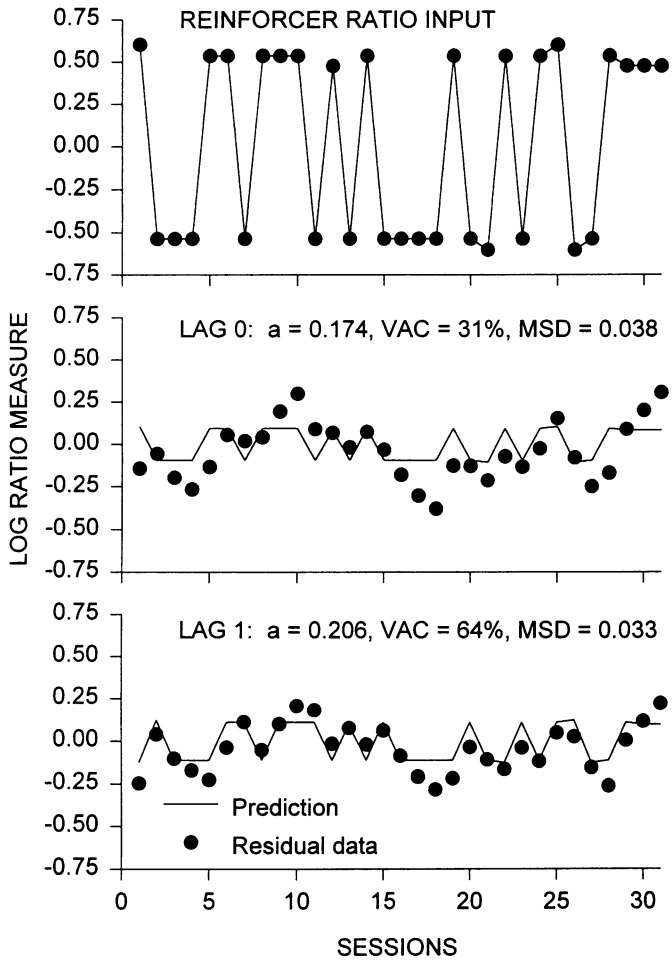

Figure 2. The original log response ratio and reinforcer ratio data for Bird 24 from Hunter and Davison (1985) (top panel). In subsequent panels are shown the results of regressions between today's log reinforcer ratio and today's log response ratio, and between yesterday's log reinforcer ratio and today's log response ratio.

rises and then falls; indeed, of the residual data, more variance is accounted for by yesterday's and the day-before-yesterday's reinforcer inputs than by today's inputs. Third, the overall variation in the data (residuals) falls. And fourth, the mean-square deviation between the predictions and the data falls to very low levels. In this way, because the generalized matching law is a reasonable model of performance in choice situations (Baum, 1974; Davison & McCarthy, 1988), we can extract almost all of the variation in the data by looking at historically distant environmental inputs.

The same process of extracting the environmental causes of current behavior is at least theoretically possible in any area of behavioral research, and can be carried out with independent variables on differing dimensions. However, a number of factors militate against the use of this procedure. First there is the need to insure that all influences occur randomly with respect to time (including the major influence being investigated). In applied research, it may not be acceptable to provide treatment and control conditions randomly with respect to sessions, but multiple baseline designs come close

