Experimental analysis of behavior, Part 1
Iversen and Lattal (eds.)

CHAPTER 4

Experimental design in the analysis of free-operant behavior

MICHAEL PERONE

Department of Psychology, West Virginia University, P.O. Box 6040, Morgantown, WV 26506-6040, U.S.A.

1. Introduction

In an account of his seminal work with Skinner on schedules of reinforcement, Ferster (1953) defined *free-operant procedures* as those allowing a subject to emit the behavior of interest at any time during a test session, unconstrained by apparatus or experimenter. The typical arrangement has an animal in a small chamber with continuous access to the response operandum – a lever for rats or a key for pigeons. Because lever presses and key pecks occupy minimal time, rate of responding can vary over a wide range, thus providing a sensitive baseline for assessing experimental manipulations. With automated equipment such as cumulative response recorders and on-line computers, rate can be monitored continuously and recorded in real time, allowing researchers to examine moment-to-moment changes in freely emitted, ongoing behavior.

Free-operant procedures, and the empirical and theoretical issues they have been used to study, are familiar features on the contemporary scene in experimental psychology. However, in the 1930s when Skinner developed his research program, analysis of free-operant behavior represented a radical departure from the methods in common use (Kimble, 1961, Chapter 2; Skinner, 1938, 1979). At that time *discrete-trial procedures* dominated psychology. These procedures, still in use, restrict responding to isolated observation periods, either by removing the subject from the apparatus during intertrial intervals, or, if the subject is allowed to remain, by removing or disabling the operandum. In a typical experiment using a maze, for example, a rat is placed in the start box, allowed to find its way to a goal box baited with food, and then removed to a holding cage until the next trial. Discrete-trial procedures also can be arranged using operant chambers. With rats, trials may be defined by mechan-

ically inserting and withdrawing the response lever. With pigeons, the chamber and pecking key are illuminated during trials and darkened during intertrial intervals (although the key is physically present throughout, this procedure is functionally equivalent to the rat procedure because pigeons normally do not peck in the dark). The natural continuity of behavior is disrupted by treating each response (e.g., traversing the maze or pressing the lever) as an isolated event. When only one such event can occur per trial, behavior cannot be measured in terms of response rate, and instead reliance is placed on such variables as the percentage of trials on which the response occurs, the latency to initiate the response, and the speed of execution.

1.1. Comparison of free-operant and discrete-trial procedures

As in the foregoing summary, the conventional distinction between free-operant and discrete-trial procedures has emphasized: (a) continuous versus discontinuous observation of behavior, and (b) rate versus percentage, latency, and speed measures of response strength (e.g., see reviews or textbooks by Bitterman, 1966; Kling, 1971; Mackintosh, 1974; Fantino and Logan, 1979; Flaherty, 1985; Yaremko et al., 1986). The importance of such differences is controversial. Skinner (1969), arguing in favor of free-operant procedures, claimed that response rate is a 'basic datum in a science of behavior' (p. 110), and he objected strenuously to procedures that measure behavior in other dimensions. His disdain for discrete-trial procedures is evident in the following advice he offered to young psychologists:

> Do not spend much time (reading) articles in which changes in behavior are followed from trial to trial or in which graphs show changes in the time or number of errors required to reach a criterion, or in amount remembered, or in percent of correct choices made, or which report scores, raw or standard... Dimensions (of measurement) are probably suspect if the work was done with mazes, T-mazes, jumping stands or memory drums. (Skinner, 1969, pp. 93–94)

By comparison, authors whose theoretical views on learning and behavior fall outside the Skinnerian camp have been less dogmatic about the procedures to be used (e.g., Logan and Ferraro, 1970; Mackintosh, 1974). Mackintosh, for example, offered this rejoinder to Skinner:

> Much nonsense has been written on the relative merits of latency, speed, or rate measures, and hence on the supposed superiority of one procedure over the other... Different procedures and different measures do not always yield similar results from apparently similar operations... We do not always understand the reasons for these differences... but the existence of such differences must tell us a great deal about the limits of the conditions under which certain effects operate and may even tell us something about the underlying processes involved. Furthermore, the analysis of instrumental behavior would be the poorer, and its conclusions less valid, if its data came exclusively from a single, standardized experimental situation: we should never know whether a particular phenomenon was simply a consequence of some peculiarity of the experimental situation if we had no information about the generality across situations. (Mackintosh, 1974, p. 144)

1.2. Methodological implications

Regardless of the position one takes on the relative merits of free-operant and discrete-trial procedures, it is clear that they have been associated with distinct research traditions within experimental psychology. The critical difference between the traditions is not the specific procedures used, but rather the conceptions of behavior that have given rise to the procedures. In the operant tradition, behavior is understood as a continuous process of interaction between an organism and its environment, to be pursued as an object of study in its own right (Skinner, 1938, 1950, 1966; Sidman, 1960; Johnston and Pennypacker, 1986). Such a view entails a commitment to seek order at the level of the individual subject by experimentally isolating sources of environmental influence. Irregular data are assumed to reflect the operation of extraneous factors, not variability intrinsic to the behavior itself, and a critical test of the current state of the science is the ability to eliminate irregularity through improvements in experimental control. Single-subject designs, involving repeated measurement of an individual's behavior under several experimental conditions, constitute the preferred method of research. Indeed, processes that are not amenable to study with single-subject designs may be seen as falling outside the boundaries of behavioral science (e.g., Sidman, 1960, pp. 52–53).

In the alternative view, behavior is studied not so much because it is interesting in its own right, but because it provides a measure of underlying processes and structures that cannot be observed directly – the modern counterpart to 19th century interest in behavior as the 'ambassador of the mind' (Wasserman, 1984). The objects of study have varied over the years, ranging from habit strength and other forms of stimulus-response associations (e.g., Hull, 1943), to cognitive processes such as attention and memory (e.g., Terrace, 1984; Rilling and Neiworth, 1986; Roitblat and Weisman, 1986), to the organization of mental structures (e.g., Luce, 1986). Interruption of behavior into discrete trials is not seen as doing violence to the operation of these underlying variables. Order is not sought in the behavior of the individual because its relation to the process of interest is only indirect; to cancel the noise inherent in behavioral data and uncover the order in the underlying process, the performances of several subjects are grouped together and statistical comparisons are made across averages of groups given different experimental treatments.

1.3. Overview

This chapter deals with the design of laboratory research in the tradition that evolved from Skinner's work on free-operant behavior, a tradition that has become known by such labels as 'operant conditioning' and, more suitably, the Experimental Analysis of Behavior. As a practical matter, an area of investigation can be defined by pointing to current work in the area (Dinsmoor, 1966, p. 421); accordingly, this chapter will favor description over prescription by characterizing the Experimental Analysis of Behavior as it is reflected in reports of contemporary research.

1.3.1. Single-subject experiments

Research on free-operant behavior relies on *single-subject designs* that compare experimental conditions imposed on individual organisms. The adjective 'single' is somewhat misleading, as the research rarely involves just one subject. At the least, procedures must be replicated with several subjects to establish the reliability and generality of the relations under study. Perhaps it would be best to remember that the term 'single' describes the unit of analysis – the behavior of the individual – not the size of the sample.

Single-subject designs include 3 characteristics: (a) repeated observation of behavior; (b) manipulation of one or more independent variables on a within-subject basis; and (c) demonstrations of stability within and across levels of the independent variable.

1.3.1.1. Repeated observation. Kazdin (1982) has described repeated observation as the 'fundamental requirement' of single-subject designs (p. 104). The ideal is to measure behavior continuously so that the effect of an independent variable can be assessed over time. Accordingly, researchers have developed automated procedures to allow continuous monitoring of increasingly subtle details of performance, and it has become conventional to conduct experimental sessions on a daily basis for months or even years. Thus, although few subjects are involved in single-subject research, each is studied extensively and yields substantial amounts of data.

1.3.1.2. Within-subject manipulation. Any psychological experiment must compare behavior in one condition with an estimate of behavior had the condition not been introduced. In single-subject research, the repeated observations conducted under control conditions provide the basis of comparison, or baseline, for assessing the effects of experimental operations imposed on the same organism. Sidman (1960) put it this way:

> Ongoing behavior gives the experimenter an important tactical advantage: he can manipulate it directly. He can introduce a new variable, or change the value of one that is already relevant, and he can observe any alterations that take place in the subject's ongoing behavior. (p. 409)

Comparisons may be conducted successively, as when each condition is imposed for a block of sessions, or simultaneously, by arranging two or more conditions within every session.

1.3.1.3. Stability. Only if differences observed across experimental conditions are reliable can the differences be confidently attributed to manipulation of the independent variable. Reliability is said to be present when behavior is stable from one observation to the next under constant conditions (e.g., Nunnally, 1978). In single-subject research, the repeated observations made in every condition allow assessment of 2 types of stability: (a) stability of behavior from observation to observation within a condition; and (b) stability of the change in behavior from one condition to another. Both types must be present before valid inferences can be made about causal relations between the experimental operations and behavior.

Sidman (1960) described stable behavior as representing a *steady state*. This term, borrowed from the physical sciences, portrays behavioral stability as a state of equilibrium in the reciprocal interaction between behavior and the variables that influence it. Because the goal of experimental analysis is to identify and control such variables, stability is the cornerstone for evaluating success: stability reflects adequate control over relevant variables, with minimal variation in the data from session to session within a condition as well as across replications of a condition.

2. Stability criteria

Since Sidman's (1960) classic book on methodology, single-subject research has been closely associated with the analysis of steady states. Indeed, by contemporary standards the production of steady states has become an essential feature of experimental method in the analysis of free-operant behavior. Each condition is continued until a steady state is observed, at which point another condition is imposed, and the effects of the experimental manipulations are evaluated by comparing the steady-state performances from the last few sessions of the various conditions. Behavior in transition from one steady state to the next is considered of secondary importance (if it is considered at all), as the nature of its control is obscure by comparison with that of steady-state performances.

To analyze steady states, researchers must overcome 3 obstacles. First, they must exert enough control over experimental and extraneous variables to engender a steady state. Second, they must impose such control long enough for the steady state to occur. Finally, they must recognize the steady state when it does occur.

Overcoming the first 2 obstacles, regarding the need for control over possibly long periods of time, depends on the state of the science. A successful experimental science is one that exerts high degrees of control over its subject matter and fosters the pursuit of research topics where the painstaking exercise of such control is deemed worthwhile. The ability to control variables that affect behavior is prerequisite to the analysis of steady states. Thus, because single-subject designs require researchers to seek strict levels of control, they encourage the development of an experimental science of behavior.

