CHAPTER 7

MULTIPLE BASELINE DESIGNS

7.1. INTRODUCTION

Withdrawal or reversal designs may be undesirable or even inappropriate when treatment variables cannot be withdrawn or reversed due to ethical considerations, practical limitations, or problems in subject or staff cooperation. The researcher's goal of identifying which treatment components are most effective and of demonstrating the controlling influence of these components by returning undesirable behavior in their absence is often at odds with the desires of the subject and clinical staff, who most often want the undesirable behaviors to go away quickly, and not to return.

Ethical considerations are of paramount importance when the treatment variable is effective in reducing self-injurious or aggressive behaviors. Here the withdrawal of treatment is obviously unwarranted, even for brief periods of time. Variations of the withdrawal design may be used in such instances, such as quasi-experimental withdrawal designs (e.g., Wallenstein & Nock, 2007), but these present limitations on the causal inferences that can be drawn. In still other instances, withdrawal of treatment, despite absence of harm to the subject or others in his or her environment, may be undesirable because of the severity of the disorder. Here the importance of preserving therapeutic gains is given priority, especially when a disorder has a lengthy history and previous efforts at remediation have failed.

A related concern is environmental cooperation. Even if the behavior in question does not have immediate destructive effects on the environment, if it is considered to be aversive (i.e., by family, parents, teachers, or clinical staff) the experimenter will not obtain sufficient cooperation to carry out withdrawal or reversal of treatment procedures. Under these circumstances, it is clear that the researcher must seek alternative experimental strategies.

Another potential limitation that can arise is one in which carryover effects appear across adjacent phases of study, such as in the case of therapeutic instructions or drugs with long-lasting effects. Also, when multiple behaviors within an individual are targeted for change, withdrawal designs may not provide the most
elegant strategy for evaluation. In summary, in many situations withdrawal reversal designs may not be the best methodological option.

There are three main design options the clinical researcher should consider as strong alternatives to withdrawal or reversal designs when presented with problems mentioned in the preceding paragraphs. One option is the use of a changing criterion design, which was discussed earlier (see chapter 6). A second option is the use of a multiple-baseline design, which is the focus of the current chapter. The third option, use of an alternating treatment design, will be discussed in the next chapter (see chapter 8).

In this chapter we present the rationale and procedures for multiple baseline designs. Examples of the three principal varieties of multiple baseline strategies will be presented for illustrative purposes. In addition, we will consider more recent varieties and permutations, including the nonconcurrent multi baseline design across subjects and the multiple-probe technique. Finally, specific issues that may arise in the use of multiple baseline designs in drug evaluation will be discussed.

7.2. MULTIPLE BASELINE DESIGNS

Historically, the first uses of multiple-baseline designs appeared in the clinical science literature approximately forty years ago. The rationale for the multiple baseline design first appeared in the applied behavioral literature in 1968 (B et al.), in which they explained:

In the multiple-baseline technique, a number of responses are identified and insured over time to provide baselines, against which changes can be evaluated. When these baselines established, the experimenter then applies an experimental variable to one of the behaviors, produces a change in it, and perhaps notes little or change in the other baselines. (p. 94)

Since that time, the multiple baseline design has become the most well-known and widely used of the three alternatives to withdrawal or reversal designs. Indeed, a brief search on PubMed (an electronic text-based article search and retrieval system available on the internet at www.pubmed.gov) reveals that as of June 6, 2007: 30 articles are listed using the term “changing criterion,” articles using the term “alternating treatments,” and 1,036 articles using term “multiple baseline.”

Multiple-baseline designs are those in which the treatment variable is introduced in temporal sequence to different behaviors, subjects, or settings. The power of such designs comes from demonstrating that change occurs when and only when, the intervention is directed at the behavior, setting, or subject question. As with reversal designs, multiple-baseline designs begin with a baseline phase that continues until behavioral stability is demonstrated, at which time the treatment variable is introduced. As mentioned above, a change in behav
that occurs only when the intervention is introduced suggests that the intervention caused the change; however, the influence of other factors (e.g., history, maturation, etc.) must be ruled out in order to increase the validity of this claim. Rather than using a withdrawal or reversal, multiple baseline designs replicate the effect of the treatment variable across different behaviors, settings, or individuals. The temporal sequencing element is vital in order to rule out the likelihood that history, maturation, or other extraneous factors could account for the observed behavior change.

Each baseline and introduction of the treatment variable can be conceptualized as separate A-B designs, with the A phase further extended for each of the succeeding baselines (i.e., behaviors, subjects, or settings) until the treatment variable is applied. The experimenter is assured that the treatment variable is effective when a change in rate appears after its application while the rate of concurrent (untreated) behaviors remains relatively constant. A basic assumption is that the targeted behaviors are independent from one another. If they happen to covary (i.e., change is observed across all baselines when the intervention is administered to the first behavior), then the controlling effects of the treatment variable are subject to question, and limitations of the A-B analysis fully apply (see chapter 5).

In some cases, when independence of behaviors is not found, application of the alternating treatment design may be recommended (see chapter 8). In other cases, application of the multiple baseline design across different subjects might yield useful information. Kazdin and Kopel (1975) offered three additional recommendations for dealing with instances in which the effects of the treatment variable may be general rather than specific. The first is to include baselines that topographically are as distinct as possible from one another, although this may be difficult to ascertain on an a priori basis. The second is to use four or more baselines rather than two or three. However, there is the statistical probability that interdependence will be enhanced with a larger number. The third (on an ex post facto basis) is to withdraw and then reintroduce treatment for the correlated baseline (as in the B-A-B design), thus demonstrating the controlling effects over that targeted response. Even though the multiple baseline strategy was implemented in the first place to avoid treatment withdrawal, the rationale for such temporary (or partial) withdrawal in the multiple baseline design across behaviors seems reasonable when independence of baselines cannot be documented.

The issue regarding how many baselines are needed to establish confidence in the controlling effects of treatment is one that has been debated a bit in the literature. Baer et al. (1968) initially considered this issue to be an "audience variable" and were reluctant to specify the minimum number of baselines required. Theoretically, only two baselines are needed to derive useful information. However, consistent with earlier views on this topic (e.g., Barlow and Hersen, 1973; Kazdin and Kopel, 1975; Wolf and Risley, 1971), we recommend using at least three baselines, with the use of four or more baselines increasing the strength of the design, if practical and experimental considerations permit.
Although demonstration of the controlling effects of a treatment variable is weaker in the multiple baseline design than in withdrawal designs, a major advantage of the former is that it facilitates the simultaneous measurement of multiple target behaviors. This is significant for two reasons. First, the monitoring of concurrent behaviors allows for a closer approximation to naturalistic conditions, where a variety of responses are occurring at the same time. Second, examination of concurrent behaviors fosters an analysis of covariation among targeted behaviors. The co-occurrence of maladaptive behaviors has long been of interest to epidemiologists (e.g., Angold, Costello, & Erkanli, 1999; Kessler, Chiu, Demler, Merikangas, & Walters, 2005), basic behavioral research (Goodwin & Gotlib, 2004; Lewinsohn, Rohde, Seeley, Klein, & Gotlib, 2000), and those studying the treatment of maladaptive behaviors (Doss & Weiss, 2000; Kazdin & Whitley, 2006). In fact, in their introduction to a special journal section focused on the treatment implications of comorbid psychopathology, Kendall and Clarkin (1992) called comorbidity “the premier challenge facing mental health professionals” (p. 833). The single-case experimental design in general, and the multiple baseline strategy in particular, is especially well suited to advancing understanding of how treatment variables may influence not only the key target behavior, but other behaviors as well. Kazdin (1973) cogently underscored the importance of measuring concurrent (untreated) behaviors when assessing the efficacy of reinforcement paradigms in applied settings in stating that:

While changes in target behaviors are the raison d’être for undertaking treatment training programs, concomitant changes may take place as well. If so, they should be assessed. It is one thing to assess and evaluate changes in a target behavior, quite another to insist on excluding nontarget measures. It may be that investigators are short-changing themselves in evaluating the programs. (p. 527)

**Types of multiple baseline designs**

There are three basic types of multiple baseline designs. In the first, the *multiple baseline design across behaviors*, the same treatment variable is applied sequentially to separate (independent) target behaviors within a single subject. A possible variation of this strategy, of course, involves the sequential application of a treatment variable to targeted behaviors for an entire group of subjects (e.g., Macklin, Glasgow, O’Neill, & Klesges, 1984). Of course, in such cases it would be best if the experimenter presents data for individual subjects, demonstrating the sequential treatment applications to independent behaviors affected most subjects in the same direction.