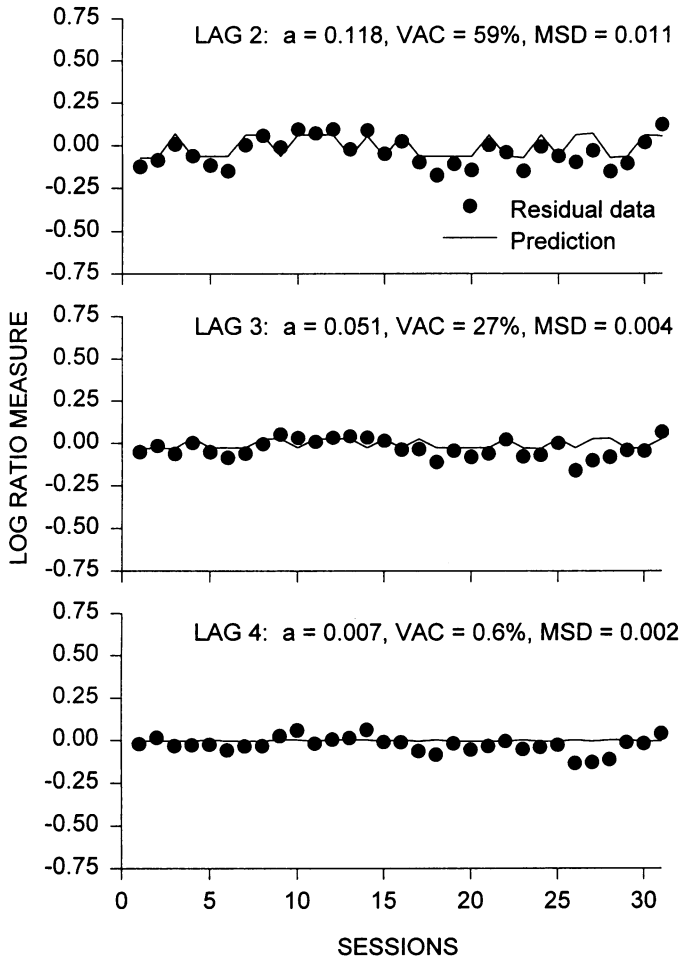

Figure 3. A continuation of Figure 2, showing the results of regressing today's log response ratio against the log reinforcer ratios obtained in more and more temporally distant sessions.

to doing this. Also, these influences must not be in phase with each other, otherwise "aliasing" will occur, in which it will appear that either one of two independent variables can each describe the data. Second, an adequate (perhaps not perfect) model of performance as influenced by these variables has to be available. We, above, were able to use a linear model of the sort $Y = mX$. If we had the wrong model, we could actually create variance in an initial regression that would then, naturally, have to be accounted for by further regressions! Third, if we are dealing with a set of different independent variables acting concurrently, we would need to be able to assess the current (and possibly past) status of each of these each time data are taken.

There is, however, an interesting way out of the necessity of looking at (and knowing, in a sense) the values of past conditions. Again taking the above example, it is evident that performance in a particular session is the product of the log reinforcer ratios in that session and in about three or four previous sessions. Thus, a particular session's *performance* contains within it historical information about previous independent-variable effects, about what happened in this and previous sessions. Thus, the reinforcer history

that modulates the next session's reinforcer input is embodied in the previous session's performance (see Davison & Hunter, 1979). We can thus effectively predict current-session performance from just the current session's reinforcer inputs *and the previous session's performance*. In equation form,

$$\log\frac{B_{1(0)}}{B_{2(0)}} = a\log\frac{R_{1(0)}}{R_{2(0)}} + b\log\frac{B_{1(-1)}}{B_{2(-1)}},$$

where the subscripts 1 and 2 refer to the response alternatives, and the numbers in parentheses refer to the sessional lag (0 being the current session). B is the number of responses and R is the number of reinforcers. The parameter a is sensitivity to the current-session reinforcer ratio, and b is the sensitivity to the previous session's *behavior* ratio. The larger that b is relative to a, the greater the contribution of history to this session's performance. Such an equation described Davison and Hunter's (1979) data well. In this case, of course, yesterday's behavior is not what we would call an independent variable; it is, after all, the animal's behavior. However, because of its summarizing properties, in many ways it does act as an independent variable for performance on the next day, and we can indeed control its value, at least in an historical sense. Yesterday's behavior summarizes the subject's reinforcement history. In a sense, yesterday's behavior is equivalent to some sort of psychological test.

