CHAPTER 9

Multiple Baseline and Changing Criterion Designs

Key Terms

changing criterion design
delayed multiple baseline design
multiple baseline across behaviors design
multiple baseline across settings design
multiple baseline across subjects design
multiple baseline design
multiple probe design

Behavior Analyst Certification Board® BCBA® & BCABA® Behavior Analyst Task List©, Third Edition

Content Area 5: Experimental Evaluation of Interventions	
5-1	Systematically manipulate independent variables to analyze their effects on treatment.
(d)	Use changing criterion design.
(e)	Use multiple baseline designs.
5-2	Identify and address practical and ethical considerations in using various experimental designs.

This chapter describes two additional experimental tactics for analyzing behavior–environment relations—the multiple baseline design and the changing criterion design. In a multiple baseline design, after collecting initial baseline data simultaneously across two or more behaviors, settings, or people, the behavior analyst then applies the treatment variable sequentially across these behaviors, settings, or people and notes the effects. The changing criterion design is used to analyze improvements in behavior as a function of stepwise, incremental criterion changes in the level of responding required for reinforcement. In both designs, experimental control and a functional relation are demonstrated when the behaviors change from a steady state baseline to a new steady state after the introduction of the independent variable is applied, or a new criterion established.

Multiple Baseline Design

The multiple baseline design is the most widely used experimental design for evaluating treatment effects in applied behavior analysis. It is a highly flexible tactic that enables researchers and practitioners to analyze the effects of an independent variable across multiple behaviors, settings, and/or subjects without having to withdraw the treatment variable to verify that the improvements in behavior were a direct result of the application of the treatment. As you recall from Chapter 8, the reversal design by its very nature requires that the independent variable be withdrawn to verify the prediction established in baseline. This is not so with the multiple baseline design.

Operation and Logic of the Multiple Baseline Design

Baer, Wolf, and Risley (1968) first described the **multiple baseline design** in the applied behavior analysis literature. They presented the multiple baseline design as an alternative to the reversal design for two situations: (a) when the target behavior is likely to be irreversible or (b) when it is undesirable, impractical, or unethical to reverse conditions. Figure 9.1 illustrates Baer and colleagues' explanation of the basic operation of the multiple baseline design.

> In the multiple baseline technique, a number of responses are identified and measured over time to provide baselines against which changes can be evaluated. With 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 no change in the other baselines. If so, rather than reversing the just-produced change, he instead applies the experimental variable to one of the other, as yet unchanged, responses. If it changes at that point, evidence is accruing that the experimental variable is indeed effective, and that the prior change was not simply a matter of coincidence. The variable then may be applied to still another response, and so on. The experimenter is attempting to show that he has a reliable experimental variable, in that each behavior changes maximally only when the experimental variable is applied to it. (p. 94)

The multiple baseline design takes three basic forms:

- The multiple baseline across behaviors design, consisting of two or more different behaviors of the same subject
- The multiple baseline across settings design, consisting of the same behavior of the same subject in two or more different settings, situations, or time periods
- The multiple baseline across subjects design, consisting of the same behavior of two or more different participants (or groups)

Although only one of the multiple baseline design's basic forms is called an "across behaviors" design, all multiple baseline designs involve the time-lagged application of a treatment variable across technically different (meaning independent) behaviors. That is, in the multiple baseline across settings design, even though the subject's performance of the same target behavior is measured in two or more settings, each behavior–setting combination is conceptualized and treated as a different behavior for analysis. Similarly, in a multiple baseline across subjects design, each subject–behavior combination functions as a different behavior in the operation of the design.

Figure 9.2 shows the same data set displayed in Figure 9.1 with the addition of data points representing predicted measures if baseline conditions were not changed and shaded areas illustrating how the three elements of baseline logic—prediction, verification, and replication—are operationalized in the multiple baseline design.[1] When stable baseline responding has been achieved for Behavior 1, a *prediction* is made that if the environment were held constant, continued measurement

[1]Although most of the graphic displays created or selected for this text as examples of experimental design tactics show data plotted on noncumulative vertical axes, the reader is reminded that repeated measurement data collected within any type of experimental design can be plotted on both noncumulative and cumulative graphs. For example, Lalli, Zanolli, and Wohn (1994) and Mueller, Moore, Doggett, and Tingstrom (2000) used cumulative graphs to display the data they collected in multiple baseline design experiments; and Kennedy and Souza (1995) and Sundberg, Endicott, and Eigenheer (2000) displayed the data they obtained in reversal designs on cumulative graphs. Students of applied behavior analysis should be careful not to confuse the different techniques for graphically displaying data with tactics for experimental analysis.