In the second design, the *multiple baseline design across subjects*, a particular treatment is applied in sequence across matched subjects presumably exposed “identical” environmental conditions. Thus, as the same treatment variable is applied to succeeding subjects, the baseline for each subject increases in length, in contrast to the multiple baseline design across behaviors (which is a within-subjects design).
multiple baseline design), in the multiple baseline design across subjects a single targeted behavior serves as the primary focus of inquiry. Of course, there is no experimental contraindication to monitoring concurrent (untreated) behaviors as well. On the contrary, it is likely that the monitoring of concurrent behaviors will lead to additional findings of merit.

As with the multiple baseline design across behaviors, a possible variation of the multiple baseline design across subjects involves the sequential application of the treatment variable across entire groups of subjects (e.g., Porritt et al., 2006). Here too, however, it is best if the experimenter demonstrates that a large majority of individual subjects from each group evidenced the same effects of treatment. Notably, the multiple baseline design across subjects has also been labeled a time-lagged control design (e.g., Allen & Shriver, 1997; Gottman, 1973). In fact, this strategy was followed by Hilgard (1933) some 75 years ago in a study in which she examined the effects of early and delayed practice on memory and motoric functions in a set of twins (method of co-twin control).

In the third design, the multiple baseline design across settings, a particular treatment is applied sequentially to a single subject or a group of subjects across independent situations. For example, in a classroom situation, one might apply time-out contingencies for unruly behavior in sequence across different classroom periods. The baseline period for each succeeding classroom period, then, increases in length before application of the treatment. As in the across-subjects design, assessment of treatment is usually based on rate changes observed in a selected target behavior. However, once again the monitoring of concurrent behaviors might prove to be of value and should be encouraged where possible. Here too, this multiple baseline strategy may be used to study groups of subjects rather than individual subjects (e.g., Lohrman & Talerico, 2004). In the following three subsections we provide illustrations of the use of these basic multiple baseline strategies, as well as variations on these designs, selected from various clinical literatures.

Multiple baseline design across behaviors

An early example of a multiple baseline design across behaviors appeared in a study by Bornstein, Bellack, and Hersen (1977). These authors used this design to assess the effects of social skills training in the social performance of an unassertive 8-year-old third grader (Tom) whose passivity led to derision by his peers. Generally, if Tom experienced conflict with a peer, he cried or reported the incident to his teacher. Three target behaviors were selected for modification as a result of role-played performance in baseline: ratio of eye contact to speech duration, number of words spoken, and number of requests made. In addition, independent evaluations of overall assertiveness, based on role-played performance, were obtained. As can be seen in Figure 7.1, baseline responding for targeted behaviors was low and stable. Following baseline evaluation, Tom received three weeks of social skills training consisting of three 15–30 minute sessions per
FIGURE 7.1 Probe sessions during baseline, social skills treatment, and follow-up for training scenes for Tom. A multiple baseline analysis of ratio of eye contact while speaking to speech duration, number of words, number of requests, and overall assertiveness. (Figure 3, p. 190, from: Bornstein, M. R., Bellack, A. S., Hersen, M. (1977). Social-skills training for unassertive children: A multiple-baseline analysis. *Journal of Applied Behavior Analysis*, 10, 183–195. Copyright 1977 by Society for Experimental Analysis of Behavior. Reproduced by permission.)

week. These were applied sequentially and cumulatively over the three-week period. Throughout training, six role-played scenes were used to evaluate the effects of treatment. In addition, three scenes (on which the subject received no training) were used to assess generalization from trained to untrained scenes.

The results for training scenes appear in Figure 7.1. Examination of the graph indicates that institution of social skills training for ratio of eye contact to
speech duration resulted in marked changes in that behavior, but rates for number of words and number of requests remained constant. When social skills training was applied to number of words itself this behavior increased, while the rate for number of requests remained the same. Finally, when social skills training was directly applied to number of requests, marked changes were noted. Thus, the finding that each behavior changed markedly when and only when the treatment variable targeted each behavior provides evidence for the effectiveness of social skills training. Independence of the three behaviors and the absence of generalization effects from one behavior to the next facilitate interpretation of these data. On the other hand, had non-treated behaviors covaried following application of social skills training, unequivocal conclusions as to the controlling effects of the training could not have been reached without resorting to Kazdin and Kopel’s (1975) solution to withdraw and reinstate the treatment.

It is notable that while overall assertiveness was not treated directly, independent ratings evinced gradual improvement over the three-week period, with treatment gains for all behaviors maintained in follow-up (see bottom of Figure 7.1). Examination of data for the untreated generalization scenes indicates that similar results were obtained, confirming that transfer of training occurred from treated to untreated items. Indeed, the patterns of data for Figures 7.1 and 7.2 are remarkably consistent.

A more recent example of a multiple baseline design across behaviors is provided in a study by Nock (2002) that examines the effects of a treatment package including modeling, graduated exposure and contingency management on increasing food consumption in a four-year-old boy with food phobia. The boy in this case consumed only water and protein drinks since the age of seven months, when he choked on baby food with chunks of solid food in it. His parents had tried repeatedly to introduce various foods and drinks without success, and several previous psychodynamic and occupational therapy interventions also had failed.

In this study, the parents learned and applied the modeling, graduated exposure and contingency management strategies in weekly therapy sessions and applied them at home. The dependent variable was number of servings of four different types of food consumed by the boy each week: fluids other than water (e.g., juice), soft foods (e.g., pudding), hard foods (e.g., cookies), and chewy foods (e.g., chicken). The treatment strategies were directed at each of these four types of foods one at a time using a multiple baseline design.

The results of this study, presented in Figure 7.3, show that the boy’s consumption of fluids increased immediately after the treatment targeted this behavior (panel A), while the other three behaviors remained unchanged. Figure 7.3 shows a similar pattern for each of the other three behaviors (panels B-D), each increasing when and only when the treatment targeted that behavior, with increases in each behavior maintained at two- and six-month follow-up. Two additional aspects of this study are notable. First, there is a slight increase in consumption of chewy foods before the treatment targeted this behavior. This may have occurred because the parents began targeting this behavior before being
instructed to do so by the experimenter. Although certainly understandable clinically, this pattern slightly decreases the strength of the findings. Second, an additional problem behavior, vomiting, was designated for study during the second baseline and a line mapping the frequency of this behavior appears in panel B. More specifically, the boy began vomiting on occasion after consuming soft foods (e.g., pudding, bread). The parents and clinician suspected that this
FIGURE 7.3 Multiple-baseline design across four food categories for number of servings per week and vomiting. Note: In all four graphs, the lines marked with diamonds represent the number of servings of each food type consumed each week. The line marked with triangles in Graph B represents the number of vomiting episodes each week. (Figure 1, p. 221, from: Nock, M. K. (2002). A multiple-baseline evaluation of the treatment of food phobia in a young boy. *Journal of Behavior Therapy and Experimental Psychiatry, 33* (3–4), 217–225.)

behavior was somewhat under the boy’s behavioral control, and so instituted a treatment variable of time-out plus re-introduction of soft food (i.e., extinction) whenever vomiting occurred. The initiation of this intervention is represented by the second vertical line in panel B, and as shown, led to an immediate cessation of vomiting. This last feature provides one example of how multiple baseline designs can be flexibly modified during the course of the study to address newly emerging problem behaviors.
In another example of a multiple baseline design across behaviors, a psychological measure (erectile strength as assessed with a penile gauge) was used to determine efficacy of covert sensitization in the treatment of a 21-year-old married male, admitted for inpatient treatment of exhibitionism and obscene phone calling (Alford, Webster, & Sanders, 1980). History of exhibitionism began at age 16, and obscene phone calling had taken place over the previous year. During baseline assessment:

Audiotapes of both deviant and non-deviant sexual scenes were used to elicit arousal during physiological monitoring sessions. Deviant stimulus material included three tapes depicting various obscene phone calls... and three tapes of exhibitionism... Two non-deviant tapes... that depicted normal heterosexual behavior were also used... They consisted of verbal descriptions designed to closely parallel the patient's own sexual behavior and fantasy. (p. 17)

These included one taped description of intercourse with his wife and another with different sexual partners.