This is what Staddon (1993a, 1993b, 1993c) was talking about when he reintroduced the concept of *state*. As he says, the implied terms of physiological state or mental state have various problems, not the least of them being the shock and disbelief engendered in the behavioral community. However, if we were to use the term *behavioral state*, we may be able to circumvent most of the unhappiness. Staddon avoids this rather trite (but useful) avoidance response, and says (1993c, p. 248) that "The behaviorist answer is state as equivalent history." State is a "class of histories that are sufficient to produce the measured differences in response" (1993c, p. 248). Thus, in order to understand and predict behavior, we can have recourse to two differing sources of data: First, we could know history, or sufficient history, to make the explanation; second, we could know the state of the organism, which is a summary of its history. Behaviorists are oddly blinkered about state explanations. On the one hand, they eschew them, and on the other, they will accept them if they are mechanistic enough. One example that readily springs to mind is Nevin, Mandell, and Atak's (1983) measurement of the strength of response via a challenge to behavior of particular types. Strength of response is clearly a state variable and was postulated as such prior to the considerable amount of work that Nevin and others have done to discover what historical variables affect this state.

In summary, then, previous behavior can be used to summarize previous reinforcer conditions, and to construct a state variable. In many areas of psychology (but generally *not* in the experimental analysis of behavior), this seems to be dimly recognized and, for example, pencil-and-paper tests of intelligence and personality and suchlike are used in this way as an aid to predicting future performance. Baseline variability is, in a sense, an engineering problem, but it is probably more valuable to the progress of a dynamical science of behavior to see it not as the random demon that wrecks experiments but as nature whispering, talking, or screaming to us about other controlling variables.

Thus, random behavior may not always be the result of chaotic behavior–environment interactions. Rather, it may be systematic behavior that is under the control of independent variables that are occurring randomly with respect to the experimental manipulations the experimenter is undertaking. If it is, then it can be analyzed and extracted.

More on the Scientific Value of Variance

It is completely accepted in science that systematic variance away from a known result or a theory must lead to further research. The real question is: What is systematic variance?

In the Hunter and Davison (1985) result, the remaining variance after the "main effect" (today's input of reinforcers) is removed does not, in any way, appear systematic. It looks, even to the trained eye of the researcher, like a jumble of uncontrolled influences. However, this apparently random variance was shown to have systematic components when analyzed according to a particular type of model (we will not say, "the correct model"), one in which previous inputs were investigated as potential independent variables. Thus, it is not just *systematic* variance away from a model (although that is most important in guiding research), it is also *apparently random* variance that contains information. In a sense, here, we are taking the strong behaviorist classical science Laplacian position that it is possible, in principle, to explain all behavior completely (although it is an impossible job to do so in actuality). And we are saying: The more variance, of any sort, in the data, the more interesting those data potentially are.

The prime example, here, is the now-common demonstration of chaos. A completely deterministic set of equations with particular starting points can rapidly and predictably (in the sense of knowing this will happen) produce a totally unpredictable and random sequence of outputs (in the sense of not knowing which output will occur when). The smallest change in the starting points, down to vanishingly small decimal numbers, will produce quite different output sequences after a while, and may even produce fundamentally different stable data. The data are random. The way in which the data are produced—the equations—have no random elements and are often very simple and straightforward equations. Are the data really unpredictable, then? No, in the sense that if we *could* start the equations at *exactly* the same point, the same results would ensue.

Ruelle (1993) has an excellent example concerning billiard balls, one real cue ball and one imaginary. We compare the trajectory of the real and the imaginary balls which are hit at a *very* slightly different angle. As these two balls hit the convex sides of other balls on the tables, their trajectories are very soon completely different from each other. It is the convexity of the balls that is important here: A slightly different point of contact between two balls changes the angle of incidence very considerably, and the difference between the trajectories grows exponentially in time. The effect of very slightly different starting angles is called *sensitivity to initial conditions*. Ruelle points out that if the distance between the real and imaginary balls doubles every 1 s, and starts at 1 micron, then after 10 s the distance will be 1,024 microns (1.024 mm), and after 30 s the distance would be greater than 1 kilometer. It is evident that over a relatively short transaction of this sort, "behavior" can vary immensely from one start of the system to the next.