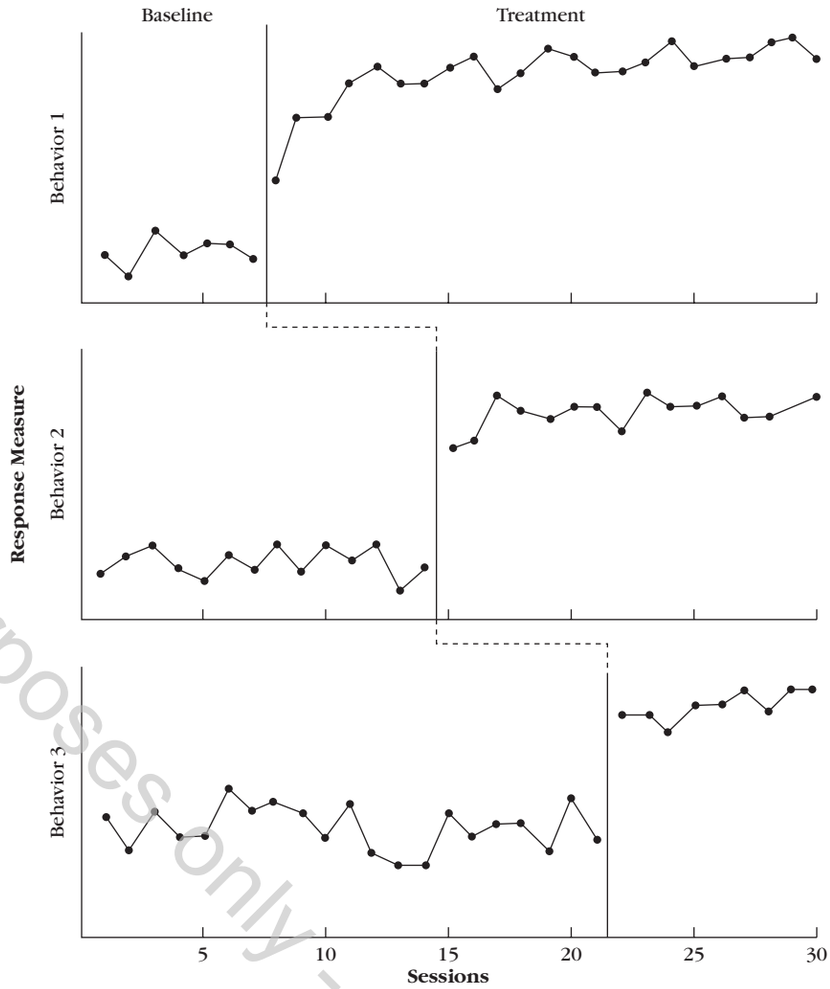

Figure 9.1 Graphic prototype of a multiple baseline design.

would reveal similar levels of responding. When the researcher's confidence in such a prediction is justifiably high, the independent variable is applied to Behavior 1. The open data points in the treatment phase for Behavior 1 represent the predicted level of responding. The solid data points show the actual measures obtained for Behavior 1 during the treatment condition. These data show a discrepancy with the predicted level of responding if no changes had been made in the environment, thereby suggesting that the treatment may be responsible for the change in behavior. The data collected for Behavior 1 in a multiple baseline design serve the same functions as the data collected during the first two phases of an A-B-A-B reversal design.

Continued baseline measures of the other behaviors in the experiment offer the possibility of verifying the prediction made for Behavior 1. In a multiple baseline design, *verification* of a predicted level of responding for one behavior (or tier) is obtained if little or no change is observed in the data paths of the behaviors (tiers) that are still exposed to the conditions under which the prediction was made. In Figure 9.2 those portions of the baseline condition data paths for Behaviors 2 and 3 within the shaded boxes verify the prediction for Behavior 1. At this point in the experiment, two inferences can be made: (a) The prediction that Behavior 1 would not change in a constant environment is valid because the environment was held constant for Behaviors 2 and 3 and their levels of responding remained unchanged; and (b) the observed changes in Behavior 1 were brought about by the independent variable because only Behavior 1 was exposed to the independent variable and only Behavior 1 changed.

In a multiple baseline design, the independent variable's function in changing a given behavior is inferred by the lack of change in untreated behaviors. However, verification of function is not demonstrated directly as it is with the reversal design, thereby making the multiple baseline design an inherently weaker tactic (i.e., less convincing from the perspective of experimental control) for revealing a functional relation between the independent