Covert sensitization sessions were conducted twice daily in the hospital at various locations. This treatment consisted of imaginarily pairing the deviant sexual approach (i.e., obscene phone calls, exhibitionism) with aversive stimuli such as suffocation, nausea, and arrest. Each session involved 20 pairings of the deviant scenarios with aversive imagery. Following baseline assessment, covert sensitization was first applied to obscene phone calling and then to exhibitionism. In addition to therapist-conducted treatment sessions, the patient was instructed to use covert imagery on his own initiative whenever he experienced deviant sexual urges.

Results from this study are presented in Figure 7.4. During baseline evaluation, penile tumescence in response to tapes of obscene phone calling and exhibitionism was quite high. Similarly, tumescence was above 75% in response to nondeviant tapes of sexual activity with females other than his wife, but only slightly higher than 25% in response to lovemaking with his wife.

Institution of covert sensitization for obscene phone calling resulted in marked diminution in penile responsivity to taped descriptions of that behavior, eventually resulting in only a negligible response. However, such treatment also appeared to affect changes in penile response to one of the exhibitionism tapes (Ex. 1), even though that behavior had not yet been specifically targeted. (We have here an instance where the baselines are not independent from one another.) However, when treatment subsequently was directed to exhibitionism itself, there was marked diminution in penile response to tapes Ex. 2 and Ex. 3 in addition to continued decreases to tape Ex. 1. During the course of treatment, penile responsivity to nondeviant heterosexual interactions remained high, increasing considerably with respect to lovemaking with the wife.

It is notable that "the patient was preloaded with 36 oz of beer 90 to 60 minutes prior to Assessments 10 and 11" (Alford et al., 1980, p. 19). This was incorporated because the subject claimed that alcohol had disinhibited deviant
sexuality. However, experimental data did not seem to confirm this. One-, two-, and 10-month follow-up assessments indicated that all gains were maintained, with the exception of decreased penile responsivity to taped descriptions of intercourse with the wife. In addition, 10-month collateral information from the subject’s wife, parents, and attorney, as well as police, court, and telephone company records revealed no incidents of sexual deviance.

Our illustration reveals a clinically successful intervention evaluated through the multiple baseline design. However, because of some correlation between the first two baselines (obscene phone calling and exhibitionism) the experimental control of the treatment over targeted behaviors remains unclear. Retrospectively, a more elegant experimental demonstration might have ensued if the experimenters had temporarily withdrawn treatment from the second baseline and then reinstated it (in B-A-B fashion), in order to show the specific controlling power of the aversive strategy. However, from the clinical standpoint, given the length of the disorder, it is most likely that the aversive intervention was responsible for ultimate change.
A variation on the multiple baseline design across behaviors in which changing criterion design was used in applying treatment to each behavior illustrated by Rassau and Arco (2003). The experimenters tested the effects of internet-based administration of cognitive behavior therapy (CBT) on school-related behaviors in an undergraduate student experiencing academic difficulties. The target behaviors in this study were hours of study per day, number of pages read per day, and quality of note-taking per day (scored using standard scoring criteria with greater points awarded for factors such as the use of subheadings, etc.). The treatment variable was administered 6 times per week in 45-minute sessions over a 57-day period. The treatment first targeted hot study per day, then number of pages read, then points for note-taking, with a target criterion for each behavior increasing over the course of the intervention.

The results of this study are presented in Figure 7.5. As shown, in the panel, hours of study increased in variability following the initiation of the treatment variable and performance far exceeded the criterion level on most days, a sharp increase several days into the treatment phase, which then immediately decreased. Similarly, there is a sharp increase in the second behavior, number of pages read, during the baseline period, which also then decreased. Both of these patterns suggest the treatment variable did not exert a controlling influence on these behaviors in the early course of the study, but rather the behavior increased due to the assignment due (marked with asterisks) at that time. Notable, however, that after the first assignment performance on both of the two behaviors closely matched the established criterion, and increased commensurate with increases in the criterion level, suggesting a controlling influence on the treatment variable. Similarly, the third behavior, points for note-taking increased substantially when the treatment variable was directed at this behavior, providing further support for the intervention.

A notable limitation of this study is the reliance on self-report of behavior over the course of the study, which introduces questions about the accuracy and validity of the data collected. Although the early stages of this study did not provide strong support for the effectiveness of this intervention, the measure of an extraneous factor that likely influenced these behaviors (i.e., assignment dates)—which was a nice feature of this study—and the subsequent increase in performance and confirmation to the established criterion levels increase strength of the findings.

An excellent, early example of a multiple baseline design across behaviors in which treatment was applied to entire group of subjects was provided by Totten, Guess, Garcia, and Baer (1970). Sixteen severely and profoundly retarded males served as subjects in an experiment designed to improve their meal behaviors through the use of time-out procedures. Several undesirable behaviors were selected as targets for study during preliminary observations. These included stealing (taking food from another resident's tray), fingers (eating with the fingers that should have been eaten with utensils), messy utensils (using a utensil to push food off the dish, spilling food), and what was ter
FIGURE 7.5 Participant’s study behaviors across baseline and on-line CBT. Assignment due dates are indicated by asterisks. (Figure 1, p. 380, from: Rassau, A., & Arco, L. (2003). Effects of chat-based on-line cognitive behavior therapy on study related behavior and anxiety. Behavioural and Cognitive Psychotherapy, 31, 377–381.)
pigging (eating spilled food from the floor, a tray, etc.; placing mouth directly over food without the use of a utensil). Observations of these behaviors were made 5 days per week during the noon and evening meals by using a time sampling procedure. Independent observations were also obtained as reliability checks. The treatment—time-out—involved removing the subject from the dining area for the remainder of a meal or for a designated time period contingent upon his evidencing undesirable mealtime behavior.

After 6 days of baseline assessment, the treatment was applied to stealing. Time-out contingencies for fingers, messy utensils, and pigging were then applied in sequence, each time maintaining the contingency in force for the previously treated behavior. During the application of time-out for fingers the contingency involved time-out from the entire meal for 11 subjects, but only 15 seconds time-out for five of the subjects. This differentiation was made in response to the nursing staff's concerns that a complete time-out contingency for the five subjects might jeopardize their health. Time-out procedures for messy utensils and pigging were limited to 15 seconds per infraction for all 16 subjects.

The results of this study are presented in Figure 7.6. Examination of the graph indicates that when time-out was applied to stealing and fingers, rates for these behaviors decreased. However, application of time-out to fingers also resulted in a concurrent increase in the rate for messy utensils. But subsequent application of time-out for messy utensils affected a decrease in rate for that behavior. Finally, application of time-out for pigging proved successful in reducing its rate.