The behavior–environment transaction is just like this, though potentially much more complicated. Not only is the cue ball hit at a slightly different angle on each strike (start of the experiment), but also the other balls on the table are at subtly different positions—maybe not greatly different for a laboratory experiment, but maybe very greatly different if part of the environment consists of moving rather than static balls (to translate, for instance, other behaving animals). But even in the laboratory, the placing of the balls will never be accurate enough to produce the same long-term result twice. The motion of balls on a billiard table is completely deterministic (or so I am assured), but predicting the future of the system has inherent limitations,

and prediction will always get exponentially worse with time since the start of the system. In behavior, we have E rules and O rules that transact dynamically over time. Although these may be relatively simple (though I doubt it), the sheer number of them (multiple causation for both the environment and behavior) means that it is highly likely that we shall always have complex dynamical systems that, even if they do not lead to chaotic behavior, will lead to fundamental unpredictability. However, if we know the E rules and the O rules, we can at least simulate the processes and discover predicted attractors, areas in which the transaction can stabilize. Clearly, with such systems, replication in the usual sense may be unlikely or even impossible.

In terms of behavior, it is quite evident that in processes like this, one could quite quickly find oneself in a seriously different environment from that predicted, and his or her behavior will then be under the control of completely different contingencies, including different organisms. Predictability of behavior may only, at best, be accurate in the very short term. In physics, the two-body problem (predicting the interaction of two celestial bodies, for instance) can be solved. The three-body (or more) problems cannot. Approaching behavior in the same terms, I suspect even the two-body problem will be beyond us.

Not all dynamical systems (systems that have a time evolution) are unpredictable. Some, like wine in a wine glass, always return to the same arrangement as long as the wine remains in the glass. Turn the wine glass upside down, however, and the predictability of the wine's behavior is lost.

In another sense, though, the behavior of a billiard ball *now* does contain within it at least a partial representation of its history, at least in the short term. The fact that it is here *now*, and that it arrived here from this angle at this speed after x s since the blow that started the system, is informative. The fact that the results are different on a replication is also informative. Moreover, were we to measure the behaviors of the other balls on the table, then it might be possible, I believe, to recreate most of the trajectory or world line of the cue ball. There is a fascinating implication that comes from these considerations: If we were to wait until *stability,* when all the balls have ceased moving, it surely would be impossible to recreate much of the world line: There would be only location data, and a time from the initial blow, and no directional or speed information. There would be myriad ways of getting to this stable point. This may indicate that the study of transitions will provide more and more useful information than can the study of stable end points to behavioral manipulations. But bear in mind one inescapable fact: Animal subjects bring to the billiard table their own personal histories, which may have large effects on the behavior of the cue and the other balls. And unlike billiard balls, animals are nonspherical, unequally weighted, and of varying hardness!

Our ability to describe apparently random variation using real, controllable, potentially *independent* variables is limited not only by our creativity and our technology, but also by our theoretical position on what needs to be done with and about "random" variance. If it is cast as having seriously negative value, as a major threat to our main effects, and we follow the appropriate technology, we will amplify our main effects and squash all other effects. If we cast it as potentially positive, our main effects may account for less variance, but the whole experiment will be potentially much richer. The experimental analysis of behavior has always been considerably better at recognizing the value of variance, as evidenced by Skinner (1959). The question is whether the experimental analysis of behavior has yet shown enough recognition of the value of variance.

In concert with the casting of random variance as demon, the experi-

mental analysis of behavior has always been very clear on not publishing unwarranted speculation after the experimental event. It is true, and I accept the view, that in the past some such speculations, usually about the purported mechanisms of the main effects, have gotten into the literature as fact and have had an unfortunate influence on subsequent work. But along with this, "unwarranted" speculation about sources of error variance and about failures to replicate either within or across experiments has also been struck out. I question this. Is not the experimenter's opinion (he or she is, of course, closest to the data) a rich source of ideas and knowledge? Could these not better guide subsequent research? Are we doing the best for science if we eliminate such admittedly informal opinion? Such opinions have no scientific status, but they certainly have a status in the psychology and sociology of science. If this were not so, why would researchers gain so much from attending conferences and discussing wild ideas informally with others late into the night?