The third problem, recognizing steady states, arises frequently. Given the present understanding of behavior and its measurement, it would be unreasonable to expect the complete absence of variability in steady-state performances. Thus, one may ask how much variability can or should be tolerated. To answer this question, researchers have devised decision rules called *stability criteria*. These rules play a pivotal role in the Experimental Analysis of Behavior. By defining the steady states that form the basis of experimental comparisons, stability criteria have a substantial impact on the standards of control that researchers strive to meet and on the quality of the data that are available for analysis. If the criteria are too lenient, conditions may be ter-

minated prematurely and uncontrolled variability may obscure or distort any effects that may be present. At the other extreme, if the criteria are too strict they may never be met and, once again, effects may go undetected. Because stability criteria are so important, a detailed discussion is in order.

2.1. Trend and bounce

Although there are several types of stability criteria, they share concerns about 2 types of variability: trend and bounce. *Trend* refers to systematic increases or decreases, whereas *bounce* refers to apparently unsystematic variance or noise. Ideally, steady-state behavior should be free of both trend and bounce over some specified number of sessions, usually 5 to 10, that may be called the *terminal sessions.*

Elimination of trend appears to be a realistic objective in most laboratory experiments; one has simply to keep conducting sessions until there are no systematic increases or decreases in the value of the dependent variable. The absence of trend can be verified in several ways: (a) by visual inspection of the graphed data; (b) by ensuring the absence of monotonic upward or downward changes, i.e., by ensuring that the data values do not rise or fall consistently across the terminal sessions; or (c) by fitting a line to the data from the terminal sessions, with the goal being a slope of about zero. The first 2 methods are more common than the third.

The treatment of bounce also differs across stability criteria. Some criteria are expressed in quantitative terms, indicating an acceptable range of variation over the terminal sessions. Others are nonquantitative; one specifies the duration of experimental conditions on the assumption that performances over the last few sessions will represent a steady state, and another relies on visual inspection of graphed data. Representative examples are considered below.

2.2. Quantitative stability criteria

Stability criteria may quantify the limits of variation in relative or absolute terms. *Relative stability criteria,* which specify an acceptable range of variation in terms of percentage change across the terminal sessions of a condition, have been used since at least the mid-1950s. An early version used at Columbia University (Schoenfeld et al., 1956; Cumming and Schoenfeld, 1960) considered response rates from the most recent 6 sessions, with means calculated on the basis of the first 3 of these sessions, the last 3, and the entire block of 6. The difference between the sub-means was divided by the overall mean, and behavior was judged stable if the resulting percentage was less than 5. By this criterion, behavior was in a steady state if variation in mean response rates within the block of sessions was sufficiently small relative to the overall mean rate across the block.

Absolute stability criteria specify the acceptable range of variation in terms of fixed units of behavior rather than a percentage of the prevailing levels. For example, in

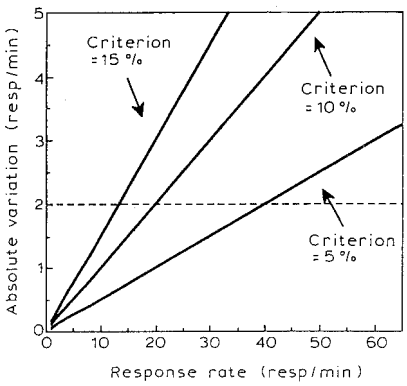
a study of concurrent schedules by Dinsmoor et al. (1981), the stability criterion considered response rates from the most recent 6 sessions, with medians calculated for blocks of 3 sessions each, and behavior was judged stable when these medians differed by no more than 2 resp/min. (The criterion had to be met by the rates on both of the concurrent schedules). Or, in a study of choice proportions by Spetch and Dunn (1987), the criterion divided the most recent 9 sessions into blocks of 3. Behavior was considered stable when the block means fell within a range of 0.10 and there was no monotonic upward (M1 < M2 < M3) or downward (M1 > M2 > M3) trend.

There is nothing sacrosanct about the number of terminal sessions considered by the criterion or the amount of variation deemed acceptable, and the experimental literature reveals a range of values for both. It should be noted that raising either factor will tend to relax the criterion. The reason is obvious in the case of raising the limit on variation, but some explanation is in order regarding the effect of raising the number of terminal sessions. This step reduces the influence of fluctuations that occur over the short term, usually from day to day, by enlarging the samples of behavior entering into the calculations. If large daily fluctuations are characteristic of a subject's performance, and cannot be traced to the influence of controllable extraneous variables, then they should be considered an aspect of the steady state. If a steady state has been attained, and if the samples of behavior are large enough so that the variability is equally distributed, then the long-term variation assessed by comparisons across large blocks of sessions should be minimal even while short-term (day-to-day) variation is substantial.

2.2.1. Adjusting for response rate

The stringency of any quantitative criterion, relative or absolute, is greatly influenced by the prevailing response rate. This is illustrated in Figs. 1 and 2. Fig. 1 shows how fixing the allowable variation in *relative* terms (i.e., at a percentage of the average response rate over the terminal sessions) permits a wide range of variability in *absolute* terms (i.e., in responses per unit of time). Specifically, the figure shows the absolute variation allowed by 3 relative criteria (5, 10 and 15%) as a function of the prevailing response rate. The absolute limit imposed by each criterion is directly proportional to the prevailing rate. Viewed in such absolute terms, then, relative criteria tolerate substantial variability in behavior occurring at high rates, but vanishingly small variability as rates decrease.

Fig. 2 illustrates how *absolute* stability criteria permit a wide range of variability in *relative* terms. The figure shows functions representing 2 criteria, an absolute limit of 2 resp/min as in the experiment of Dinsmoor et al. (1981) and, for comparison, a limit of 1 resp/min. In each case, the limit on relative variability is a negatively accelerated, decreasing function of the prevailing response rate. Thus, viewed in relative terms, absolute criteria tolerate substantial variability in low response rates, but little variability in high rates.



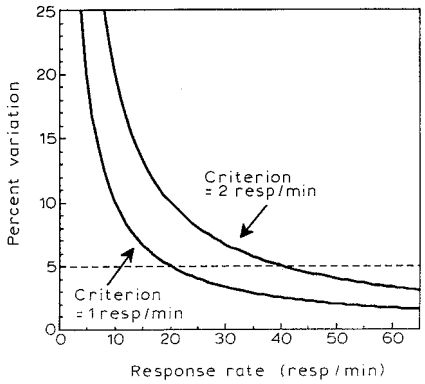


Fig. 1. (left) Variation in absolute response rate allowed by 3 relative stability criteria (5%, 10%, 15%) as a function of the prevailing response rate. For each criterion, the allowable variation is a direct function of the prevailing rate. The horizontal dashed line provides a basis for comparing the relative criteria with an absolute criterion of 2 resp/min; e.g., if the prevailing response rate is less than 40 resp/min, a 5% criterion will be more stringent (allow less absolute variation) than the 2 resp/min criterion.

Fig. 2. (right) Relative level of variation in response rate allowed by 2 absolute stability criteria (1 and 2 resp/min) as a function of the prevailing response rate. The allowable variation is a negatively accelerated, decreasing function of the prevailing rate. The horizontal dashed line provides a basis for comparing the absolute criteria with a relative criterion of 5%; e.g., if the prevailing response rate is less than 40 resp/min, a 2 resp/min criterion will be less stringent (allow more relative variation) than the 5% criterion.

When response rates are low, even minor disturbances from session to session may exceed the limits imposed by some relative criteria. Relaxing the criterion may be wise under such circumstances; indeed, it may be necessary to complete the experiment. An example is provided by the data in Fig. 3, which shows responding by rats on multiple schedules arranging brief periods of timeout from an avoidance schedule that was programmed on another lever (Perone and Galizio, 1987). The 6 conditions in the figure represent different combinations of variable-interval (VI) and extinction (EXT) components associated with the timeout response; for most of the conditions, only data from the last 10 sessions are shown. Across animals, rates averaged 2.0 to 7.4 resp/min in the VI component and 0.1 to 1.3 in the EXT component. Fig. 1 can be used to estimate the absolute limits to variability that would be imposed by a 5% criterion in this experiment: even with the highest prevailing rate of 7.4 resp/min, a discrepancy of just 0.4 resp/min would exceed the limit. For rats with the lowest rates, the limit would dwindle to 0.1 in the VI component and 0.005 in the EXT component. Meeting such narrow limits would have been impractical.

The solution adopted in this case involved the application of two criteria over 10 terminal sessions. First, responding in the VI component was judged according to a 15% criterion. As shown in Fig. 1, given the low response rates in this experiment, a 15% criterion is more stringent than the absolute limit of Dinsmoor et al. of 2 resp/min.

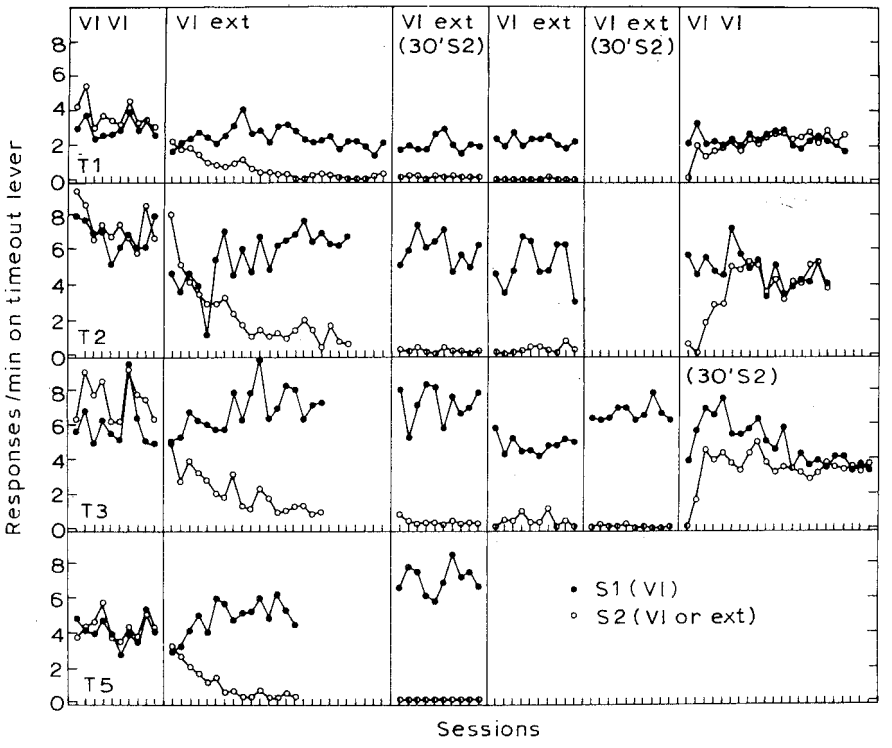


Fig. 3. Responding by each of 4 rats on multiple VI VI and multiple VI EXT schedules of timeout from avoidance. Unless otherwise noted, component durations were 6 min (exclusive of timeout periods). Data are presented from all sessions in the first multiple VI EXT and final multiple VI VI conditions; otherwise the data are from the last 10 sessions of the condition. The low prevailing rates make it difficult to meet relative stability criteria, especially in the EXT component, despite the consistency of the data from the terminal sessions of each condition. (Reproduced from Perone and Galizio, 1987, with permission of the publisher.)