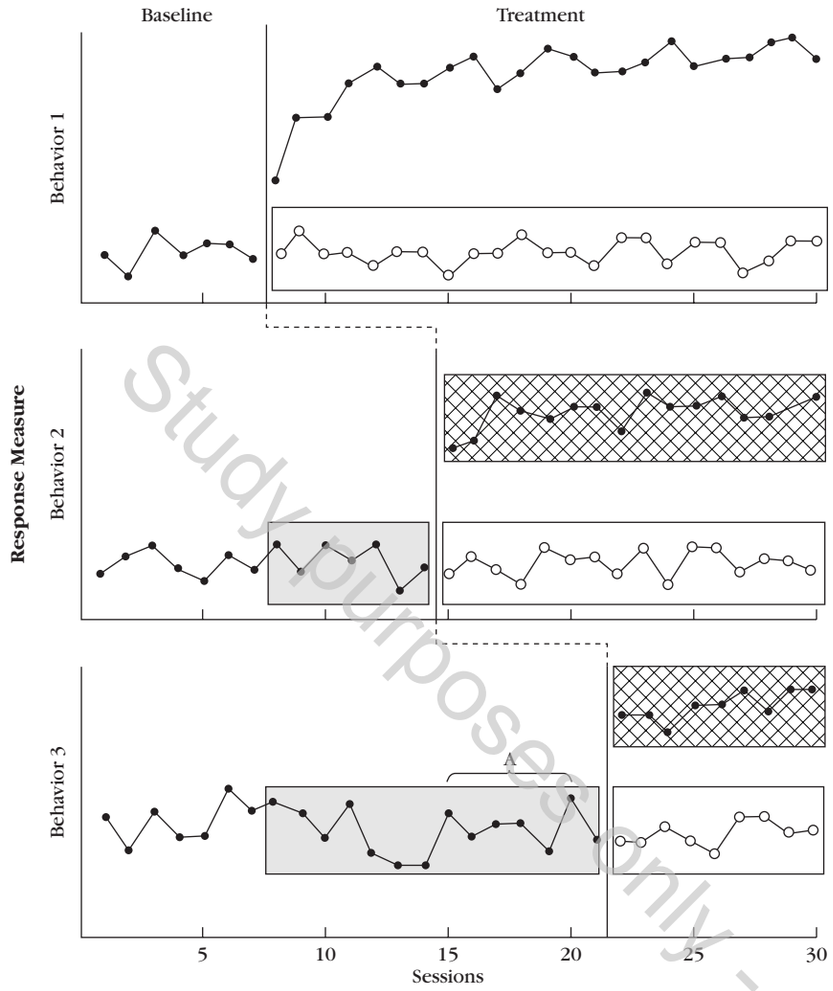

Figure 9.2 Graphic prototype of a multiple baseline design with shading added to show elements of baseline logic. Open data points represent predicted measures if baseline conditions were unchanged. Baseline data points for Behaviors 2 and 3 within shaded areas verify of the prediction made for Behavior 1. Behavior 3 baseline data within Bracket A verify the prediction made for Behavior 2. Data obtained during the treatment condition for Behaviors 2 and 3 (cross-hatched shading) provide replications of the experimental effect.

variable and a target behavior. However, the multiple baseline design compensates somewhat for this weakness by providing the opportunity to verify or refute a series of similar predictions. Not only is the prediction for Behavior 1 in Figure 9.2 verified by continued stable baselines for Behavior 2 and 3, but the bracketed portion of the baseline data for Behavior 3 also serves as verification of the prediction made for Behavior 2.

When the level of responding for Behavior 1 under the treatment condition has stabilized or reached a predetermined performance criterion, the independent variable is then applied to Behavior 2. If Behavior 2 changes in a manner similar to the changes observed for Behavior 1, *replication* of the independent variable's effect has been achieved (shown by the data path shaded with cross-hatching). After Behavior 2 has stabilized or reached a predetermined performance criterion, the independent variable is applied to Behavior 3 to see whether the effect will be replicated. The independent variable may be applied to additional behaviors in a similar manner until a convincing demonstration of the functional relation has been established (or rejected) and all of the behaviors targeted for improvement have received treatment.

As with verification, replication of the independent variable's specific effect on each behavior in a multiple baseline design is not manipulated directly. Instead, the generality of the independent variable's effect across the behaviors comprising the experiment is demonstrated by applying it to a series of behaviors. Assuming accurate measurement and proper experimental control of relevant variables (i.e., the only environmental factor that changes during the course of the experiment should be the presence—or absence—of the independent variable), each time a behavior changes when, and only when, the independent variable is introduced, confidence in the existence of a functional relation increases.

How many different behaviors, settings, or subjects must a multiple baseline design include to provide a believable demonstration of a functional relation? Baer, Wolf, and Risley (1968) suggested that the number of

replications needed in any design is ultimately a matter to be decided by the consumers of the research. In this sense, an experiment using a multiple baseline design must contain the minimum number of replications necessary to convince those who will be asked to respond to the experiment and to the researcher's claims (e.g., teachers, administrators, parents, funding sources, journal editors). A two-tier multiple baseline design is a complete experiment and can provide strong support for the effectiveness of the independent variable (e.g., Lindberg, Iwata, Roscoe, Worsdell, & Hanley, 2003 [see Figure 23.2]; McCord, Iwata, Galensky, Ellingson, & Thomson, 2001 [see Figure 6.6]; Newstrom, McLaughlin, & Sweeney, 1999 [see Figure 26.2]; Test, Spooner, Keul, & Grossi, 1990 [see Figure 20.7]). McClannahan, McGee, MacDuff, and Krantz (1990) conducted a multiple baseline design study in which the independent variable was sequentially implemented in an eight-tier design across 12 participants. Multiple baseline designs of three to five tiers are most common. When the effects of the independent variable are substantial and reliably replicated, a three- or four-tier multiple baseline design provides a convincing demonstration of experimental effect. Suffice it to say that the more replications one conducts, the more convincing the demonstration will be.