Independence of the target behaviors was observed, with the exception of messy utensils, which increased in rate when the time-out contingency was applied to fingers. Although group data for the 16 subjects were presented, it would have been desirable if the authors had presented data for individual subjects. Unfortunately, the time-sampling procedure used by Barton et al. (1970) precluded obtaining such information. However, this factor should not overshadow the clinical and social significance of this study, in that (1) mealtime behaviors improved significantly; (2) a result of improved mealtime behaviors was a concomitant improvement in staff morale, facilitating more favorable interactions with the subjects; and (3) staff beyond those involved with the study subjects were sufficiently impressed with the results of this study to begin to implement similar mealtime programs for their own residents.

It is notable that studies demonstrating behavior change for an entire group of subjects can be very useful; however, in co-opting behavior analytic procedures, one must be careful to present as much individual data as possible. Whenever group data are presented, all of the problems of averaging described in previous chapters apply to the obtained data. That is, some subjects could show the very steady changes apparent in the group data across measurement sessions, whereas others might demonstrate opposite or cyclic patterns. Presenting data in this way does not allow one the option of examining sources of variability (e.g., treatment moderators) where it might be important. Moreover, when it is not clear how many individuals changed in clinically significant ways, estimates of the replicability of the procedures used across individuals and iden-
FIGURE 7.6 Concurrent group rates of Stealing, Fingers, Utensils, and Pigging behaviors, and the sum of Stealing, Fingers, and Pigging (Total Disgusting Behaviors) through the baseline and experimental phases of the study. (Figure 1, p. 80, from: Barton, E. S., Guess, D., Garcia, E., & Baer, D. M. [1970]. Improvement of retardates' mealtime behaviors by time-out procedures using multiple baseline techniques. *Journal of Applied Behavior Analysis*, 3, 77–84. Copyright 1970 by Society for Experimental Analysis of Behavior, Inc. Reproduced by permission.)

The differentiation of individual predictors of success and failure are not possible (see chapter 10). In summary, when using group designs the presentation of as much individual data as possible is strongly recommended.

In an interesting solution to the problem of averaging when a number of subjects are treated simultaneously, Kelly (1980) argued for the use of a design referred to as the simultaneous replication design, which can be used within a
multiple baseline format. The specific example cited involves application of social skills training in group format to 6 subjects for three components of social skill on a time-lagged basis. However, although applied on a group basis, behavioral assessment of each subject follows each group session. Thus individual data for each treated subject are available and can be plotted individually (see Figure 10.6). As noted by Kelly (1980):

The use of this group multiple baseline-simultaneous replication design is particularly useful in applied clinical settings for several reasons. First, it eliminates the need for elaborate and/or untreated control groups to establish group treatment effects and rule out many alternative hypotheses which cannot be adequately controlled by other one group designs. Second, by analyzing the social skills behavior change effects of a group treatment procedure, it is possible to demonstrate more compellingly cost- or time-effectiveness than if each subject had been laboriously handled as an individually treated case study using single subject procedures. Because subjects all received the same group training but are individually evaluated after each group, it is possible to examine "within subject" response to group treatment with greater specificity than in "between groups" designs. Since data for each subject in the training group is individually measured and graphed, each subject also serves as a simultaneous replication for the training procedure and provides important information on the generality (or specificity) of the treatment. (pp. 206–207)

(See also section 10.2 for a discussion of issues arising from this strategy relevant to replication.)

Although the multiple baseline design is frequently used in clinical research when withdrawal of treatment is considered to be detrimental to the patient, on occasion withdrawal procedures have been instituted following the sequential administration of treatment to target behaviors (e.g., Russo & Koegel, 1977). If treatment is reintroduced after a withdrawal, a powerful demonstration of its controlling effects can be documented.

A recent example of a multiple baseline design across behaviors with an embedded withdrawal (A-B-A-B) design is provided in a study by Clayton, Helms, and Simpson (2006). In this study, the experimenters tested the effectiveness of active prompting (i.e., posting street signs) on decreasing cell phone use and increasing seat belt use among motorists passing through a four-way intersection. The primary dependent variables in this study were cell phone use and cell phone hang-ups while driving through the intersection, as well as seat belt use and buckling up while driving through the intersection. Using a multiple baseline design across behaviors, the experimenters tested the effectiveness of first posting a sign asking motorists to hang up their cell phones (i.e., “Please Hang Up, I Care”), and subsequently to buckle their seat belts (i.e., “Please Buckle Up, I Care”). Each of these behaviors was recorded by independent observers and scored with strong inter-observer agreement (>95% for all behaviors).

The results of this study are presented in Figure 7.7. As shown, cell phone use occurred at a fairly stable rate of 6% during the baseline phase, with 0% of
motorists actually hanging up their cell phone when they passed through the intersection. The percentage of motorists hanging up their cell phone increased following posting of the cell phone hang up sign, although the overall rate of cell phone use remained unchanged. The experimenters then instituted a withdrawal of the hang up sign, during which the rate of hang ups returned to 0%, followed by a subsequent reinstatement of the hang up sign, which again increased the rate of hang ups. The hang up sign was removed in the final phase, again returning the rate of hang ups to 0%. Thus, the experimenters used an A-B-A-B-A design for the cell phone behaviors to demonstrate the controlling effect of the intervention on cell phone hang ups, with minimal impact on overall cell phone use.

Notably, the use of seat belts and rate of buckling up were not affected during the phases involving posting of the hang up sign. Instead, both buckling up and overall seat belt use increased when and only when the buckle up sign was posted.
(bottom panel of Figure 7.7). The experimenters again used an A-B-A-B-A design for seat belt use and buckling up to demonstrate the controlling influence of active prompting (i.e., posting street signs) on these behaviors.

In summary, this study by Clayton, Helms, and Simpson (2006) illustrates the use of the multiple baseline design across behaviors, demonstrating clear independence of target behaviors. Sequential application of active prompting showed the controlling effects of the intervention. Additional manipulations (withdrawal and reintroduction of the intervention) further confirmed the controlling effects of the treatment.

In our final example of a multiple baseline design across behaviors, the effect of booster sessions on subject behavior (i.e., rather than a full withdrawal or reversal design) subsequent to deterioration during follow-up was tested (Van Hasselt, Hersen, Kazdin, Simon, & Mastantuono, 1983). The subject in this case was a blind female child attending a special school for the blind. Baseline assessment of social skills through role playing revealed deficiencies in posture and gaze, a hostile tone of voice, inability to make requests for new behavior, and a general lack of social skills (see Figure 7.8).

The sequential and cumulative application of social skills training resulted in marked improvements in role-played performance, thus documenting the controlling effects of the treatment. However, data for the four-week post-treatment follow-up revealed a decrement for two of the four specific behaviors targeted: gaze and requests for new behavior. Examination of Figure 7.8 shows that treatment provided in booster sessions for those behaviors resulted in a renewed improvement, extending through the eight- and 10-week follow-up assessments. Thus, our multiple baseline analysis permitted a clear assessment of which behaviors were maintained after treatment in addition to those that required booster treatment. This study again highlights the specificity and flexibility afforded by the use of single-case experimental designs.