Second, responding in the EXT component was judged solely by visual inspection (visual criteria will be discussed in the next section). Other researchers have acknowledged the need to abandon relative criteria when rates approach zero (e.g., Lattal et al., 1989). In the final analysis, the criteria apparently were successful, as there is little doubt about the stability of the terminal performances illustrated in Fig. 3.

Absolute criteria automatically adjust for decreases in prevailing response rates by relaxing the limit on relative variation. In the experiment of Dinsmoor et al. (1981), the mean rates during the terminal sessions varied from about 4 to 44 resp/min across animals and conditions, with most falling between 10 and 25. Fig. 2 shows that at 25 resp/min their absolute criterion of 2 resp/min is comparable to a relative criterion of 8%, but at 10 resp/min the relative equivalent rises to 20%. Of course, this property of absolute criteria does pose a risk of setting a limit that is too lax in conditions with low rates or too strict in conditions with high rates. Nevertheless, it is clear that

a judiciously selected absolute criterion could be serviceable over a wide range of rates.

The general issues raised by low response rates also pertain to high rates, but here relative criteria are more forgiving than absolute ones. Fig. 1 shows that at 60 resp/min, for example, a 5% criterion would tolerate variations of 3 resp/min, with further widening of the limit occurring in direct proportion to increases in rate (e.g., the criterion would tolerate variations of 4 at 80 resp/min, 5 at 100, and so on). By comparison, Fig. 2 shows that an absolute criterion of, say 2 resp/min would become more stringent than a 5% criterion when the prevailing response rates exceed 40 resp/min.

The foregoing discussion illustrates an important point about absolute and relative stability criteria: each controls one type of variability at the expense of giving up control over the other type. It might appear, then, that the choice between absolute and relative criteria should depend on the experimenter's views about the nature of behavioral variability. For example, if one believed that the allowable level of variability is directly proportional to response rate, a relative criterion would be used. But lacking a well-developed theory of variability, one might be better advised simply to consider the response rates that are likely to be generated by the experimental manipulations. As a practical matter, the consequences of the relations depicted in Figs. 1 and 2 may be expressed roughly as follows: as response rates decrease, relative criteria tend to become stricter, whereas absolute criteria tend to relax. Conversely, as rates increase, relative criteria relax and absolute criteria become stricter.

2.3. Other types of stability criteria

Also in common use are 2 additional types of criteria, identified by Sidman (1960) as 'fixed time-interval' criteria and 'criteria-by-inspection'.

2.3.1. Fixed time-intervals

Fixed time-interval criteria simply specify the overall duration of experimental conditions and the size of the sample of behavior to be considered as representing steady-state performance. For example, an experimenter might conduct each condition for 50 sessions, with the data from the last 5 providing the basis for analysis. Decisions about the exact parameters rest on judgments about the time course of the anticipated effects. Such judgments may be guided by published reports of studies with similar procedures and subjects, or by the experimenter's direct experience with the variables in question (e.g., in previous research).

Fixed time-interval criteria offer several advantages over the criteria considered so far. They make it easy to project the completion of an experiment, and thus to plan the allocation of resources in one's laboratory. They simplify the day-to-day operation of the laboratory by eliminating the need to calculate indices of stability after every session. And they reduce the number of decisions that must be made over the course of an experiment.

These attractive features are offset by 2 disadvantages. First, in setting the criterion, the experimenter must anticipate individual differences among subjects in the amount of time needed to reach stability. The criterion must be designed to accommodate the slowest subject. This will tend to prolong the experiment, as most subjects will stabilize ahead of the fixed time-interval, and laboratory resources will be tied up needlessly. Second, fixed time-interval criteria provide no direct evaluation of trend and bounce, raising the possibility that performances may not reach satisfactory levels of stability. The only way to address this matter is after the fact: when reporting the results, the experimenter may be able to demonstrate the adequacy of the criterion by presenting data or summary statistics (e.g., standard deviations) describing the levels of stability actually attained.

2.3.2 Visual inspection

The last type of stability criterion to be discussed here, *criterion-by-inspection*, involves reaching decisions about steady states by visual examination of the data. At one time the decisions were made on the basis of cumulative response records (e.g., Ferster and Skinner, 1957), but now it is more common to inspect graphs showing average response rates from each session in an ordered time series. Presumably, the experimenter considers both trend and bounce when examining these graphs, but how they are considered is not often described in published reports, most of which simply indicate that a visual criterion was used.

As with fixed time-interval criteria, the adequacy of criteria-by-inspection hinges on the quality of the experimenter's judgment, which, in turn, hinges on experience with the variables under study. In light of this limitation, Sidman (1960) offered 2 recommendations for researchers using visual criteria. First, such criteria should be restricted to the study of variables expected to have relatively large, simple effects, i.e., effects large enough to override the substantial noise that may be allowed by qualitative stability criteria, and effects simple enough to be specified as either present or absent. It would be risky to use criterion-by-inspection in parametric studies, where the objective is to provide a precise description of a variable's effect over a range of levels. Exceptions can be found, however. An influential parametric study of VI schedules by Catania and Reynolds (1968) used criterion-by-inspection. But these investigators reported that the response rates they judged stable by inspection also met, after the fact, a relatively stringent quantitative criterion, and this added to the credibility of their results. Sidman's other recommendation was that researchers report the data that guide the decisions about stability. For example, if the decisions are based on session-by-session graphs of response rate, then those graphs should be included in the report so that readers can make their own evaluations. Unfortunately, the scarcity of journal space often makes this impractical. In such cases, it is especially important for the experimenter to describe how the judgments were made and to report some empirical evidence of stability, e.g., in the form of standard deviations or other descriptive statistics.

2.4. Additional considerations

A few additional issues regarding stability need to be considered. These include (a) placing limits on the number of sessions in each experimental condition, (b) deciding which aspect of behavior should stabilize before changing conditions, and (c) recognizing the interplay between trend and bounce.

2.4.1. Limiting the number of sessions

Researchers may augment quantitative or visual stability criteria by setting limits on the number of sessions to be conducted in each experimental condition. At the start of a new condition, behavior may be slow to change, so that behavior in the initial stages of transition from one steady state to another may still meet the stability criterion. Clearly, it would be a mistake to terminate the condition at this point. To prevent such mistakes, it is common to specify some minimum number of sessions to be completed before applying the criterion. The exact number will depend on the nature of the anticipated effects, but the literature gives the impression that about 10 sessions will suffice in many cases.

Setting an upper limit on the length of conditions can be helpful in deciding when to give up trying to meet an elusive criterion. In practice, this amounts to substituting a fixed time-interval criterion for some quantitative or visual criterion, but only after repeated failures to meet the original criterion. The considerations in setting an upper session limit are the same as those in setting a fixed time-interval criterion.

When a condition is terminated because the upper limit has been reached, should the data be included in the analysis? To decide, the experimenter should consider the reasons for failing to meet the original stability criterion as well as the levels of stability actually attained. Perhaps the original criterion was missed because it was too stringent given the prevailing response rates or the nature of the variables under study – in the overall experiment or just the condition in question. Either way, it is probably safe to include the data, if they meet some other reasonable, albeit less stringent, criterion. Indeed, it may be wise to consider adopting the less stringent criterion for the subsequent conditions. But if the original criterion was missed because of failures in experimental control, the experimenter is best advised to make the necessary corrections and start over.

2.4.2. Which aspect of behavior should stabilize?

Many experiments collect data on several aspects of behavior at once. Experiments using multiple schedules measure response rates in 2 or more components, those using concurrent schedules measure rates on 2 or more operanda, and so on. Several measures may be taken even when a single schedule is studied. For example, studies of fixed-interval (FI) performance may measure postreinforcement pauses, running response rates, quarter-lives, and sequential patterns of interresponse times. In cases such as these, which aspects of behavior should be required to meet the stability cri-

terion? Examination of the literature suggests that various approaches are acceptable.

In experiments with parallel measures of a small number of separate responses, it is common to apply the stability criterion to all of them. For example, in studies of concurrent schedules, the stability criterion may be applied separately to rates on both operanda (e.g., Dinsmoor et al., 1981). Or to simplify matters, the criterion may be applied to a single composite variable such as a choice proportion (responses on one operandum divided by the combined responses; e.g., Spetch and Dunn, 1987).

The way is less clear when the several measures represent different aspects of a single performance. One might apply the stability criterion to every measure, continuing conditions until each one stabilizes. Although this approach may be justified on theoretical grounds, it seems impractical. If the criterion is stringent, it is likely that random fluctuations will prevent a large number of measures from meeting it simultaneously.

Perhaps the most common solution is to define the steady state in terms of the most global measure of behavior relevant to the aims of the research. Once stability is attained on this basis, the experimenter is free to examine more molecular measures in an effort to characterize what might be regarded as the underlying structure of the steady state. For example, interresponse time distributions might be examined after the overall response rate stabilizes (e.g., Arbuckle and Lattal, 1988), or distributions of pre-ratio pauses might be examined after the median pause stabilizes (e.g., Perone et al., 1987).

An alternative is to identify a relatively small number of variables central to the aims of the research, and to apply the stability criterion to them. A major advantage of this approach is that the analyses carrying the bulk of the theoretical or empirical load will be based on demonstrably stable data. For example, if the focus of a study on FI performance is on moment-to-moment response patterns, then the stability criterion should be applied to data that reflect the pattern, e.g., the index of curvature, quarter life, or postreinforcement pause (for a discussion of the adequacy of these measures and related ones, see Fry et al., 1960; Gollub, 1964; Dukich and Lee, 1973). Although this may seem obvious, there have been experiments in which the stability criterion was applied to data that were clearly of subsidiary interest, e.g., to overall response rates in studies where the nature of the response patterns was at issue.