Some of the earliest examples of the multiple baseline design in the applied behavior analysis literature were studies by Risley and Hart (1968); Barrish, Saunders, and Wolf (1969); Barton, Guess, Garcia, and Baer (1970); Panyan, Boozer, and Morris (1970); and Schwarz and Hawkins (1970). Some of the pioneering applications of the multiple baseline technique are not readily apparent with casual examination: The authors may not have identified the experimental design as a multiple baseline design (e.g., Schwarz & Hawkins, 1970), and/or the now-common practice of stacking the tiers of a multiple baseline design one on the other so that all of the data can be displayed graphically in the same figure was not always used (e.g., Maloney & Hopkins, 1973; McAllister, Stachowiak, Baer, & Conderman, 1969; Schwarz & Hawkins, 1970).

In 1970, Vance Hall, Connie Cristler, Sharon Cranston, and Bonnie Tucker published a paper that described three experiments, each an example of one of the three basic forms of the multiple baseline design: across behaviors, across settings, and across subjects. Hall and colleagues' paper was important not only because it provided excellent illustrations that today still serve as models of the multiple baseline design, but also because the studies were carried out by teachers and parents, indicating that practitioners "can carry out important and significant studies in natural settings using resources available to them" (p. 255).

Multiple Baseline across Behaviors Design

The **multiple baseline across behaviors design** begins with the concurrent measurement of two or more behaviors of a single participant. After steady state responding has been obtained under baseline conditions, the investigator applies the independent variable to one of the behaviors while maintaining baseline conditions for the other behavior(s). When steady state or criterion-level performance has been reached for the first behavior, the independent variable is applied to the next behavior, and so on (e.g., Bell, Young, Salzberg, & West, 1991; Gena, Krantz, McClannahan, & Poulson, 1996; Higgins, Williams, & McLaughlin, 2001 [see Figure 26.8]).

Ward and Carnes (2002) used a multiple baseline across behaviors design to evaluate the effects of self-set goals and public posting on the execution of three skills by five linebackers on a college football team: (a) *reads,* in which the linebacker positions himself to cover a specified area on the field on a pass play or from the line of scrimmage on a run; (b) *drops,* in which the linebacker moves to the correct position depending on the offensive team's alignment; and (c) *tackles.* A video camera recorded the players' movements during all practice sessions and games. Data were collected for the first 10 opportunities each player had with each skill. Reads and drops were recorded as correct if the player moved to the zone identified in the coaches' playbook; tackles were scored as correct if the offensive ball carrier was stopped.

Following baseline, each player met with one of the researchers, who described the player's mean baseline performance for a given skill. Players were asked to set a goal for their performances during practice sessions; no goals were set for games. The correct performances during baseline for all five players ranged from 60 to 80%, and all players set goals of 90% correct performance. The players were informed that their performance in each day's practice would be posted on a chart prior to the next practice session. A *Y* (yes) or an *N* (no) was placed next to each player's name to indicate whether he had met his goal. A player's performance was posted on the chart only for the skill(s) in intervention. The chart was mounted on a wall in the locker room where all players on the team could see it. The head coach explained the purpose of the chart to other players on the team. Players' performances during games were not posted on the chart.

The results for one of the players, John, are shown in Figure 9.3. John met or exceeded his goal of 90% correct performance during all practices for each of the three skills. Additionally, his improved performance generalized to games. The same pattern of results was obtained for each of the other four players in the study, illustrating