**Multiple baseline design across subjects**

Our first example of the multiple baseline strategy across subjects is a fairly basic application of this design taken from the child literature. Miltenberger and colleagues (Miltenberger et al., 2005) tested the effectiveness of teaching firearm safety skills on the prevention of playing with guns found in the home or school setting. Participants in this study were 10 children recruited from a local preschool program. The dependent variable in this study was demonstration of firearm safety skills scored on a four-point scale (e.g., 0 = touches gun, 1 = doesn’t touch gun, 2 = leaves area, 3 = tells adult about gun) in response to finding a gun at home or at school. The guns were unloaded and placed in designated locations by the experimenters where the children were expected to find them, and the children’s behavior was observed and rated by hidden video-camera and/or an experimenter out of sight from the child. Inter-observer agreement on the dependent variable was 100%. The treatment variable was firearm
FIGURE 7.8  Probe sessions during baseline, social skills treatment, follow-up, and booster assessments for training scenes for S1. A multiple baseline analysis of posture, gaze, hostile tone, requests for new behavior, and overall social skill. (Figure 1, p. 201, from: Van Hasselt, V. B., Hersen, M., Kazdin, A. E., Simon, J., & Mastantuono, A. K. [1983]. Social skills training for blind adolescents. Journal of Visual Impairment and Blindness, 75, 199-203. Copyright 1983. Reproduced by permission.)
safety training that used instructions, modeling, rehearsal, and feedback to
the children not to touch firearms and to inform an adult if a firearm was
found. Over the course of the study, the experimenters placed firearms in
dedicated places at the children’s school and home and observed the effects of
treatment, which was administered in a multiple baseline fashion across chil-

The results of this study are presented in Figure 7.9. The first three chil-
dren were observed over two baseline sessions and all touched the gun in each
(i.e., scored a 0 on the safety skills scale). After introduction of the treat-
mantime and consistently going to 3 for Alan and Jessie, but with a one-
 latency for Steph (who notably left the pre-school program after five ses-
next three children were observed over three baseline sessions before receiv-
treatment. Notably, they showed no improvement when the first three chil-
d received the intervention, but did show an improvement when the treat-

FIGURE 7.9 (A) Rating scale scores for 5 participants during baseline and in situ-
ing phases. Circles are day-care assessments, triangles are home assessments, and
are dyad assessments. The last home or day-care data point for all participants
Steph is a 3-month follow-up assessment. (B) Rating scale scores for 5 participants
baseline and in situ training phases. Circles are day-care assessments, triangles are
assessments, and squares are dyad assessments. The last home or day-care data po
all participants is a 3-month follow-up assessment. (Figure 1, p. 397, from: Milten
g R. G., Gatheridge, B. J., Scatterlond, M., Egemo-Helm, K. R., Johnson, B. M., Jast
et al. (2005). Teaching safety skills to children to prevent gun play: An evaluation of
directed at them, with Kari and Bob showing an immediate and sustained improvement to 3, but with a one-session latency for Donald. The next three children received the treatment after a four-session baseline and the last child after a five-session baseline. The results of this study by Miltenberger and colleagues (2005) clearly demonstrate that the children’s behavior improved markedly when and only when they received the treatment, providing evidence for the effectiveness of this intervention. A three-month follow-up assessment was conducted for five of the 10 children and in each case maintenance of safety skills was demonstrated.

Miller and Kelley (1994) used an interesting variation of a multiple baseline strategy across subjects in their assessment of goal setting and contingency contracting on improving homework performance among four children (9–11 years old) who were at risk for academic problems due to poor homework performance. Among the dependent variables measured was accuracy of completed work (percentage of correct solutions on homework problems) and percentage of time spent on-task and off-task while engaged in homework after school each day. Accuracy was calculated by the child’s parent each day, and on-task and off-task behaviors were carefully defined and observed by trained graduate and undergraduate raters with strong inter-observer agreement (97% for on-task and 87% for off-task behaviors). Treatment consisted of the experimenter teaching the child’s parent goal setting and contingency contracting for homework assignments, and the parent managing the homework intervention with the child (see Miller & Kelley, 1994).

Data on homework accuracy are presented in Figure 7.10. As shown, improved performance only occurred when the treatment was directly applied to each child, thus documenting the controlling effects of treatment. Data clearly indicate that the four baselines were independent of one another. Moreover, additional confirmation of the controlling effects of treatment was noted when introduction of the treatment resulted in improved performance, followed by deterioration when withdrawn and renewed improvement when reinstated. Thus, for each child we have an A-B-A-B demonstration, but carried out sequentially and cumulatively across the four. In short, the study by Miller and Kelley (1994) is an excellent example of the combined use of the A-B-A-B design in multiple baseline fashion across subjects.

Recent technological advances have introduced exciting methodological innovations that can be incorporated into single-case experimental designs, and several recent examples have done so within the context of multiple baseline designs across subjects. As one recent example, Dallery and Glenn (2005) tested the effects of a new internet-based voucher reinforcement program on abstinence from cigarette smoking. Voucher reinforcement programs have been shown to be effective at decreasing smoking; however, such programs often rely on breath carbon monoxide (CO) output as a measure of smoking abstinence (with CO output < 4 parts per million [ppm] indicating abstinence). This is problematic because it requires in-person measurement in order to verify the authenticity of the CO sample.
FIGURE 7.10 Percentage of homework completed accurately during baseline and treatment conditions across subjects. Sessions correspond to sequential school days (i.e., Monday through Thursdays) on which subjects were assigned homework. Data were not collected on days on which homework was not assigned. (Figure 2, p. 80, from: Miller, D. L., & Kelley, M. L. (1994). The use of goal setting and contingency contracting for improving children's homework performance. *Journal of Applied Behavior Analysis, 27* (1), 73–84.)
To address this problem and thus increase the potential application of such protocols, Dallery and Glenn (2005) developed an internet-based protocol in which subjects submitted CO results via the internet (i.e., using a home-based monitor, web-cam, and e-mailing results to the experimenters) and were awarded vouchers via the internet as well (i.e., decreased CO levels were rewarded with monetary vouchers to an on-line shopping mall: www.amazon.com). Subjects in this study were four healthy adults with an 8–30 year history of smoking, who were currently smoking 20–40 cigarettes per day.

As displayed in Figure 7.11, each of the four subjects evidenced levels of CO consistent with cigarette smoking during the baseline phase. Next, a shaping program was administered sequentially across subjects in which each subject earned vouchers for increasingly lower levels of CO. As shown, in each case CO levels decreased during this phase. In the next phase, subjects received vouchers for smoking abstinence (i.e., CO < 4 ppm) and results indicate that three of the four subjects regularly met this criterion. During the next phase the reinforcement schedule was thinned, and finally there was a return to baseline.

Several aspects of these results are notable. First, there was a decrease in CO levels for the second subject during the baseline phase. This weakens the inferences that can be drawn about the controlling influence of the intervention in this case; however, the replication of decreased CO associated with presentation of the intervention across the three other participants provide much stronger evidence for the intervention than would be possible in a single A-B design. Second, the first subject showed a significant increase in smoking during the third (abstinence) phase. This pattern also weakens inferences that can be drawn about the effectiveness of the treatment; however, here the measurement of an extraneous factor that may have contributed to this increase—namely, the experience of significant family-related stress—provides some increased understanding about this increase. Third, although it would be nice clinically to see long-term maintenance of treatment effects during the final phase of the study, the increase in CO levels for the first and third subject as well as increased variability for the fourth, while not supporting the durability of behavior change, do provide some evidence for the controlling influence of the treatment variable. Overall, these results generally support the effectiveness of this internet-based treatment program—which has been further examined and compared with alternative treatment strategies by these authors (Glenn & Dallery, 2007), provide an illustration of the benefits of examining results at the level of each individual, highlight the potential usefulness of single-case designs for demonstrating the feasibility of a newly developed intervention (see also Choate, Pincus, Eyberg, & Barlow, 2005; Moras, Telfer, & Barlow, 1993), and also demonstrate the innovations possible in single-case experiments using recent technological advances.

Another example of a multiple baseline design across subjects that incorporates recent technological advances is provided by Ingvarsson and Hanley (2006). In this study, the experimenters tested the effectiveness of a fully computerized assessment and training tool designed to teach preschool teachers the names and faces of parents so that the teachers could greet the parents by name
FIGURE 7.11 CO (ppm) values for all participants across baseline, shaping, abstinence induction, thinning, and return to baseline. The dashed horizontal line marks a reading of 4 ppm (4 ppm or below is considered abstinent). The open circles in the shaping condition indicate when the shaping criteria were met. The asterisk indicates the day when C0037 started to experience significant family-related stress. Note that Participant H0040 has a different y axis. (Figure 1, p. 354, from: Dallery, J., & Glen, I. M. (2005) Effects of an Internet-based voucher reinforcement program for smoking abstinence: A feasibility study. *Journal of Applied Behavior Analysis, 38* (3), 349-357.)
when they dropped their children off at preschool each morning, thus improving parent-teacher communication. Subjects in this study were four student teachers (1 male, 3 female) working at a local preschool. The primary dependent variable was percentage of opportunities during which parents were greeted by name by each teacher (inter-observer agreement = 93%). The computerized training program was administered sequentially to each teacher using a multiple baseline design across subjects.