Given multiple causation, the complexity of the behavior–environment system, and the dynamical nature of the behavior–environment transaction, both the longer term prediction of behavior and experimental replication may well be either difficult or impossible.

Single-Subject Experimental Design

The effectiveness of a single-subject, or subjects-as-their-own-controls, experimental design in terms of convincing scientists of a real effect of an independent variable (or, for that matter, of the noneffect of a variable) is completely predicated on one or more assumptions: (a) There is no effect of historical manipulations on present behavior. (b) Minor fluctuations in the levels of putatively extraneous independent variables will have no catastrophic effects on the behavior of some or all subjects. (c) The baseline measurement will not, itself, either sensitize or desensitize the system to particular levels of independent variables or to the introduction of an independent variable. (d) Alternatively, the strength of the contingencies applied will overcome the effect of the above.

These assumptions apply to any experimental design in which the subject is exposed successively to different experimental and baseline conditions (ABAB designs, multiple baseline designs, changing criterion designs, etc.). Oddly enough, in discovering major effects, designs that use control groups may be less affected by historical differences because the histories will be more diverse. However, I would not argue for the use of control groups when the interesting information (at least according to my view) is in discovering the historical contingencies that modulate the effects of current independent variables, which often seem to have an import equal to the effects of the currently applied variables.

Can we make the above assumptions? I argue that they are dangerous to make, as are all assumptions. In the applied literature, we all know of experiments in which baselines could not be recovered, often for clear reasons. Indeed, an effective therapeutic intervention (as distinct from a scientific experiment) rests more on the possibility of trapping a behavior change, and therefore leaving the therapist in a position to cease intervention in the knowledge that the effected behavior change will be maintained, than on the demonstration of a recovered baseline. In order to control splay feet, we can continue to reinforce the maintenance of an appropriate angle between the feet and punish an inappropriate angle, without making any long-term change if the natural proximate and historical contingencies do not support the behavior change. We need to understand such "extraneous" contingencies (other E rules in the environment) if we are to be able to provide an adequate technology, and we also need to understand

them if we are to provide an adequate science.

Conclusion

Error variance in behavior is simply the control of behavior by variables we do not understand or have not identified, the interaction of behavior and environmental systems that we do not understand or have not identified, or both. In the former case, by better experimental control, or by statistics, we can attempt to highlight the main effect. In the latter case, we will have to accept multiple causation and massive dynamical transactions between behavior and environment. We will have to look for the equations that describe the components of the transaction, and try to recreate the major features of the interaction; the appropriate data here may be data variance and periodicities in the data. Whichever science we do, it seems to me that more sensitivity towards variance is required. One aspect of this sensitivity is the need to know the number of failures of replication as well as the number of successful replications, which requires some changes to the sociology of our science.

If there is a dynamical transaction between behavior and its environment, we must also confront the very real possibility that systems may be able to stabilize at more than a single point, and that replication, both direct and systematic, may be considerably less important than we had previously thought. Indeed, a demonstration that a system can stabilize at a number of different points (failing to replicate, negative results) is most important for our understanding of behavior. We should cherish such findings rather than push them under the carpet.

Finally, we must understand that behavioral history can have large current effects under certain circumstances, principally those in which current stimulus situations are similar to previous situations, and in which differential reinforcement is weak: flat feedback-function situations. Indeed, flat feedback functions provide data that not only may not be predictable from current contingencies but also are more predictable from recent training. But flat feedback functions are the method of choice for assessing behavioral history. The notion of behavioral state may help us to understand current behavior and predict future behavior in these situations and many others.

I fully agree with Sidman (1960, p. 142), who wrote,

> Most psychologists accept the premise that the subject matter itself is intrinsically variable over and above experimental error. As a direct consequence of this presupposition, confidence-level statistics have been substituted for replication as a means of evaluating data. . . . Because the doctrine of natural behavioral variability appeared to be sound, until recently the data upon which most current systematic interest is centered have been produced by experimenters operating within this doctrine. . . . Meanwhile, the premature acceptance of intrinsic variability as a basic property of behavior has led to the adoption of experimental designs whose nature effectively prevents further investigation of the problem.