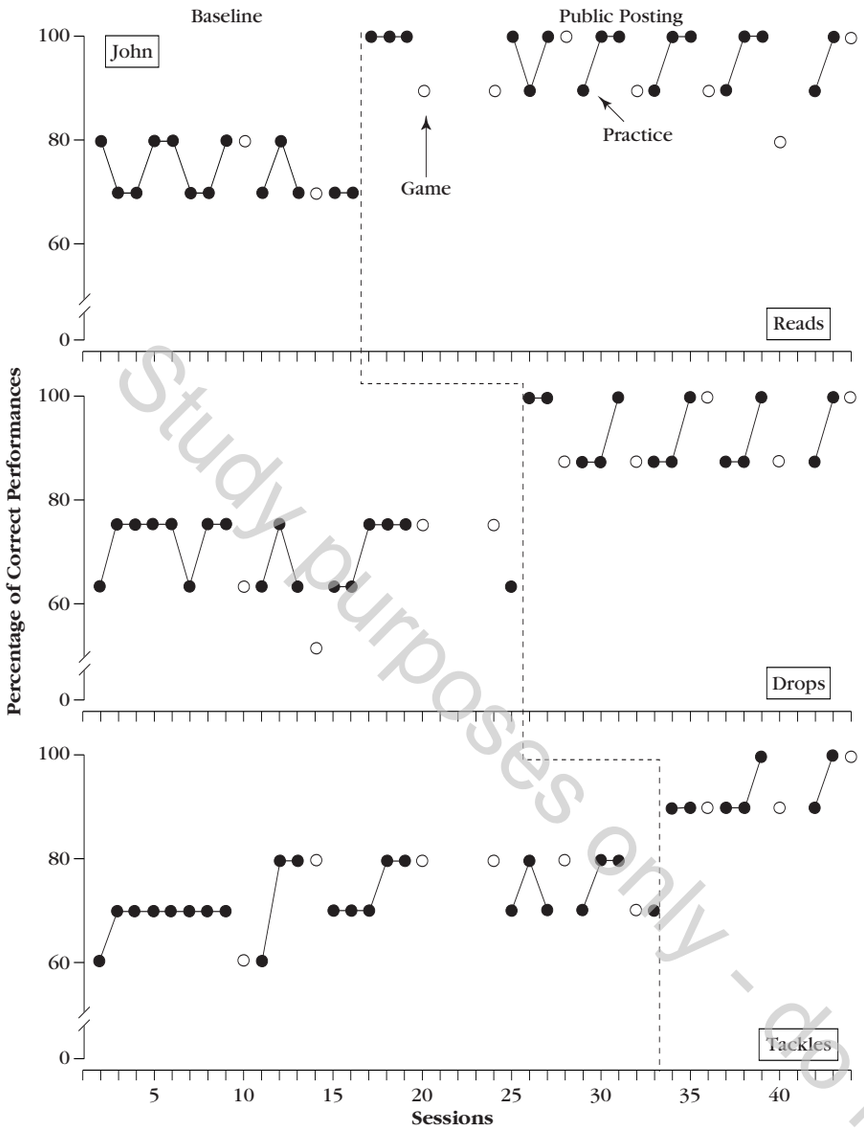

Figure 9.3 A multiple baseline across behaviors design showing percentage of correct reads, drops, and tackles by a college football player during practices and games.

From "Effects of Posting Self-Set Goals on Collegiate Football Players' Skill Execution During Practice and Games" by P. Ward and M. Carnes, 2002, *Journal of Applied Behavior Analysis, 35,* p. 5. Copyright 2002 by the Society for the Experimental Analysis of Behavior, Inc. Reprinted by permission.

that the multiple baseline across behaviors design is a single-subject experimental strategy in which each subject serves as his own control. Each player constituted a complete experiment, replicated in this case with four other participants.

Multiple Baseline across Settings Design

In the **multiple baseline across settings design,** a single behavior of a person (or group) is targeted in two or more different settings or conditions (e.g., locations, times of day). After stable responding has been demonstrated under baseline conditions, the independent variable is introduced in one of the settings while baseline conditions remain in effect in the other settings. When maximum behavior change or criterion-level performance has been achieved in the first setting, the independent variable is applied in the second setting, and so on.

Roane, Kelly, and Fisher (2003) employed a multiple baseline across settings design to evaluate the effects of a treatment designed to reduce the rate at which an 8-year-old boy put inedible objects in his mouth. Jason, who had been diagnosed with autism, cerebral palsy, and moderate mental retardation, had a history of putting objects such as toys, cloth, paper, tree bark, plants, and dirt into his mouth.

Data on Jason's mouthing were obtained concurrently in a classroom, a playroom, and outdoors—three settings that contained a variety of inedible objects and