2.4.3. Interplay between trend and bounce

Researchers seem less worried about trend than bounce. Published descriptions of stability criteria often fail to mention trend at all (e.g., Perone and Galizio, 1987; an interesting exception appears in a paper by Zeiler and Buchman, 1979, where the criterion is defined strictly in terms of trend). There are probably 2 reasons for the omission. First, many researchers would argue that requiring the data to be free of trend before changing conditions 'goes without saying'. From this perspective, there is no more point in describing the trend component of a stability criterion than in saying

the calculations were double-checked (it simply 'goes without saying'). Second, if the data meet a reasonably stringent criterion in terms of bounce, then it is likely that they also are free of significant trend. This is especially so if the criterion encompasses a relatively large number of sessions.

2.5. Replication in the evaluation of stability criteria

The absence of a standardized stability criterion may upset readers who seek to conduct research by a small set of specific formulas or who believe that scientific objectivity depends on wide consensus in such matters. Criteria may vary from one experiment to another, depending on the judgment of the researcher and the nature of the variables under study, but this state of affairs should not be misunderstood as anarchy. The adequacy of any criterion can be empirically tested within a properly designed experiment. According to Sidman (1960, pp. 258–259), a criterion is adequate if it allows researchers to select a reproducible state of behavior and thus leads to orderly and replicable functional relations between independent and dependent variables.

Even the most stringent criterion may be met by chance, i.e., by behavior that has not yet reached a true steady state. No change in the criterion itself can eliminate this possibility, but replication can reveal whether it has happened. If marked differences in outcomes are observed when experimental conditions are replicated, the stability criterion may be inadequate.

Replication, then, is the key to establishing the ability of a criterion to identify steady-state performances. Replication can be accomplished on a between-subject basis, e.g., by exposing 2 animals to one set of experimental conditions, or on a within-subject basis, e.g., by exposing 1 animal to the same conditions twice. Virtually every published steady-state experiment includes between-subject replication, and most include some form of within-subject replication as well. Exactly how replications should be arranged is an important aspect of experimental design, which is the topic to be considered next.

3. Experimental design

Much of the graduate training devoted to shaping sound research has given experimental design a bad name. The negative connotations come from long hours studying esoteric terms, rules, and statistical formulas. Overwhelmed by such matters, students hoping to complete a thesis or dissertation have created considerable demand for 'cookbook' approaches. But students who rely on pretested 'recipes' may lose sight of the central – and simple – objective of experimental design: to collect, analyse, and present data leading to valid conclusions about causal relations between variables. The study of experimental design is just an analysis of cases that will persuade critics that the conclusions are indeed valid.

3.1. Internal validity

If an experiment provides persuasive evidence that manipulation of the independent variable caused the changes observed in the dependent variable, the experiment is said to have *internal validity*. Campbell and Stanley (1963) and Cook and Campbell (1979) identified 8 classes of extraneous variables that, if not controlled, might produce effects that are confounded with the effects of the independent variable. Research designs can be judged by the degree to which they either eliminate these *threats to internal validity* or, failing that, equalize their effects across experimental conditions.

Several threats, i.e., history, maturation, testing, and instrumentation, represent collections of variables that operate as a function of time or repeated exposure to laboratory procedures. The experimenter should take account of such variables whenever the same subject is exposed to multiple conditions, which of course is always the case in single-subject designs. Relevant *historical variables* might include an animal's previous experience with drugs or schedules of reinforcement, as well as extra-experimental experience in its living quarters or in interaction with other animals. *Maturational variables* include not only long-term effects associated with biological aging but also short-term effects such as satiation or fatigue. *Testing* is a concern when the same procedure is administered repeatedly and thus may have effects of its own in addition to the ones it is supposed to assess. An example is the use of repeated extinction sessions to probe the strength of an operant. The *instrumentation* threat arises if there is reason to suspect that data may have been contaminated by drift in the calibration of measuring devices over the course of a study.

The prudent experimenter will make every reasonable attempt to eliminate these threats or at least hold them constant. For example, many historical variables can be controlled by physically isolating subjects from unwanted influences; indeed, that is the purpose of testing animals in sound- and light-proof chambers and housing them in individual cages under conditions of tightly-regulated temperature and humidity. But some extraneous variables cannot be eliminated, either because they arise in the experimental procedure itself (e.g., when the effects of one condition may carry over to the next) or because they are inextricably tied to the passage of time (e.g., aging). Valid experiments are designed so that the effects of such factors can be distinguished from those of the independent variable.

In single-subject designs, internal validity is established through replication. For example, in experiments that compare behavior at different points in time, conditions should be replicated on a within-subject basis. If the replications are successful in reestablishing previously observed behavior, then changes may be attributed to the experimental manipulation rather than the coincidental operation of some time-related extraneous factor. Between-subject replication also may enhance internal validity, especially when the timing or order of conditions is varied across the replications. In the sections that follow, both forms of replication are discussed in relation to illustrative designs.

Several additional threats to internal validity are peculiar to experiments involving group comparisons. These involve problems with nonequivalent groups, which can arise in either of two ways: (a) biased assignment of subjects to the experimental conditions (*selection* and *interactions between selection and maturation*) or (b) differential loss of subjects from the conditions over the course of the study (*mortality*). Such concerns are irrelevant in single-subject research because the same subject is exposed to every experimental condition.

The final threat to internal validity, *statistical regression*, is at issue when an experimental treatment is evaluated by comparing a single pretest with a single posttest, under circumstances in which the subjects were selected for study on the basis of extreme scores on the pretest. The extended sequence of observations required in single-subject research and the emphasis on stable data render this threat irrelevant as well.

3.1.1. Statistical versus experimental control

The dominating presence of group-statistical designs in the behavioral, biological and social sciences has led to the widespread misunderstanding that large numbers of subjects are needed for internal validity (for illuminating critiques of group-statistical approaches, see Meehl, 1967; Lykken, 1968; Bakan, 1970). But single-subject research is up to the task if it is properly designed and there is adequate control over the relevant variables. Results should meet the following criteria: (a) minimal variation should be seen from session to session within a condition; (b) clear changes should be evident when conditions are manipulated; and (c) the changes should be repeatable, i.e., replications of a condition should yield comparable measures of behavior.

Instead of improving control to demonstrate the reliability of effects experimentally, by way of replication, some researchers have sought statistical methods to help them detect effects that might otherwise be obscured by substantial levels of uncontrolled variability. A range of statistical tests have been devised for single-subject research (for a review, see Kazdin, 1984), with particular interest in time-series analysis (e.g., Jones et al., 1977; Gottman and Glass, 1978; Kratochwill, 1978). This approach seems reasonable because many single-subject experiments can be viewed as interrupted time series. In addition, because measures obtained from a single subject are not independent of one another, more familiar tests such as the analysis of variance appear to be inappropriate. At present, however, there is some question about whether serial dependence is a common characteristic of single-subject data (for a range of opinion, see Baer, 1988; Busk and Marascuilo, 1988; Huitema 1986a, 1988; Sharpley and Alavosius, 1988). If serial dependence is absent, it would be possible to apply analysis of variance after all.

Not surprisingly, the use of statistical inference has arisen mainly in applied research, where practical constraints may hamper efforts to achieve steady states and so difficulties may be encountered in identifying treatment effects without statistical assistance (e.g., DeProspero and Cohen, 1979; Knapp, 1983). Nevertheless, reliance

on statistical inference has been criticized as a distraction from developing more effective techniques of direct behavioral control (Michael, 1974; Baer, 1977; Parsonson and Baer, 1986), and many investigators of free-operant behavior would view them as wholly unacceptable in basic research.

Taking a pragmatic approach to statistical inference, Huitema (1986b) observed that "statistical tests have very high credibility to most people in most fields; to act otherwise is to commit political suicide" (p. 229). Thus, when publishing in journals that favor conventional group methods, some researchers may find it expedient to supplement single-subject data with group statistics. Even with just 3 or 4 subjects, the data can be organized in the traditional group format for repeated-measures analysis of variance, and the comparisons should meet the usual standards of statistical significance (for an example of such an analysis, see Galizio et al., 1986). This is not surprising, as group-statistical criteria of reliability are less stringent than criteria based on visually compelling demonstrations of experimental control at the level of the individual organism (cf. Parsonson and Baer, 1986).

3.2. Designs with successive comparisons

At the heart of every experiment is a comparison across sets of data collected under 2 or more conditions. In the most straightforward single-subject designs, the conditions are arranged successively (*successive-comparisons designs*). For example, Conditions A and B may be alternated across successive phases of the experiment, with a block of sessions devoted to each. Designs of this type are distinguished in terms of the number of alternations. The simplest, and least acceptable, version involves a single alternation: the A-B design. Minor extensions yield better designs, including the familiar A-B-A and A-B-A-B designs. These are superior to the A-B design because they arrange within-subject replications of one or more conditions and thus assess the reliability of any differences observed between A and B.

Textbook discussions often characterize the initial phase of single-subject designs as a baseline defined in terms of the natural frequency of the behavior under study (e.g., Barlow and Hersen, 1984; Christensen, 1985). Such characterizations may accurately depict applied research where real problems of human adjustment are under study. But in basic laboratory research, where most behavior is engendered by schedules of reinforcement and expressed through the movement of levers, keys and the like, the notion of natural frequency seems out of place. In laboratory experiments, a *baseline* is simply a point of departure, a standard of comparison for assessing the effects of the subsequent manipulations (cf. Sidman, 1960, p. 409).

Data from an experiment with an A-B-A-B design are presented in Fig. 4 (Morris, 1987). The experiment compared 2 ways of testing the effectiveness of a contingency to reinforce response variability (different sequences of 4 pecks on two keys). The figure shows that both subjects were more successful in meeting the contingency (as shown by the percentage of reinforcers obtained) under the 'discrete-response' condi-

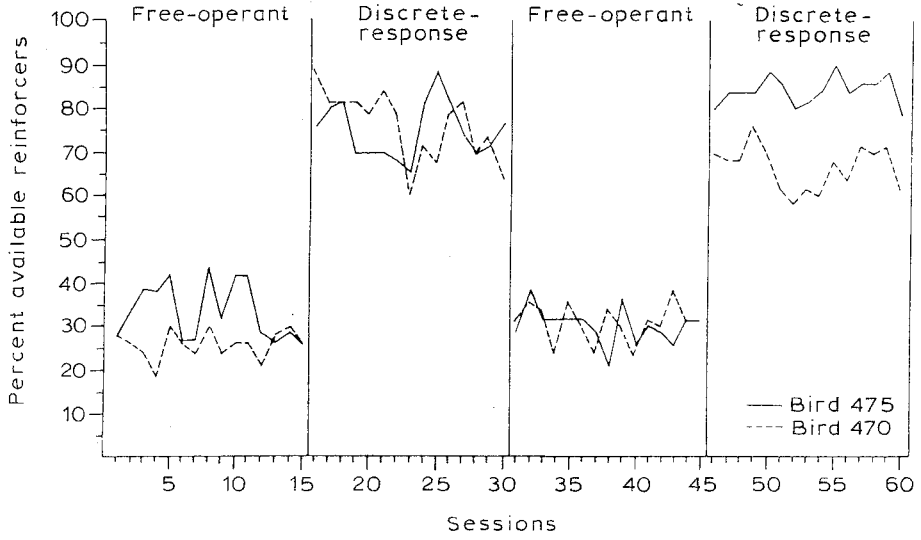


Fig. 4. *An A-B-A-B design.* The percentage of available reinforcers obtained by 2 pigeons under 2 procedures to encourage response variability. (Reproduced from Morris, 1987, with permission of the publisher.)

tion than under the 'free-operant' condition: the success rates rose and fell reliably as the 2 conditions were alternated across successive blocks of sessions.