The results from this study are presented in Figure 7.12. As shown, although each of the four teachers had been instructed to greet parents by name each morning, they rarely did so during the baseline phase. For three of four teachers (1, 2, and 4) the percentage of greetings using parents’ names increased markedly when and only when the treatment variable was applied to that teacher. The results from these first two phases suggest the treatment was effective for three of four teachers. In the case of teacher 3, due to lack of a response to the treatment variable, a feedback phase was quickly introduced during which the teacher’s supervisor reviewed the previous data for that teacher, encouraged her to use parents’ names, and praised that teacher for any instances of parental name use. These procedures were repeated daily for the duration of the feedback phase. The results reveal that addition of this feedback phase led to an increase in name use by teacher 3. The feedback phase was subsequently used to enhance treatment effects for teacher 1 and 4 also, whereas teacher 2 performed at a high level throughout the study and so no feedback phase was added. This study by Ingvarsson and Hanley (2006) provides another example of the flexibility possible using single-case designs and offers an additional illustration of how recent technological advances can enhance such designs, as well as improve behavioral interventions.

A classic example of a multiple baseline design across groups of subjects was provided by Epstein et al. (1981). The effects of a behavioral treatment program to increase the percentage of negative urine tests among 19 children with insulin dependent diabetes. Treatment aimed to decrease intake of simple sugars and saturated fats, decrease stress, increase exercise, and adjust insulin intake. Parents were taught to use praise and token economic techniques to reinforce improvements in the children’s self-regulating behavior. When treatment began, 10 of the children (ages 8 to 12) were self-administering their insulin; the remaining 9 were receiving shots from their parents.

The primary dependent measure involved a biochemical determination of any glucose in the urine. As noted by Epstein et al. (1981), this “…suggests that greater than normal glucose concentrations are present in the blood, and the renal threshold has been exceeded” (p. 367). Such testing was carried out on a daily basis during baseline, treatment, and follow-up. The 19 families were assigned on a random basis to one of three groups, with treatment begun under time-lagged conditions two, four, or six weeks after initiation of the 12-week program.

Examination of Figure 7.13 indicates that percentage of negative urines was relatively low for each of the three groups during baseline. Institution of treatment resulted in marked improvements in percentage of negative urines, indicating the
FIGURE 7.12 Teachers' use of parents' names during classroom observations. Each panel represents data from an individual teacher. Each data point represents the percentage of opportunities in which a teacher greeted a parent by name. A data point is not presented for days on which a teacher had no opportunities to greet parents. An asterisk denotes that a refresher training session was implemented. (Figure 2, p. 209, from: Ingvarsson, E. T., & Hanley, G. P. (2006). An evaluation of computer-based programmed instruction for promoting teachers' greeting of parents by name. Journal of Applied Behavior Analysis, 39 (2), 203–214.)
controlling effects of the strategy. Moreover, it appears that these gains were maintained post-treatment, as indicated by the follow-up assessment at 22 weeks.

In summary, Epstein et al. (1981) presented a powerful demonstration of the effects of a behavioral treatment over a biochemical dependent measure (that has serious health implications). From a design standpoint, this study is an excellent illustration of the multiple baseline strategy across small groups of subjects, suggesting how the particular experimental strategy can be used to evaluate treatments in the area of behavioral medicine. However, from a design standpoint, the aforementioned cautions associated with averaging data certainly apply.
Multiple baseline across settings

Our first example of a multiple baseline strategy across settings involves a recent study by Kay, Harchik and Luiselli (2006) that tests the effectiveness of a multi-component treatment package designed to reduce drooling in a 17-year-old high school student, George, diagnosed with autism and mental retardation. The boy had a long history of drooling since childhood that was causing problems at school (e.g., presence of saliva on school materials) and with peers (e.g., other students avoided the subject and complained about his drooling).

During baseline, George's drooling was monitored by his assigned academic aide, who counted the pools of saliva (i.e., saliva present on the surfaces his immediate area measuring ≥1 inch in diameter) present in three different school locations each day: his classroom (three hours per day), community vocational site (two hours per day), and cooking class (one hour per day). Inter-observer agreement was 91–96% across all three sites. The aide wiped each saliva pool immediately after it was recorded. As shown in Figure 7.14, during baseline George produced an average of 6–14 pools of saliva per hour in the classroom at community settings, and between 8–16 pools of saliva per hour in cooking class.

Following baseline measurement and formulation of hypotheses about factors maintaining George's drooling (i.e., skills deficit), George was administered an intervention including instructions, skill acquisition (i.e., swallowing and wiping his mouth), and differential reinforcement of other behavior (DRO; i.e., small edible treats were provided for dry mouth on periodic checks by the aide). The treatment was applied sequentially across settings, focusing initially on George drooling in the classroom, then in the community setting, then in cooking class.

As shown in Figure 7.14, George's drooling in the classroom decreased substantially when the intervention was applied to that setting, but there was no change in drooling in the other two settings. However, drooling in the community and cooking class settings decreased when and only when the intervention was targeted at these settings. Behavior change was maintained when the period checks and associated reinforcement strategies were discontinued, demonstrating maintenance of behavior change.

A multiple baseline design across settings with an embedded withdraw strategy (A-B-A-B) is demonstrated in a study by Singh, Dawson, and Grego (1980) that tested treatment of hyperventilation in a 17-year-old girl diagnosed with profound mental retardation. The subject in this case suffered from epilepsy (controlled pharmacologically) and had a 6-year history of hyperventilation. Prior attempts to treat her symptoms, which included deep, heavy breathing, accompanied by a grunting noise and up-and-down head movements, were unsuccessful. These symptoms were observed in four separate settings at the residential unit in the state facility in which she lived: classroom, dining room, bathroom, and day room. Data were recorded in 10-second intervals throughout 30-minute sessions.

Baseline data were obtained for five sessions in the classroom, 10 in the dining room, 15 in the bathroom, and 20 in the dayroom. Then, under time-lagged conditions, treatment was introduced, removed, and reintroduced in each
FIGURE 7.14 Number of saliva pools (average per hour) recorded during baseline and intervention phases. (Figure 1, p. 26, from: Kay, S., Harchik, A. E., & Luiselli, J. K. (2006). Elimination of drooling by an adolescent student with autism attending public high school. *Journal of Positive Behavior Interventions*, 8 (1), 24–28.)
setting—yielding an A-B-A-B design in each setting. Treatment consisted of the application of response-contingent aromatic ammonia held under her nose for three seconds whenever an instance of hyperventilation was observed. Finally, during the eight weeks of the generalization phase, nurses were requested to carry out the punishment procedure on an eight-hour-per-day basis. This is in contrast to original treatment that was carried out for only four 30-minute sessions per day.

Results of this study are presented in Figure 7.15. These data clearly indicate the controlling effects of the treatment, both in terms of its initial application on

---

**FIGURE 7.15** Number of hyperventilation responses per minute and condition means across experimental phases and settings. (Figure 1, p. 565, from: Singh, N. N., Dawson, J. H., & Gregory, P. R. [1980]. Suppression of chronic hyperventilation using response-contingent dramatic ammonia. *Behavior Therapy, 11*, 561–566. Copyright 1980 by Association for Advancement of Behavior Therapy. Reproduced by permission.)
time-lagged basis (baselines were independent) and when it was removed and reintroduced simultaneously in all four settings. Rate of hyperventilation episodes increased dramatically when the punishment contingency was removed in the second baseline and decreased to near zero levels when it was reintroduced. Moreover, the positive effects of treatment were prolonged and enhanced as a result of the more extensive punishment approach followed in the generalization phase.