I would reinterpret his last point, however, which is absolutely true from the viewpoint of classical science. From the dynamical science viewpoint, we would accept that variability, in some sense intrinsic, will be a basic property of some behavior–environment systems. The answer is not statistics, but rather it is the understanding of the interaction of the environmental and behavioral systems. Engineering out the variability may lead us into the same problem as statistics; in both, the system is simplified sufficiently for classical science. We must be very careful in doing this.

As Sidman says, "Variability may be measured, and even used as a datum" (1960, p. 142). I would venture that variability *should* be measured, and should be used as a datum, because it contains within it important clues to the properties of the system that we study. We need to explain and predict variability. Finally, Sidman states "In order to treat any instance of variability as a manifestation of an or-

derly process, we must not only identify the source of the variability but also control it" (1960, p. 143). Yes, indeed, we should certainly identify it, and then derive the equations that produce the variability. But we should always remember that controlling an "extraneous" source of variance may seriously and fundamentally change the operation of the system in which we are interested, and may lead us towards an incomplete and rather sterile understanding of behavior.

REFERENCES

Alsop, B. (1987). A failure to obtain magnetic discrimination in the pigeon. *Animal Learning & Behavior, 15,* 110–114.

Anger, D. (1956). The dependence of inter-response times upon the relative reinforcement of different inter-response times. *Journal of Experimental Psychology, 52,* 145–161.

Azrin, N. H., & Hake, D. F. (1969). Positive conditioned suppression: Conditioned suppression using positive reinforcers as the unconditioned stimuli. *Journal of the Experimental Analysis of Behavior, 12,* 167–173.

Baum, W. M. (1973). The correlation-based law of effect. *Journal of the Experimental Analysis of Behavior, 20,* 137–153.

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

Baum, W. M. (1992). In search of the feedback function for variable-interval schedules. *Journal of the Experimental Analysis of Behavior, 57,* 365–375.

Bookman, M. (1977). Sensitivity of the homing pigeon to an earth-strength magnetic field. *Nature, 267,* 340–342.

Catania, A. C., & Reynolds, G. S. (1968). A quantitative analysis of the responding maintained by interval schedules of reinforcement. *Journal of the Experimental Analysis of Behavior, 11,* 327–383.

Charman, L. F. (1983). *Performance in multiple schedules.* Unpublished doctoral dissertation, University of Auckland, New Zealand.

Charman, L., & Davison, M. (1982). On the effects of component durations and component reinforcement rates in multiple schedules. *Journal of the Experimental Analysis of Behavior, 37,* 417–439.

Cooper, J. O., Heron, T. E., & Heward, W. L. (1987). *Applied behavior analysis.* Columbus, OH: Merrill.

Davison, M. C., & Hunter, I. W. (1979). Concurrent schedules: Undermatching and control by previous experimental conditions. *Journal of the Experimental Analysis of Behavior, 32,* 233–244.

Davison, M., & Kerr, A. (1989). Sensitivity of time allocation to an overall reinforcer rate feedback function in concurrent interval schedules. *Journal of the Experimental Analysis of Behavior, 51,* 215–231.

Davison, M., & McCarthy, D. (1988). *The matching law: A research review.* Hillsdale, NJ: Erlbaum.

Davison, M., Sheldon, L., & Lobb, B. (1980). Positive conditioned suppression: Transfer of performance between contingent and non-contingent reinforcement situations. *Journal of the Experimental Analysis of Behavior, 33,* 51–57.

Ferster, C. B., & Skinner, B. F. (1957). *Schedules of reinforcement.* New York: Appleton-Century-Crofts.

Herrnstein, R. J. (1966). Superstition: A corollary to the principles of operant conditioning. In W. K. Honig (Ed.), *Operant behavior: Areas of research and application* (pp. 33–51). New York: Appleton-Century-Crofts.