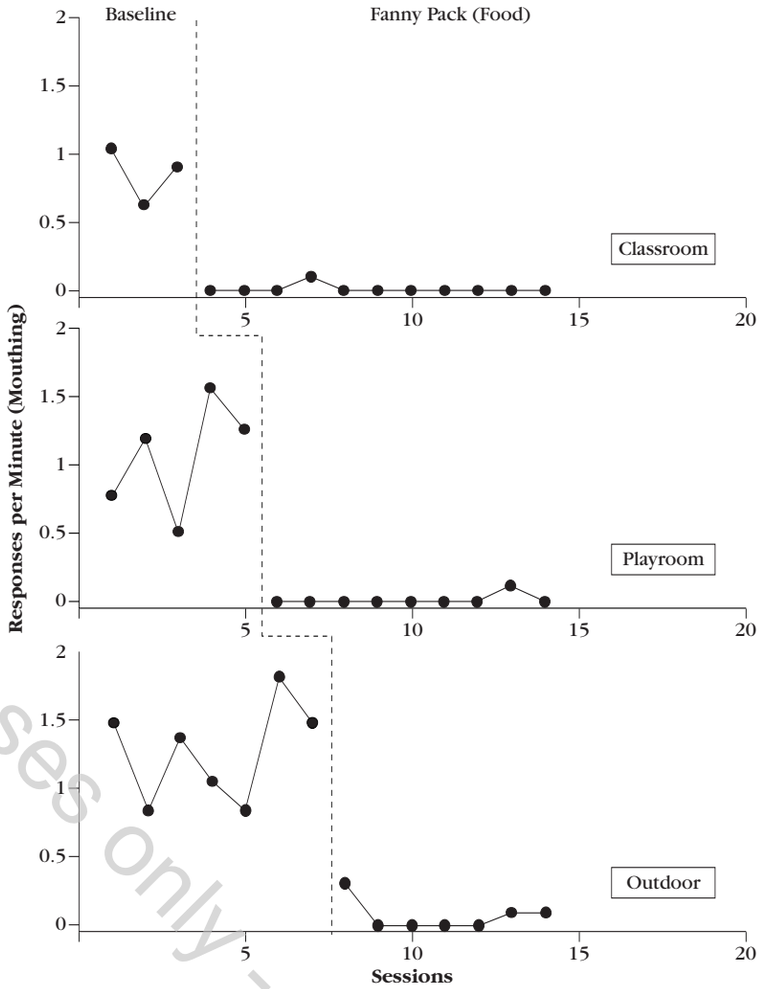

Figure 9.4 A multiple baseline across settings design showing the number of object mouthing responses per minute during baseline and treatment conditions.

From "The Effects of Noncontingent Access to Food on the Rate of Object Mouthing across Three Settings" by H. S. Roane, M. L. Kelly, and W. W. Fisher, 2003, *Journal of Applied Behavior Analysis, 36,* p. 581. Copyright 2003 by the Society for the Experimental Analysis of Behavior, Inc. Reprinted by permission.

where caretakers had reported Jason's mouthing to be problematic. Observers in each setting unobtrusively tallied the number of times Jason inserted an inedible object past the plane of his lips during 10-minute sessions. The researchers reported that Jason's object mouthing usually consisted of a series of discrete episodes, rather than an extended, continuous event, and that he often placed multiple objects (inedible objects and food) in his mouth simultaneously.

Roane and colleagues (2003) described the baseline and treatment conditions for Jason as follows:

> The baseline condition was developed based on the functional analysis results, which showed that mouthing was maintained by automatic reinforcement and occurred independent of social consequences. During baseline, a therapist was present (approximately 1.5 to 3 m from Jason), but all occurrences of mouthing were ignored (i.e., no social consequences were arranged for mouthing, and Jason was allowed to place items in his mouth). No food items were available during baseline. The treatment condition was identical to baseline except that Jason had continuous access to foods that had been previously identified to compete with the occurrence of object mouthing: chewing gum, marshmallows, and hard candy. Jason wore a fanny pack containing these items around his waist. (pp. 580–581)[2]

The staggered sequence in which the treatment was implemented in each setting and the results are shown in Figure 9.4. During baseline, Jason's mouthed objects at mean rates of 0.9, 1.1, and 1.2 responses per minute in the classroom, a playroom, and outdoor settings, respectively. Introduction of the fanny pack with food in each setting produced an immediate drop to a zero or near zero rate of mouthing. During treatment, Jason put items of food from the fanny pack into his mouth at mean rates of 0.01, 0.01, and 0.07 responses per minute in the classroom, a playroom, and outdoor settings, respectively. The multiple baseline across settings design revealed a clear functional relation between the treatment and the frequency of Jason's object mouthing. No measures obtained during the treatment condition were as

[2]Functional analysis and automatic reinforcement are described in Chapters 24 and 11, respectively.

high as the lowest measures in baseline. During 22 of 27 treatment sessions across the three settings, Jason put no inedible objects in his mouth.

As was done in the study by Roane and colleagues (2003), the data paths that comprise the different tiers in a multiple baseline across settings design are typically obtained in different physical environments (e.g., Cushing & Kennedy, 1997; Dalton, Martella, & Marchand-Martella, 1999). However, the different "settings" in a multiple baseline across settings design may exist in the same physical location and be differentiated from one another by different contingencies in effect, the presence or absence of certain people, and/or the different times of the day. For example, in a study by Parker and colleagues (1984), the presence or absence of other people in the training room constituted the different settings (environments) in which the effects of the independent variable were evaluated. The attention, demand, and no-attention conditions (i.e., contingencies in effect) defined the different settings in a multiple baseline design study by Kennedy, Meyer, Knowles, and Shukla (2000, see Figure 6.4). The afternoon and the morning portions of the school day functioned as different settings in the multiple baseline across settings design used by Dunlap, Kern-Dunlap, Clarke, and Robbins (1991) to analyze the effects of curricular revisions on a student's disruptive and off-task behaviors.

In some studies using a multiple baseline across settings design, the participants are varied, changing, and perhaps even unknown to the researchers. For example, Van Houten and Malenfant (2004) used a multiple baseline design across two crosswalks on busy streets to evaluate the effects of an intensive driver enforcement program on the percentage of drivers yielding to pedestrians and the number of motor vehicle–pedestrian conflicts. Watson (1996) used a multiple baseline design across men's rest rooms on a college campus to assess the effectiveness of posting signs in reducing bathroom graffiti.