Soenksen and Alper (2006) present an interesting application of a multiple baseline design across settings in which they target two behaviors within each setting. They tested a social skills training intervention presented in a story format to a five year-old boy, named TJ, with hyperlexia—a condition characterized by an ability to read far beyond one’s comprehension, often accompanied with social deficits including poor eye contact and poor social communication skills. The treatment consisted of reading TJ and several of his classmates a story about a person who communicates with his friends by looking at them and saying their names. This story was read to TJ and his classmates each day, first immediately before recess, then before choice time (during which students could choose from various activities in the classroom), then before math time. The dependent variables in this study were TJ’s daily frequency of (1) saying a peer’s name and (2) looking at a peer’s face when attempting to speak to that person (inter-observer agreement = 94.3%).

The results of this study are presented in Figure 7.16. As shown, TJ very rarely engaged in either of the target behaviors during baseline. When the treatment variable was applied to recess, TJ’s frequency of both saying peers’ names and making eye-contact with them increased to match that of his peers’ performance (represented by the solid horizontal line), and remained at that level during the maintenance phase. The frequency of these two behaviors also increased for TJ when the treatment variable was applied to choice time, and subsequently to math time, and in both instances TJ’s behavior increased to match that of his peers.

Overall, the results of this study by Soenksen and Alper (2006) suggest this treatment was effective at increasing these two social skill behaviors across these three different settings. Several aspects of these findings warrant comment. First, although the data generally support independence of behavior across these three settings, the brief increases during the baseline phase of choice time and math time slightly weaken the case for independence and for the controlling influence of the intervention. However, given the brevity of the increases one could conclude that the intervention is supported overall. Second, a positive aspect of this study is the maintenance of behavior change, which extends out to follow-up for choice time and math time and highlights the durability of the improvements in social skills associated with this treatment.

A particularly socially relevant example of a multiple baseline design across settings that also measured two related behaviors but in addition included an embedded withdrawal (A-B-A) design was provided by Ludwig, Gray, and Rowell (1998). The experimenters tested the effects of moving recycling receptacles (and associated recycling signs) from hallways to classrooms (where beverages
FIGURE 7.16 Frequency of saying a peer’s name and looking at a peer’s face across three different settings. (Figure 1, p. 41, from: Soenksen, P., & Alper, S. (2006). Teaching a young child to appropriately gain attention of peers using a social story intervention. Focus on Autism and Other Developmental Disabilities, 21 (1), 36-44.)
were typically consumed) in academic classrooms in two different buildings on: (1) percentage of recyclable aluminum cans placed in recycling receptacles and (2) percentage of cans thrown in trash containers—and thus not recycled.

The results of this study are presented in Figure 7.17. As shown, during baseline cans were more often thrown away in trash containers than in recycling receptacles, and this was true across both buildings studied. After 10 days of baseline measurement, the treatment variable was administered in Building A, at which time the percentages of cans in each type of container switched, with the majority of cans now being thrown away in the recycling receptacles. There was no such change in Building B during this time; however, a similar change was observed.

**FIGURE 7.17** The percentage of cans counted in recycling receptacles and trash containers per day in Buildings A and B. Daily observation sessions were conducted Monday through Friday. Filled circles represent cans counted in recycling receptacles, and open circles represent cans counted in trash containers. Vertical dashed lines represent the phase changes. Thick dashed horizontal lines represent the mean number of cans counted in recycling receptacles for a given phase. Thin dashed and dotted horizontal lines represent the mean number of cans counted in trash containers for a given phase. (Figure 1, p. 685, from: Ludwig, T. D., Gray, T. W., & Rowell, A. (1998). Increasing recycling in academic buildings: A systematic replication. *Journal of Applied Behavior Analysis, 31* (4), 683–686.)
in Building B when the treatment variable was administered in that setting (after 40 day hiatus in both settings due to winter break), although with greater variability observed during both baseline and treatment phases in that setting.

Notably, there was a withdrawal of the treatment variable after 18 days in Building A and after 66 days in Building B and in both settings this resulted in return to baseline levels of behavior, providing strong support for the controllin influence of the treatment variable. The A-B-A confirmation of the controllin power of the intervention adds substantially to documentation of the time-lagge contingency. Thus we have a very powerful demonstration of this treatment in multiple baseline design across settings that incorporates A-B-A withdrawal features.

### 7.3. VARIATIONS OF MULTIPLE BASELINE DESIGNS

**Nonconcurrent multiple baseline design**

As noted in section 7.2, in the multiple baseline design across subjects, each individual targeted for treatment is exposed to the same environment. Treatment delayed for each successive subject in time-lagged fashion because of the increased length of baselines required for each. The functional relationship between treatment and target behavior can be determined only when such treatment is applied to each subject in succession. Thus, since subjects are simultaneously available for assessment and treatment, this design is able to control for history, a possible experimental contaminant.

However, there are times when one is unable to obtain concurrent observations for several subjects, in that they may be available only in succession (e.g. for less frequently observed conditions such as food phobia or hyperlexia). Following strictures of the multiple baseline design across subjects, this design ordinarily would not be considered appropriate under these circumstances. However, Watson and Workman (1981) have proposed an alternative—**nonconcurrent multiple baseline design across individuals**:

In this design, the researcher initially determines the length of each of several baseline designs (e.g., 5, 10, 15 days). When a given subject becomes available (e.g. a client referred who has the target behavior of interest, and is amenable to the use of a specific treatment of interest), s/he is randomly assigned to one of the predetermined baseline lengths. Baseline observations are then carried out; assuming the responding has reached acceptable stability criteria, treatment is implemented at the pre-determined point in time. Observations are continued through the treatment phase, as in a simple A-B design. Subjects who fail to display stable responding would be dropped from the formal investigation; however, the eventual reaction to treatment might serve as useful replication data.

The logic of this variation is graphically portrayed in Figure 7.18. Of course, the major problem with this strategy is that the control for histot
(i.e., the ability to assess subjects concurrently) is greatly diminished. Therefore, we view this approach as less desirable than the standard multiple baseline design across subjects and suggest that it should be employed only when the standard approach is not feasible. Moreover, under such circumstances, an increased number of replications (i.e., number of subjects so treated) would enhance the confidence one has in the results. Of course, in the case of rarer disorders this may not be possible. For instance, Frank, Spirito, Stark and Owens-Stively (1997) report on the use of a nonconcurrent multiple baseline design across subjects to demonstrate the effectiveness of scheduled awakenings to eliminate childhood sleepwalking in three young children. The nonconcurrent multiple baseline design also has been proposed as an especially useful strategy in educational and other settings when more rigorous designs may not be feasible (Harvey, May, & Kennedy, 2004).

**Multiple-probe technique**

In all of the multiple baseline designs described above, baseline measurement has been continuous. However, there are situations in which repeated measurements
will result in reactivity to assessment, in which change occurs simply as a result of repetition of the assessment (Kazdin, 2003). When treatment is subsequently introduced under these circumstances, changes may not be detected or may be masked due to alterations in the baseline as a function of reactivity. In addition, there are instances when continuous measurement is not feasible and when on the basis of prior experimentation an "a priori assumption of stability can be made" (Horner & Baer, 1978, p. 193). This being the case, instead of having 6, 9, and 12 assessments in three successive baselines, for example, these can be more interspersed, resulting in perhaps two, three, and four measurement points. An example of this approach is presented in Figure 7.19. Hypothetical probes in our example are represented by closed triangles, whereas actual reported data appear as open circles.