Designs comparing successive conditions also may be distinguished in terms of the number of different conditions and the sequence in which they are presented. For example, 3 conditions are compared in the A-B-A-C-A design. Or the combined effects of two or more variables can be studied. The interaction between Conditions A and B, for example, can be evaluated by arranging this sequence: A-AB-A, where 'AB' designates the conditions in combination.

The study of multiple variables raises special issues. Textbooks (e.g., Barlow and Hersen, 1984) generally advise that valid conclusions can be reached only if successive conditions differ in terms of just 1 variable, as in the A-AB-A sequence mentioned previously. If conditions differ in multiple ways (e.g., A-ABC-A), it may be impossible to disentangle the contributions of the various factors. This *1-variable rule* can be followed even in experiments with quite a few independent variables. Table 1 shows the design of part of a study on choice (Ito and Fantino, 1986) involving 2 schedules (designated as the 'search' and 'handling' schedules) and 2 reinforcers ('short' and 'long'). Note that of the 4 variables, only 1 changes across successive pairs of conditions.

The importance of the 1-variable rule may be considerable if the design objective is to assess the effect of an experimental manipulation by comparing across pairs of successive conditions. But this is not always the case. In some experiments the objective is to study the overall pattern of outcomes from a range of related conditions.

TABLE 1
Sequence of experimental conditions in a portion of Ito and Fantino's (1986) experiment

Condition	VI schedule (s)		Reinforcer (s)	
	Search	Handling	Short	Long
1	30	5	3	6
2	5	5	3	6
3	5	20	3	6
4	15	20	3	6
5	30	20	3	6
6	30	20	2	6
7	5	20	2	6
8	15	20	2	6

Certainly this is so when quantitative variables are manipulated parametrically. As an example, consider Herrnstein's (1961) classic experiment on concurrent VI performances. As shown in the left half of Fig. 5, there were 3 independent variables: the schedule parameter on Key A, the schedule parameter on Key B, and the presence or absence of a 1.5-s changeover delay to penalize switching between the keys. Com-

Sequence of Procedures

Subject	VI on Key A (min)	VI on Key B (min)	No. of Sessions	COD
055	3	3	20	No
	2.25	4.5	18	No
	2.25	4.5	43	Yes
	3	3	44	Yes
	3	3	25	No
	9	1.8	35	Yes
	1.5	EXT*	37	Yes
	9	1.8	20	Yes
	1.8	9	39	Yes
231	3	3	35	Yes
	3	3	17	No
	9	1.8	35	Yes
	1.5	EXT*	37	Yes
	9	1.8	17	Yes
	1.8	9	40	Yes
	4.5	2.25	38	Yes
641	3	3	17	No
	2.25	4.5	16	No
	2.25	4.5	45	Yes
	3	3	34	Yes
	3	3	16	No

* Extinction

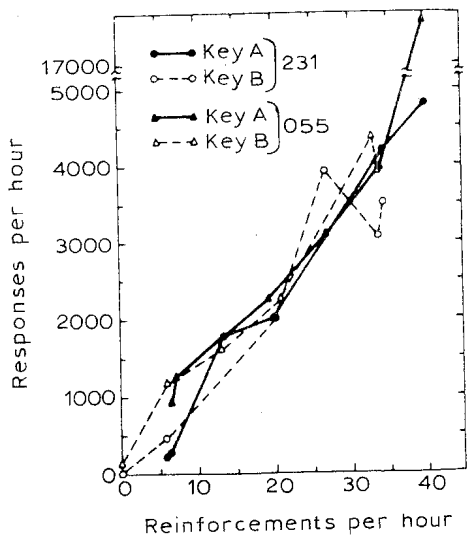


Fig. 5. Experimental conditions and illustrative results (based on terminal sessions of Pigeons 231 and 055) from a study of concurrent VI VI schedules. In addition to the 2 schedule values, the experiment also manipulated the presence of a changeover delay (COD). The violation of the 1-variable rule across pairs of successive conditions poses no problem for interpreting the pattern of results in this parametric experiment. (Reproduced from Herrnstein, 1961, with permission of the publisher.)

parison of successive sets of conditions reveals that the 1-variable rule was followed in a few instances, e.g., the VI schedules sometimes were held constant as the changeover delay was imposed or removed. But it was more common for the 2 schedule variables to be changed together and sometimes all 3 variables were changed. The right half of Fig. 5 presents the results from the 2 pigeons exposed to the most extensive manipulation of the schedule parameters. Each point represents the mean response rate during the 5 terminal sessions in the conditions with the changeover delay in effect. On both keys, response rate approximates a linear function of reinforcement rate. It is this linear pattern, not the changes in rate across any particular pair of successive conditions, that provides the basis for reaching conclusions about the effects of the reinforcement rates provided by the 2 schedules. The violation of the 1-variable rule is irrelevant.

Later sections of this chapter provide fuller discussions of experiments that manipulate quantitative variables over a wide range (Section 3.2.2, 'Parametric designs') or that manipulate two or more variables in combination (Section 3.4, 'Factorial designs').

3.2.1. Reversal designs

It is common to refer to A-B-A, A-B-A-B and similar designs as 'reversal' designs. The term might be used because each successive phase reverts to the condition that preceded it, or because the usual goal of such designs is to show that changes in behavior can be reversed by switching conditions, e.g., as in Fig. 4. Leitenberg (1973), however, suggested that the term *reversal design* be restricted to procedures (not results) in which experimental conditions are switched between incompatible behaviors.

As an example, consider an experiment in which pigeons were exposed to concurrent schedules of discriminative stimulus production (Mulvaney et al., 1974). Fig. 6, which shows the results from 1 animal, illustrates the experimental design. In the baseline phase, pecks on either of 2 side keys occasionally produced either of 2 colors correlated with the components of a multiple schedule of food reinforcement arranged on the center key – green (S+) during the VI component or red (S−) during the EXT component. Thereafter, the consequences were changed: although pecks on either key produced S+, pecks on just one of the keys could produce S−. Across conditions, the availability of S− was shifted back and forth between the right and left keys. Fig. 6 shows that responding changed systematically across the reversals, as rates on the S− key were suppressed relative to rates on the S+ key. Because the distribution of responding tracked the stimulus consequences as they were switched from key to key, it is safe to conclude that responding was controlled by these consequences rather than extraneous factors such as key bias or carryover effects from the preceding condition. Indeed, the results show that S− functioned as a conditioned punisher.

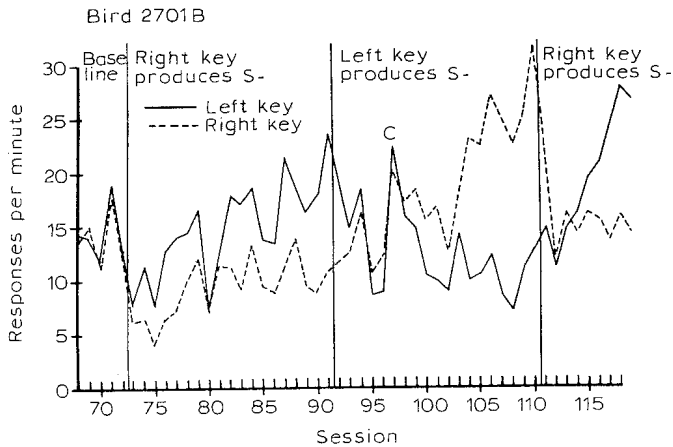


Fig. 6. *A reversal design.* Response rates of a single pigeon on 2 concurrently available keys that produced discriminative stimuli correlated with the components of a mixed VI EXT schedule programmed on a third key. Either key produced S+ (during the VI component), but only 1 produced S− (during EXT). As indicated in the panel headings, the availability of S− was reversed between the keys. Both keys produced S− and S+ during the baseline sessions, of which only a few are shown. (Reproduced from Mulvaney et al., 1974, with permission of the publisher.)

3.2.2. Parametric designs

Many designs are not given letter designations, in particular, designs involving *parametric manipulation.* When several levels of a quantitative independent variable are to be compared, the levels may be imposed across phases in ascending order, descending order, or irregular order. Often these design options are combined. For example, an ascending order of conditions may be followed by a descending order (e.g., Timberlake and Peden, 1987, Experiment 1). Or the sequence may vary across animals (e.g., Timberlake and Peden, 1987, Experiment 2). At issue is whether the outcomes depend on a specific sequence, and of course the only way to find out is by replicating the conditions in different sequences.

If the conditions are numerous or each is carried out for many sessions, the biological development of the subject may become confounded with the independent variable. The problem is most likely when a short-lived species is studied. As a case in point, consider Sidman's (1953) landmark experiment on free-operant avoidance in rats. The research involved extensive manipulation of 2 schedule parameters (the response-shock and shock-shock intervals), and each rat was exposed to about 50 experimental conditions over a period of nearly a year. This represents a significant proportion of the rat's lifespan, which is only about 2 to 3 years (see Part 1, Ch. 1). If the conditions had been arranged in a single sequence, it would have been difficult to separate the effects of the experimental manipulations from the influence of advancing age. To address this problem, Sidman varied the order across the 3 animals with the specific objective of removing the age-sequence confound. The similarity in

the data functions of the different rats established that age *per se* had negligible influence. This outcome also removes concern about any systematic influence of history, testing, and instrumentation.

If extensive variation of sequence is impractical, then at least a subset of the conditions should be replicated in a temporally distant phase of the experiment, to separate effects of the conditions from effects associated with their timing. This strategy is illustrated in Fig. 7, which shows running response rates (i.e., rates after the postreinforcement pause) of rats on variable-ratio (VR) schedules (Reed and Wright, 1988). Each animal was exposed to an ascending sequence of reinforcer magnitudes, from 1 to 4 pellets, followed by a replication of the 1-pellet condition. Response rates increased with the number of pellets, and then dropped sharply when the 1-pellet condition was reinstated, indicating that reinforcer magnitude was responsible for the changes in responding rather than, say, experience with the schedule. The importance of the replication is underscored by noting that schedule experience was perfectly confounded with magnitude across the first 4 conditions: as experience increased, so did magnitude. Thus, the replication was essential to dissociate these variables and demonstrate that experience *per se* had little influence on responding.