Multiple Baseline across Subjects Design

In the **multiple baseline across subjects design,** one target behavior is selected for two or more subjects (or groups) in the same setting. After steady state responding has been achieved under baseline conditions, the independent variable is applied to one of the subjects while baseline conditions remain in effect for the other subjects. When criterion-level or stable responding has been attained for the first subject, the independent variable is applied to another subject, and so on. The multiple baseline across subjects design is the most widely used of all three forms of the design, in part because teachers, clinicians, and other practitioners are commonly confronted by more than one student or client needing to learn the same skill or eliminate the same problem behavior (e.g., Craft, Alber, & Heward, 1998; Kahng, Iwata, DeLeon, & Wallace, 2000 [see Figure 23.1]; Killu, Sainato, Davis, Ospelt, & Paul, 1998 [see Figure 23.3]; Kladopoulos & McComas, 2001 [see Figure 6.3]). Sometimes a multiple baseline design is conducted across "groups" of participants (e.g., Dixon & Holcomb, 2000 [see Figure 13.7]; Lewis, Powers, Kelk, & Newcomer, 2002 [see Figure 26.12]; White & Bailey, 1990 [see Figure 15.2]).

Krantz and McClannahan (1993) used a multiple baseline across subjects design to investigate the effects of introducing and fading scripts to teach children with autism to interact with their peers. The four participants, ages 9 to 12, had severe communication deficits and minimal or absent academic, social, leisure skills. Prior to the study each of the children had learned to follow first photographic activity schedules (Wacker & Berg, 1983) and later written activity schedules that prompted them through chains of academic, self-care, and leisure activities. Although their teachers modeled social interactions, verbally prompted the children to interact, and provided contingent praise and preferred snacks and activities for doing so, the children consistently failed to initiate interactions without adult prompts.

Each session consisted of a continuous 10-minute interval in which observers recorded the number of times each child initiated and responded to peers while engaged in three art activities—drawing, coloring, and painting—that were rotated across sessions throughout the study. Krantz and McClannahan (1993) described the dependent variables as follows:

> *Initiation to peers* was defined as understandable statements or questions that were unprompted by an adult, that were directed to another child by using his or her name or by facing him or her, and that were separated from the speaker's previous vocalizations by a change in topic or a change in recipient of interaction. . . . *Scripted interactions* were those that matched the written script, . . . e.g., "Ross, I like your picture." *Unscripted interactions* differed from the script by more than changes in conjunctions, articles, prepositions, pronouns, or changes in verb tense; the question, "Would you like some more paper?" was scored as an unscripted initiation because the noun "paper" did not occur in the script. A *response* was defined as any contextual utterance (word, phrase, or sentence) that was not prompted by the teacher and that occurred within 5 s of a statement or question directed to the target child. . . . Examples of responses were "what?" "okay," and "yes, I do." (p. 124)

During baseline, each child found art materials at his or her place and a sheet of paper with the written instructions, "Do your art" and "Talk a lot." The teacher prompted each child to read the written instructions, then moved away. During the script condition, the two written

instructions in baseline were supplemented by scripts consisting of 10 statements and questions such as, "{Name}, did you like to {swing/rollerskate/ride the bike} outside today?" "{Name}, do you want to use one of my pencils/crayons/brushes}?" (p. 124). Immediately before each session, the teacher completed blank portions of the scripts so that they reflected activities the children had completed or were planning and objects in the classroom environment. Each child's script included the three other children's names, and the order of the questions or statements varied across sessions and children.

The script condition was implemented with one child at a time, in staggered fashion (see Figure 9.5). Initially the teacher manually guided the child through the script, prompting him or her to read the statement to another child and to pencil a check mark next to it after doing so. Krantz and McClannahan (1993) described the prompting and script-fading procedures as follows:

> Standing behind a participant, the teacher manually guided him or her to pick up a pencil, point to an instruction or a scripted statement or question, and move the pencil along below the text. If necessary, the teacher also manually guided the child's head to face another child to whom a statement or question was addressed. If the child did not verbalize the statement or questions within 5 s, the manual guidance procedure was repeated. If the child read or said a statement or read or asked a question, the teacher used the same type of manual guidance to ensure that the child placed a check mark to the left of that portion of the script.
>
> Manual prompts were faded as quickly as possible; no prompts were delivered to Kate, Mike, Walt, and