In commenting on this graph, Horner and Baer (1978) argued that:

The multiple-probe technique, with probes every five days, would have provided one, two, three, and five probe sessions to establish baselines across the four subjects. The multiple-probe technique probably could have provided a stable baseline with five or fewer probe sessions for the subject who had 15 days of continuous baseline in the original study. The use of the multiple-probe procedure might have precluded the increase in irrelevant and competing behaviors by this subject because such behavior began to increase after the tenth baseline session. (p. 195)

It should be noted that several different researchers have applied this variant of baseline assessment in the multiple baseline design over the years (Baer & Guess, 1971; Schumaker & Sherman, 1970; Striefel, Bryan, & Aikins, 1974; Striefel & Wetherby, 1973). In each of these studies the design used was the multiple baseline design across behaviors. But, as in Figure 7.19, it could be across subjects, and it could also be used across settings.

If concerns about reactivity are the primary reason for using this variant, the probe technique should be continued when treatment is instituted. However, if feasibility is questionable in baseline or if an a priori assumption of baseline stability can be made, more frequent measurements during treatment may be desirable.

An example of a multiple baseline design across settings that used the multiple probe technique is provided by De La Paz (1999), in which the experimenter tested the effects of specific instructional procedures on 22 students' ability to write longer, more complete, and higher quality essays. Because the dependent variables in this study were derived from student essays, and it was not feasible for students to provide essays on a continuous basis, a multiple probe approach was employed in which students provided essays several times per week over the course of the study. Subjects recruited from three different classroom settings completed a minimum of six essays during the baseline phase, three of which were required to be completed within the week prior to administration of the treatment variable (i.e., the instructional training). All subjects then completed three
FIGURE 7.19  Number of toothbrushing steps conforming to the definition of a correct response across 4 subjects. (Figure 2, p. 194, from: Horner, R. D., & Baer, D. M. [1978]. Multiple-probe technique: A variation of the multiple baseline. Journal of Applied Behavior Analysis, 11, 189–196. Copyright 1978 by Society for Experimental Analysis of Behavior. Reproduced by permission.)
essays within one week after receiving the treatment variable, and students from the first two classrooms completed a follow-up essay four weeks after the treatment variable. No follow-up was conducted for the third classroom because school ended within one week after the post-instruction essays were completed.

The results from this study regarding essay quality (coded by teachers blind to the study design and hypotheses) are presented in Figure 7.20. Data

from each classroom are averaged across subjects; however, subjects in each classroom were classified into subgroups based on cognitive ability in order to examine whether effects of the treatment variable differ among those with a learning disability relative to those classified as low-achieving, average-achieving, and high-achieving. As shown, writing quality was fairly stable for all subgroups in all three classroom settings across the baseline phase. The treatment variable was administered in the first classroom after six probes, at which time writing quality increased markedly for all subgroups in this classroom, with no changes in the other classrooms. Similar improvements in writing quality were observed when and only when the treatment variable was administered to subjects in the second and third classrooms, supporting the effectiveness of the instructional training session. The improvements observed were maintained at follow-up.

The probe technique can be quite useful in a number of instances. However, as in the case of the nonconcurrent multiple baseline design, it should not be employed as a substitute for continuous measurement when that is feasible, as data accrued from use of probe measures are suggestive rather than confirmatory of the controlling effects of a given treatment.

7.4. ISSUES IN DRUG EVALUATIONS

With the exception of the multiple baseline design across subjects, the multiple baseline strategies are often not well suited for the evaluation of pharmacological agents on behavior. For example, at this point in time it is quite unlikely that any drugs will have the specificity of action to, within the same person, sequentially change different behaviors or even to sequentially change the same behavior across different settings. In this regard, psychological interventions are currently far superior to pharmacological interventions.

However, it would be possible to apply different drugs under time-lagged conditions to separate behaviors following baseline placebo administrations for each. However, such a design would involve a radical departure from the basic assumptions underlying the multiple baseline strategy across behaviors and would only permit very tentative conclusions based on separate A1-B designs for each targeted behavior. In addition, the possible interactive effects of drugs might obfuscate specific results. Indeed, the interaction design (see chapter 6) is better suited for evaluation of combined effects of therapeutic strategies.

Overall, of the three multiple baseline design strategies currently in use, the multiple baseline design across subjects is most readily adaptable to drug evaluations. The application of such designs could be most useful when withdrawal procedures (return to A1—baseline placebo) are unwarranted for either ethical or clinical considerations, such as in the case of suicidal or self-injurious behaviors. Using this type of strategy across matched subjects, baseline administration of a placebo (A1) could be followed by the sequential administration
(under time-lagged conditions) of an active drug (B). Thus a series of A₁-E (quasi-experimental) designs would result, with inferences made in accordance with changes observed when the B (drug) condition was applied.

Many other design options are possible in the application of the multiple baseline design across subjects when evaluating pharmacological effects. For example, Breuning and colleagues (1980) used a multiple baseline design across subjects (small groups) to evaluate the effects of drug, placebo, and response cost conditions. This yields a B (drug), A₁ (placebo), C (response cost) design. Subjects were hospitalized individuals diagnosed with mental retardation who displayed inappropriate behaviors. After three weeks on active neuroleptic drugs, Subjects 12, 15, and 16 were switched to placebo for 10 weeks. After six weeks on active neuroleptic drugs, Subjects 13 and 19 were switched to placebo for seven weeks. Finally, after nine weeks on active neuroleptic drugs, Subjects 14 and 17 were switched to placebo for seven weeks.

The results of this study are presented in Figure 7.21. Examination of drug and placebo data reveals no apparent improvements in inappropriate behavior. However, as might be expected, the switch to placebo for Subject 18 led to an increase in inappropriate behavior, suggesting at least some controlling effects of the drug. When response-cost procedures were instituted in Week 14 for Subjects 12, 13, 15, 16, and 18, and in Week 17 for Subjects 14 and 17, marked improvements in appropriate behavior were observed, beginning almost immediately. Thus, this rather complicated experimental analysis confirmed the efficacy of response cost procedures under time-lagged conditions, but only when the contingency was directly applied. Both neuroleptic drugs and placebo generally seemed to be ineffective.

In this type of drug evaluation it is important to underscore that the prolonged placebo phases are important in that they provide a needed “wash-out” period for possible carryover effects of drugs. This, of course, would have been much more critical had neuroleptic drugs substantially decreased the behavior targeted for change (i.e., inappropriate behavior).

An alternative use of the multiple baseline design in drug evaluations is one in which the investigator tests the effectiveness of combined treatment relative to using either drug treatment or psychological treatment alone. As an example of such a design, Bach, Barlow, and Winze (2004) used a multiple baseline design to test the effectiveness of adding cognitive-behavioral treatment to administration of Sildenafil on erectile dysfunction in a series of six heterosexual couples. This study revealed large and consistent treatment gains associated with administration of cognitive-behavioral treatment above and beyond the effects of Sildenafil (see chapter 10 for further description of this study). Taken together, these studies of drug evaluations provide only a few illustrations of the many variations on the multiple baseline design that can be used to study the effects of psychological and pharmacological treatments on behavior.
FIGURE 7.21 Frequencies of inappropriate behaviors for Subjects 12-18 plotted as total occurrences per week (summed daily interval totals). During the D condition, the subjects received their drug; during the P condition, the subjects received a placebo, were no longer receiving their drug, and the response cost procedure was not in effect. Drugs were discontinued during the first 3 weeks of the P condition. During the RC condition, the response cost procedure was in effect, and the subjects were not receiving their drug. The dotted vertical lines separate the conditions. (Figure 2, p. 261, from: Breuning, S. E., O'Neill, M. J., & Ferguson, D. G. [1980]. Comparison of psychotropic drug, response cost, and psychotropic drug plus response cost procedures for controlling institutionalized mentally retarded persons. Applied Research in Mental Retardation, 1, 253–268. Copyright 1980. Reproduced by permission.)