Another strategy is to combine variation in the sequence of conditions *between* subjects with replication of selected conditions *within* subjects. This is illustrated in Table 2, which shows the design of an experiment concerned with the effect of reinforcement rate on responding maintained by VI schedules (Catania and Reynolds, 1968, Experiment 1). Reinforcement rate was manipulated by presenting schedules with different mean intervals. The 6 subjects were exposed to the schedules in different orders, but in each case the sequence included replication of at least 1 of the schedules (e.g., Pigeon 118 had 2 exposures to the VI 108-s schedule).

To minimize carryover effects from one experimental condition to the next, any of these strategies may be augmented by interpolating a common baseline between the levels of the independent variable. For example, Lattal (1984, Experiment 3) imposed a simple VI schedule before and after conditions with delayed reinforcers. Across conditions, signals accompanied different percentages of the delays (0, 33, 66

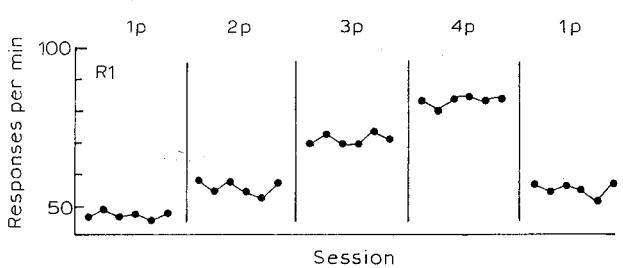


Fig. 7. *A parametric manipulation.* Terminal running response rates of 1 rat on a VR schedule, with manipulation of reinforcer magnitude across conditions (1 to 4 food pellets). Note the replication of the 1-pellet condition. (Reproduced from Reed and Wright, 1988, with permission of the publisher.)

TABLE 2
Sequence of experimental conditions in Catania and Reynold's (1968) Experiment 1. Entries are the means of VI schedules in s

Condition	Pigeon					
	118	121	129	278	279	281
1	108.0	45.5	108.0	23.5	427.0	23.5
2	45.5	23.5	216.0	12.0	216.0	45.5
3	23.5	12.0	427.0	45.5	108.0	12.0
4	12.0	108.0	23.5	216.0	23.5	427.0
5	323.0	23.5	45.5	108.0	12.0	45.5
6	108.0		12.0	45.5	45.5	12.0
7			23.5	427.0	108.0	
8			108.0			

and 100%), and the effect of this manipulation was assessed by expressing response rates in each signaled-delay condition as a proportion of the rate in the preceding VI baseline.

3.2.3. Probe designs

Ferster and Skinner (1957) and Sidman (1960) described experiments in which brief *probe manipulations* were superimposed on steady-state behavior maintained under otherwise constant conditions. Perhaps the most common use of such designs today is to study the effects of drugs on schedule-maintained responding. Several steps are involved. (a) A steady-state baseline is established under some experimental condition, either simple or complex. (b) In some sessions the experimental treatment (e.g., a drug) is added. If several levels of the probe variable are to be studied (e.g., a range of drug dosages), concerns about the sequence of exposure (e.g., ascending, descending, or random) are the same as those described for parametric designs. In any case, steps should be taken to replicate each probe a number of times, to assess the reliability of the outcomes within and across levels of the probe variable. (c) Interspersed among the probe sessions are additional baseline sessions in which behavior is allowed to regain its previous characteristics. This is to ensure that each probe is applied to comparable behavior. In practice, the number of baseline sessions is likely to match or exceed the number of probe sessions. Examples of probe designs in the analysis of drug effects can be found in Part 2, Ch. 2.

3.2.4. Multiple-baseline designs

The designs described so far are based on the assumption that the behavioral changes under study are reversible. In experiments with an A-B-A design, for example, changes from the first phase to the second can be attributed to the experimental ma-

nipulation (Condition B) only if behavior changes back to baseline levels (or nearly so) when Condition A is reinstated in the third phase. If the behavior does not change back, the result is ambiguous, i.e., it is not clear whether the initial change reflects the irreversible effect of the experimental treatment or the operation of history, maturation, testing, or instrumentation.

Multiple-baseline designs allow the study of irreversible effects. A multiple-baseline design is an extension of the A-B design, with the following modifications: (a) 2 or more independent behaviors are studied simultaneously, and (b) Condition B is applied to the behaviors at different times. If each behavior changes only upon introduction of the experimental manipulation, regardless of its timing, then it is reasonable to attribute the change to the manipulation rather than time-related extraneous factors.

Although multiple-baseline designs are used often in research with humans, especially when ethical considerations preclude the withdrawal of therapeutic treatments (for a review see Barlow and Hersen, 1984, Part 1, Ch. 7), they are quite rare – perhaps nonexistent – in research with animals. This probably reflects the current emphasis on reversible processes rather than deficiencies in the design strategy itself. At the moment, then, multiple-baseline designs constitute an untapped resource for the experimental analysis of irreversible phenomena.

3.3. Designs with simultaneous comparisons

Comparisons need not be made across successive phases of an experiment. They also can be made more-or-less simultaneously by way of *multi-element designs* (Sidman, 1960, Part 2, Ch. 3) that arrange two or more conditions within each session of the experiment.

3.3.1. Multi-element manipulations

In some cases the purpose is to compare the conditions directly, much as one would in a successive-comparisons design. According to Sidman (1960, pp. 326–330), this strategy involves *multi-element manipulations*. For example, Mazur and Hyslop (1982) assessed the effects of a timeout on fixed-ratio (FR) performance by inserting 30-s dark-key periods before a random half of the ratios in each session. In so doing, they were able to compare pre-ratio pauses with and without timeout in a single block of sessions. In contrast, successive comparison of the same conditions would require at least 3 blocks of sessions, arranging timeout and no-timeout conditions in A-B-A fashion, as in a study by Richards and Blackman (1981).

3.3.2. Multi-element baselines

In other cases the purpose is to allow simultaneous evaluation of a variable's effects on two or more different forms of behavior, by using what Sidman (1960, pp. 323–326) called a *multi-element baseline*. Fig. 8 shows the results of an experiment that used

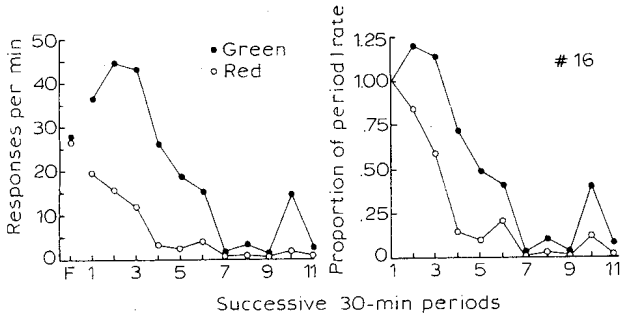


Fig. 8. *A multi-element baseline design. Left panel*: Response rates of 1 pigeon at the end of training on a multiple schedule with VI 2-min (green key) and VI 6-min (red key) components (marked by F on the X-axis), and during a 5.5-h test session in which the VI schedules were replaced with EXT. *Right panel*: Rates during the test session, expressed as a proportion of the rate during the first 30-min period of the test. (Reproduced from Nevin, 1974, with permission of the publisher.)

a multiple schedule to compare resistance to extinction in behavior that had been maintained by different rates of reinforcement (Nevin, 1974, Experiment 2). Pigeons were trained with a VI 2-min schedule in one component (associated with a green keylight) and a VI 6-min schedule in the other (red keylight). The 2 components alternated every 30 s. After responding stabilized, both schedules were changed to extinction (but the colors continued to alternate) during a single 5.5-h test session. Responding in both components decreased over the course of the test, but responding in the component previously associated with the richer schedule (filled circles in Fig. 8) was slower to change – i.e., more resistant to extinction – than responding in the other component.

3.3.3. Concurrent multi-element designs

In the multi-element experiments discussed so far, the different conditions alternated within the session. Another approach is to arrange the conditions concurrently (*concurrent multi-element designs*). In one such study, the influence of drugs on 2 forms of aversively motivated behavior was studied by training rats with concurrent schedules of negative reinforcement (Galizio and Perone, 1987). Presses on one lever postponed shock according to a free-operant avoidance schedule, while presses on the other produced 2-min timeout periods during which the avoidance schedule was suspended. The administration of a drug during probe sessions allowed a simultaneous comparison of its effects on the concurrent baseline responses; e.g., chlordiazepoxide enhanced rates on the timeout lever at doses that did not affect rates on the avoidance lever.

Another example of a concurrent multi-element design – this time with concurrent manipulations instead of concurrent baselines – is provided by a study in which pigeons responded on concurrent chain schedules with identical initial links but different terminal links (Fantino, 1968). The initial links consisted of equal, but indepen-

dent, VI schedules. In one of the terminal links food was contingent on completing an FI; in the other it was contingent on fast-paced responding (differential reinforcement of high rates). The objective was to compare the effects of these 2 terminal-link conditions on responding in the concurrent initial links. The response rate was higher in the initial-link leading to the unpaced terminal link.

3.3.4. Special issues

Multi-element designs offer at least 2 advantages over successive-comparison designs. First, and most obvious, is the efficiency introduced by comparing effects more-or-less simultaneously, i.e., within a single session or block of sessions, rather than across successive blocks. Second, the simultaneous nature of the comparisons equalizes the influence of time-related extraneous factors across the conditions of interest.

A potential disadvantage is that the juxtaposition of conditions allows interactions between them (*schedule interactions*). In other words, when 2 conditions are arranged within the same session, the resulting behavior may differ from that engendered by the conditions in isolation of one another. Numerous studies have shown that responding in one component of a multiple schedule can be affected by the contingency arranged in the other component (Schwartz and Gamzu, 1977), and that responding in one member of a pair of concurrent schedules can be affected by the other member of the pair (Davison and McCarthy, 1988). Such findings suggest a need for caution in interpreting the results from multi-element experiments.

As a practical matter, interactions in multiple schedules can be reduced by increasing the duration of the components (minimizing the number of transitions across them) or by programming timeouts between the components (separating them in time); details are provided in Part 1, Ch. 3. Thus, multiple schedules are well suited to multi-element designs, so much so that multi-element designs are sometimes called 'multiple-schedule designs' (cf. Barlow and Hersen, 1984, pp. 254–256).