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

Hinde, R. A. (1970). *Animal behaviour: A synthesis of ethology and comparative psychology.* New York: McGraw-Hill.

Hunter, I., & Davison, M. (1985). Determination of a behavioral transfer function: White-noise analysis of session-to-session response-ratio dynamics on concurrent VI VI schedules. *Journal of the Experimental Analysis of Behavior, 43,* 43–59.

Johnston, J. M., & Pennypacker, H. S. (1993). *Strategies and tactics of behavioral research.* Hillsdale, NJ: Erlbaum.

Mitchell, P., & White, K. G. (1977). Responding in the presence of free food: Differential exposure to the reinforcement source. *Bulletin of the Psychonomic Society, 10,* 121–124.

Nevin, J. A., & Baum, W. M. (1980). Feedback functions for variable-interval reinforcement. *Journal of the Experimental Analysis of Behavior, 34,* 207–217.

Nevin, J. A., Mandell, C., & Atak, J. R. (1983). The analysis of behavioral momentum. *Journal of the Experimental Analysis of Behavior, 39,* 49–59.

Ruelle, D. (1993). *Chance and chaos.* Harmondsworth, UK: Penguin Books.

Schofield, G., & Davison, M. (1997). Non-stable concurrent choice in pigeons. *Journal of the Experimental Analysis of Behavior, 68,* 219–232.

Shimp, C. P., & Wheatley, K. L. (1971). Matching to relative reinforcement frequency in multiple schedules with a short component duration. *Journal of the Experimental Analysis of Behavior, 15,* 205–210.

Sidman, M. (1960). *Tactics of scientific research.* New York: Basic Books.

Simon, H. A. (1957). *Models of man.* New York: Wiley.

Skinner, B. F. (1948). "Superstition" in the pi-

geon. *Journal of Experimental Psychology, 38,* 168–172.

Skinner, B. F. (1950). Are theories of learning necessary? *Psychological Review, 57,* 193–216.

Skinner, B. F. (1959). A case history in scientific method. In S. Koch (Ed.), *Psychology: A study of a science: Vol. 2. General systematic formulations, learning, and special processes* (pp. 359–379). New York: McGraw-Hill.

Staddon, J. E. R. (1980). Optimality analyses of operant behavior and their relation to optimal foraging. In J. E. R. Staddon (Ed.), *Limits to action* (pp. 101–141). New York: Academic Press.

Staddon, J. E. R (1993a). The conventional wisdom of behavior analysis. *Journal of the Experimental Analysis of Behavior, 60,* 439–447.

Staddon, J. E. R. (1993b). The conventional wisdom of behavior analysis: Response to comments. *Journal of the Experimental Analysis of Behavior, 60,* 489–494.

Staddon, J. E. R. (1993c). Pepper with a pinch of psalt. *The Behavior Analyst, 16,* 245–250.

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

Todorov, J. C. (1972). Component duration and relative response rates in multiple schedules. *Journal of the Experimental Analysis of Behavior,* 17, 45–49.

Vaughan, W., Jr. (1981). Melioration, matching, and maximization. *Journal of the Experimental Analysis of Behavior, 36,* 141–149.

Wagner, G. A., & Morris, E. K. (1987). "Superstitious " behavior in children. *Psychological Record, 37,* 471–488.

Wanchisen, B. A. (1990). Forgetting the lessons of history. *The Behavior Analyst, 13,* 31–37.

Wanchisen, B. A., Tatham, T. A., & Mooney, S. E. (1989). Variable-ratio conditioning history produces high- and low-rate fixed-interval performance in rats. *Journal of the Experimental Analysis of Behavior, 52,* 167–179.

Weiner, H. (1964). Conditioning history and human fixed-interval performance. *Journal of the Experimental Analysis of Behavior, 7,* 383–385.

White, K. G., & Mitchell, P. (1977). Preference for response contingent versus free reinforcement. *Bulletin of the Psychonomic Society, 10,* 125–127.