Comparable remedies are not available with concurrent schedules, however. Here the schedules are truly simultaneous and they compete directly for control over the subject's behavior. Thus, most contemporary use of concurrent schedules is directed not to multi-element comparisons, but rather to the analysis of schedule interaction as an object of study in its own right. Interaction between concurrent responses is an important characteristic of free-operant behavior as it may occur in complex natural environments, and the analysis of such interaction has provided the primary empirical and theoretical basis for operant models of choice, as well as a link to related topics such as animal foraging (e.g., Part 2, Ch. 4), economics (e.g., Hursh, 1984), and human judgment and decision-making (e.g., Rachlin, 1989).

But when schedule interactions are regarded as a nuisance, in multiple- or concurrent-schedule arrangements, the conservative researcher may consider assessing their contribution empirically by replicating the conditions of interest in an experiment with successive comparisons (cf. Dinsmoor, 1966). To the extent that the results are consistent, concerns about interactions are reduced, and the results may be regarded as having what Dinsmoor called 'procedural generality'.

3.4. Factorial designs

The basic designs described so far can be merged in various ways to address the operation of several independent variables in combination. When a design includes all possible combinations of the levels of 2 or more independent variables, it is called a *factorial design*. Such designs are ubiquitous in the behavioral and social sciences, as even a cursory glance through the literature will attest. Yet the term 'factorial design' rarely appears in the context of single-subject research on free-operant behavior, even though it would seem to describe much of that research. Perhaps the term has been avoided intentionally because of its origins in the groups-comparison approach and its close association with the analysis of variance. Unfortunately, the omission may have contributed to an impression which, judging from numerous undergraduate textbooks on research methods, appears to be widespread: namely, that single-subject designs are somehow less sophisticated than group-stastistical ones. On the contrary, the literature of the Experimental Analysis of Behavior abounds with examples of factorial designs, even if they are not identified as such by the authors.

3.4.1. Designs with two factors

For an example of a single-subject factorial design, consider Mazur and Hyslop's (1982) experiment on FR performance in which a 30-s timeout was inserted before a random half of the ratios. This multi-element manipulation was combined with a parametric manipulation of ratio size, allowing study of 2 independent variables or *factors*: timeout condition, assessed within sessions, and ratio size, assessed across blocks of sessions. There were 2 levels of the timeout factor (no timeout, timeout) and 3 of the ratio factor (50, 100, 150), for a total of 6 combinations – *a 2 × 3 factorial design*. The results are presented in Fig. 9, which shows data on pre-ratio pausing averaged over the 10 terminal sessions in each condition.

As is customary in factorial experiments, Fig. 9 allows assessment of main effects and interaction effects (the latter of which should not be confused with the schedule

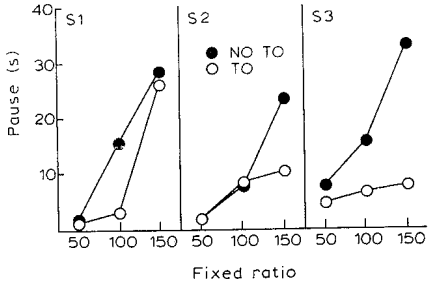


Fig. 9. *A 2 × 3 factorial design.* Median pre-ratio pauses of 3 pigeons as a function of FR size and the presence or absence of a 30-s timeout after the reinforcer. Data are from the 10 terminal sessions in each condition. (Reproduced from Mazur and Hyslop, 1982, with permission of the publisher.)

interactions discussed previously). A *main effect* is the overall influence of 1 factor, averaged over the levels of the other factors. The data in Fig. 9 can be used to evaluate 2 main effects, corresponding to the influence of the timeout condition and the ratio size (i.e., 1 main effect for each factor in the design). Main effects of both factors are evident. The effect of the timeout condition can be seen by averaging across the ratio sizes; pausing was shorter with the timeout (open circles) than without it (filled circles). The effect of the ratio size can be seen by averaging across the timeout conditions; pausing increased as the ratio size was raised.

The importance of the main effects in Fig. 9 is diminished by the presence of an interaction between the factors. In a factorial experiment, an *interaction* is said to be present when the effect of one factor depends on the level of another factor. When such dependencies are present, descriptions of the data in terms of main effects will tend to be oversimplifications. In this case, the effect of the timeout depended on the ratio size; the exact pattern varies somewhat across the birds, but in general it seems fair to say that reliability of the timeout effect was enhanced as the ratio was raised (only one bird shows a difference between timeout and no-timeout conditions at the small ratio, two show a difference at the medium ratio, and all three show a difference at the large ratio). This dependency serves to underscore the importance of the parametric manipulation. If the experiment had been restricted to a single ratio, e.g., FR 50, one might have concluded that the timeout had no reliable effect at all. At the same time, the subject-to-subject variation in the ratio with the largest timeout effect suggests a need for a more extensive manipulation of both ratio size and timeout duration.

3.4.2. Designs with three factors

An experiment from the author's laboratory illustrates a more complex factorial design. Pigeons were trained on a multiple schedule in which completing an FR 80 produced either 1-s or 7-s access to grain, with different key colors accompanying the 2 types of ratio. As in Mazur and Hyslop's (1982) experiment, half of the reinforcers were followed by a timeout. This arrangement made it possible to study the effects of 3 factors on pre-ratio pausing: the magnitude of the reinforcer delivered prior to the pause (the 'past reinforcer', small or large), the magnitude of the reinforcer to be delivered upon completion of the next ratio (the 'upcoming reinforcer', small or large), and the timeout condition (no timeout, timeout). The experiment had a *2 × 2 × 2 factorial design* and, therefore, 8 combinations of the levels of the 3 factors. Despite its complexity, the experiment was not time-consuming. Because all 3 factors were manipulated within each session, data collection was complete within a single block of 33 sessions.

Results from 1 pigeon are shown in Fig. 10. Each point represents the median pause during the 10 terminal sessions, and each vertical line represents the associated interquartile range (25th to 75th percentiles of the pause distribution). Including indices of variability in the figure allows the reader to assess the stability of the patterns

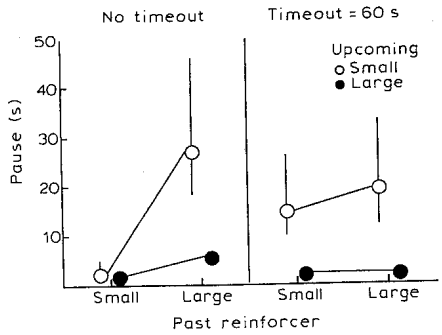


Fig. 10. *A 2 × 2 × 2 factorial design.* Median pre-ratio pauses of a single pigeon responding on an FR 80 schedule, as a function of the magnitude of the past reinforcer, magnitude of the upcoming reinforcer (as signaled by discriminative stimuli), and the presence or absence of a 60-s timeout after the reinforcer. Vertical lines extend from the 25th to 75th percentiles unless these values are encompassed by the circle. Data are from the 10 terminal sessions, each of which included all 8 conditions. (Unpublished data.)

of pausing, and hence the reliability of any differences across the conditions. The left panel shows the results from the conditions without timeout, where there was a clear interaction between the past and upcoming reinforcers. When the upcoming reinforcer (which, remember, was signaled by key color) was small, the past reinforcer had a substantial influence on pausing, with short pauses after the small reinforcer and long ones after the large reinforcer (open circles). By comparison, when the upcoming reinforcer was large, the past reinforcer had relatively little influence, with short pauses after both small and large reinforcers (filled circles).

Main effects of the past and upcoming reinforcers also are evident in the left panel of Fig. 10. On average, pause duration was directly related to the magnitude of the past reinforcer and inversely related to the upcoming reinforcer. But, in light of the interaction between the past and upcoming reinforcers, these main effects do not adequately describe the findings – in other words, anything that might be said about the effects of the past reinforcer must be qualified by some consideration of the upcoming reinforcer, and vice versa.

The right panel of Fig. 10 shows the results from the conditions when a timeout intervened between the delivery of a reinforcer and the opportunity to begin a new ratio. In this situation the past reinforcer had minimal influence on pre-ratio pausing, regardless of the size of the upcoming reinforcer, and so there was no interaction between the past and upcoming reinforcers (although the medians differ when the upcoming reinforcer was small, the substantial overlap in the interquartile ranges indicates that the difference is not reliable). The upcoming reinforcer, however, did have an effect of its own: when the key color signaled that the next reinforcer was large, pauses were short (filled circles), but when the next reinforcer was small, pauses were long (open circles). In other words, in the timeout conditions the only clear effect was a main effect of the upcoming reinforcer.

The factorial manipulation of the 3 independent variables in this experiment allows more precise – and thus more general – statements about their effects than would have been possible had any 1 or 2 of the variables been studied in isolation. Of particular interest is the discovery of a *higher order interaction* among the 3 variables. In a higher order interaction, the degree of interaction between a set of factors depends on the level of another factor. In the present case, the interaction between the magnitudes of the past and upcoming reinforcers depended on the presence or absence of the timeout. This complex outcome suggests that pausing between FRs results from competition between 2 sources of control: the inhibitory aftereffects of the past reinforcer (cf. Harzem and Harzem, 1981) and discriminative control by stimuli correlated with the upcoming reinforcer (cf. Griffiths and Thompson, 1973; Inman and Cheney, 1974).

3.4.3. Additional considerations

The primary advantage of factorial designs is that they provide a framework to guide the search for interactions among the variables that control behavior. The search is important, because knowledge of the presence or absence of interactions will tend to restrict or expand the scope of statements describing experimental outcomes. If an interaction is present, it means that the operation of one variable is bounded in some way by the operation of another (or, in the case of higher order interactions, the interaction among some set of variables is bounded by another variable). By comparison, the absence of an interaction implies that simple descriptions of effects will have broad application, spanning the levels of the other factor or factors that have been studied.

The considerations described above will lead experimenters to consider including in their designs as many relevant factors as possible. From the standpoint of design *per se*, the only limits on the number of factors (and the levels of each) are those imposed by the physical resources of the laboratory and the ingenuity of the experimenter. From a more practical standpoint, however, it should be recognized that the interpretative burden does increase with the number of factors. It may be difficult indeed to take account of more than 3 or 4 factors at once.

3.5. Designs with yoked comparisons

Discussions of experiments on free-operant behavior sometimes include 'yoked-control designs', and so brief mention is warranted here even though yoking is more aptly described as a technique of experimental control rather than a form of experimental design. Indeed, yoking procedures can be imbedded within any of the design types discussed so far (e.g., multi-element design: Peele et al., 1984, Experiment 1; A-B-A designs: Pellon and Blackman, 1987).

Yoked-control procedures are used to equate the occurrence of certain events across experimental conditions that vary in some other way. The first use was by Ferster

and Skinner (1957, pp. 399–407) in an analysis of the differences in response rates maintained by ratio and interval schedules. On VR schedules, changes in responding are accompanied by changes in reinforcement frequency, and it is possible that the high rates typically engendered by these schedules are due to the frequent reinforcement rather than the ratio contingency per se. To separate these factors, Ferster and Skinner 'yoked' 2 experimental chambers. In the 'master' chamber, a pigeon responded on a VR schedule. Every time the master bird received a reinforcer, a reinforcer was set up in the 'yoked' chamber, to be delivered upon the next peck of the pigeon there. This arrangement equated the number and temporal distribution of reinforcement in the 2 chambers, but the master schedule was a VR while the yoked was a VI.

In a well-known critique of the yoked control, Church (1964) noted that individual differences between master and yoked subjects (e.g., in reactivity to the experimental event) can bias the results in favor of the master. Recent technological advances reduce this concern, however, as yoking can be arranged on a within-subject basis (Lattal and Ziegler, 1980), and even on a within-*session* basis. In one study, for example, the availability of reinforcement in one component of a multiple schedule was yoked to the distribution of reinforcement produced by the subject's performance in a prior component (Peele et al., 1984). A range of opinion on yoked controls can be found in Gardner and Gardner's (1988) article, which includes noteworthy commentaries by Church, Dickinson and Mackintosh, Thomas, and Wasserman.

3.6. External validity

When the objective is to generalize the outcome of an experiment across populations, settings, independent variables, or dependent variables, *external validity* is at issue. Establishing the external validity of single-subject research is difficult, but in this respect single-subject research does not differ from group research. As Campbell and Stanley (1963) noted, the question of external validity "is never completely answerable" (p. 5). Assessing external validity is an inductive process, and as such it cannot be established on logical grounds. Rather, it is established by verifying initial observations in an ever widening set of conditions, e.g., by showing empirically that relations observed in one setting also occur in different settings.

Cook and Campbell (1979) identified 3 major *threats to external validity*: *interaction of selection and treatment*, *interaction of setting and treatment*, and *interaction of history and treatment*. In the experimental analysis of free-operant behavior, the concerns are whether an experimental outcome is dependent on the use of subjects with some specific characteristics, on a particular laboratory manipulation, or on the conditioning histories of the subjects. According to Cook and Campbell (1979, pp. 78–79), these issues can be resolved only through replication.

Sidman (1960) stressed the importance of replication by devoting two chapters of his book to the topic. The need for within-subject and between-subject replication in

controlling threats to internal validity has already been noted. Sidman also described 2 forms of between-subject replication that provide a mechanism for establishing external validity. The first is *direct replication*, which is accomplished simply by repeating an experiment with new subjects. Adding subjects to a study does not represent abandonment of a single-subject design; rather it is an effort to replicate the entire experiment and thereby assess the generality of the findings across individuals (Sidman, 1960, pp. 74–75). The second form of between-subject replication is called *systematic replication*. This involves an effort to replicate a functional relation under circumstances that differ from those under which the relation was originally discovered. The change can be relatively minor – e.g., the experimental conditions might be scheduled in a different order – or relatively major – e.g., a different species of subject might be used, a different form of behavior might be studied, or major constructs might be operationalized in new ways.

4. Conclusion

Experimental methods to analyze free-operant behavior have evolved in an intellectual tradition that regards behavior as a continuous interaction between an individual organism and its environment, to be pursued as an object of study in its own right rather than a proxy for unobservable processes or structures. The emphasis on single-subject designs represents the tradition's commitment to seek order at the level of the individual by identifying relevant variables and bringing them under experimental control.

Single-subject experiments compare two or more steady states engendered in the same animal. Control is demonstrated by the reliable production of behavior that changes systematically across the levels of the independent variable but remains stable within each level. Decisions about stability are based on criteria that define steady states by setting tolerances on systematic and non-systematic variation in the data (trend and bounce). Quantitative stability criteria may specify the limits in absolute or relative terms. Both types vary in stringency as response rates change; absolute criteria become more difficult to meet as rates rise and relative criteria become more difficult as rates fall. Thus, it may not be feasible to adhere to a single criterion in experiments that generate a wide range of rates across the experimental conditions. Non-quantitative criteria may be based on visual inspection of the data or on fixed lengths of exposure to each condition. Results obtained with these methods should be accompanied by some account of the factors guiding the experimenter's judgment as well as empirical evidence on the levels of stability actually attained.

Because each individual is exposed to several levels of the independent variable, it is possible for experimental manipulations to be confounded with extraneous factors operating as a function of time or repeated exposure to the laboratory procedures (history, maturation, testing, instrumentation). Although many of these threats

to internal validity can be controlled by eliminating them or holding them constant, others will continue to operate because their link to the research protocol or the passage of time is intrinsic (e.g., schedule experience, biological development). Single-subject experiments must be designed to dissociate the effects of such factors from those of the independent variable. Many design options are available. Conditions may be arranged successively, as in A-B-A, reversal, parametric, probe and multiple-baseline designs, or simultaneously, as in the various multi-element designs. Or these strategies may be merged in factorial designs that allow for the analysis of interactions between 2 or more independent variables.

In every case, the key to valid single-subject research is in judicious use of replication, whether within-subject or between-subject, direct or systematic. If successful in demonstrating the ability to reproduce previously-observed states of behavior, replication establishes the experimenter's success in identifying and controlling relevant variables and confirms the adequacy of the stability criteria guiding decisions about the attainment of steady states. By allowing experimental comparisons to be made across sets of conditions arranged in different sequences, replication separates the effects of the conditions themselves from effects associated with their ordinal position or timing, thereby freeing the comparisons of various threats to internal validity. Finally, by seeking to reproduce functional relations across a range of individuals, species, responses, and operations, replication allows the experimenter to discover the boundaries of a relation's external validity.

This chapter has discussed many of the general issues that an experimenter is likely to confront when designing and conducting research on free-operant behavior. In addition, the interested reader would do well to study Sidman's (1960) classic treatise, which remains the definitive statement on experimental method in the operant tradition. However, in moving from general issues to specific details, the reader will discover that there are no simple rules to decide which stability criterion or which design strategy is 'best' in a given situation. Whatever the plan, it will have advantages and disadvantages. A good way to tip the balance in favor of the advantages is to consult the professional literature in the area of interest, where answers will be found to many of the unique problems that arise in that area. Still, as valuable as these efforts may be, there is a limit to what one can learn by reading. As Bachrach (1981) has noted, "People don't usually do research the way people who write books about research say that people do research" (p. 2), which is just to say that the formal reconstructions of research that appear in chapters and journal articles cannot capture everything important about the processes of science. In the final analysis, there is no substitute for experience in the laboratory, where the subject's behavior will come to control the experimenter's.

Acknowledgements

The experiment reported in Fig. 10 was conducted in collaboration with Barbara Metzger. I thank Barbara Metzger and Barbara J. Kaminski for their gracious assistance in preparing Figs. 1, 2, 9 and 10.

References

Arbuckle, J.L. and Lattal, K.A. (1988) Changes in functional response units with briefly delayed reinforcement. J. Exp. Anal. Behav. 49, 249–263.

Bachrach, A.J. (1981) Psychological Research, 4th Edn. Random House, New York.

Baer, D.M. (1977) 'Perhaps it would be better not to know everything.' J. Appl. Behav. Anal. 10, 167–172.

Baer, D.M. (1988) An autocorrelated commentary on the need for a different debate. Behav. Assess. 10, 295–297.

Bakan, D. (1970) The test of significance in psychological research. In: D.E. Morrison and R.E. Henkel (Eds.), The Signifiance Test Controversy. Aldine, Chicago, pp. 231–251.

Barlow, D.H. and Hersen, M. (1984) Single Case Experimental Designs. Pergamon Press, New York.

Bitterman, M.E. (1966) Animal learning. In: J.B. Sidowski (Ed.), Experimental Methods and Instrumentation in Psychology. McGraw-Hill, New York, pp. 451–484.

Busk, P.L. and Marascuilo, L.A. (1988) Autocorrelation in single-subject research: a counter-argument to the myth of no autocorrelation. Behav. Assess. 10, 229–242.

Campbell, D.T. and Stanley, J.C. (1963) Experimental and Quasi-Experimental Designs for Research, Rand McNally, Chicago.

Catania, A.C. and Reynolds, G.S. (1968) A quantitative analysis of the responding maintained by interval schedules of reinforcement. J. Exp. Anal. Behav. 11, 327–383.

Christensen, L.B. (1985) Experimental Methodology, 3rd. Edn. Allyn and Bacon, Boston.

Church, R.M. (1964) Systematic effect of random error in the yoked control design. Psychol. Bull. 62, 122–131.

Cook, T.D. and Campbell, D.T. (1979) Quasi-Experimentation. Rand McNally, Chicago.

Cumming, W.W. and Schoenfeld, W.N. (1960) Behavior stability under extended exposure to a time-correlated reinforcement contingency. J. Exp. Anal. Behav. 3, 71–82.

Davison, M. and McCarthy, D. (1988) The Matching Law. Erlbaum, Hillsdale, NJ.

DeProspero, A. and Cohen, S. (1979) Inconsistent visual analysis of intrasubject data. J. Appl. Behav. Anal. 12, 573–579.

Dinsmoor, J.A. (1966) Operant conditioning. In: J.B. Sidowski (Ed.). Experimental Methods and Instrumentation in Psychology. McGraw-Hill, New York, pp. 421–449.

Dinsmoor, J.A., Mulvaney, D.E. and Jwaideh, A.R. (1981) Conditioned reinforcement as a function of duration of stimulus. J. Exp. Anal. Behav. 36, 41–49.

Dukich, T.D. and Lee, A.W. (1973) A comparison of measures of responding under fixed-interval schedules. J. Exp. Anal. Behav. 20, 281–290.

Fantino, E. (1968) Effects of required rates of responding upon choice. J. Exp. Anal. Behav. 11, 15–22.

Fantino, E. and Logan, C.A. (1979) The Experimental Analysis of Behavior: a Biological Perspective. W.H. Freeman & Co., San Francisco, CA.

Ferster, C.B. (1953) The use of the free operant in the analysis of behavior. Psychol. Bull. 50, 263–274.

Ferster, C.B. and Skinner, B.F. (1957) Schedules of Reinforcement. Prentice-Hall, Englewood Cliffs, NJ.

Flaherty, C.F. (1985) Animal Learning and Cognition. Knopf, New York.

Fry, W., Kelleher, R.T. and Cook, L. (1960) A mathematical index of performance on fixed-interval schedules of reinforcement. J. Exp. Anal. Behav. 3, 193–199.