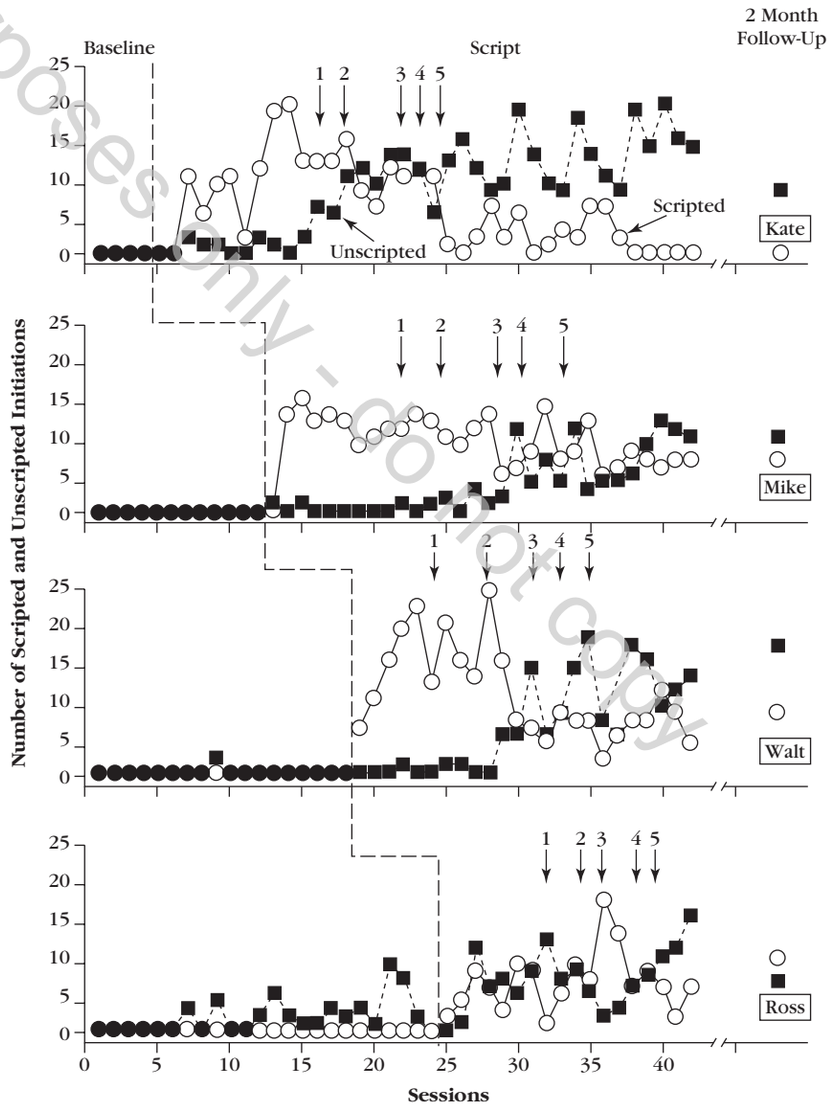

Figure 9.5 A multiple baseline across subjects design showing the number of scripted and unscripted initiations to peers and responses by four children with autism during baseline, script, and follow-up sessions. Arrows indicate when fading steps occurred.

From "Teaching Children with Autism to Initiate to Peers: Effects of a Script-Fading Procedure" by P. J. Krantz and L. E. McClannahan, 1993, *Journal of Applied Behavior Analysis, 26,* p. 129. Copyright 1993 by the Society for the Experimental Analysis of Behavior, Inc. Reprinted by permission.

> Ross after Sessions 15, 18, 23, and 27, respectively, and the teacher remained at the periphery of the classroom throughout subsequent sessions. After manual guidance had been faded for a target child, fading of the script began. Scripts were faded from end to beginning in five phases. For example, the fading steps for the question "Mike, what do you like to do best on Fun Friday?" were (a) "Mike, what do you like to do best," (b) "Mike, what do you," (c) "Mike, what," (d) "M," and (e) "." (p. 125)

Kate and Mike, who never initiated during baseline, had mean initiations per session of 15 and 13, respectively, during the script condition. Walt's initiations increased from a baseline mean of 0.1 to 17 during the script condition, and Ross averaged 14 initiations per session during script compared to 2 during baseline. As the scripts were faded, each child's frequency of unscripted initiations increased. After the scripts were faded, the four participants' frequency of initiations were within the same range as that of a sample of three typically developing children. The researchers implemented the script-fading steps with each participant in response to his or her performance, not according to a predetermined schedule, thereby retaining the flexibility needed to pursue the behavior–environment relations that are the focus of the science of behavior.

However, because each subject did not serve as his or her own control, this study illustrates that the multiple baseline across subjects design is not a true single-subject design. Instead, verification of predictions based on the baseline data for each subject must be inferred from the relatively unchanging measures of the behavior of other subjects who are still in baseline, and replication of effects must be inferred from changes in the behavior of other subjects when they come into contact with the independent variable. This is both a weakness and a potential advantage of the multiple baseline across subjects design (Johnston & Pennypacker, 1993a), discussed later in the chapter.

Variations of the Multiple Baseline Design

Two variations of the multiple baseline design are the multiple probe design and the delayed multiple baseline design. The multiple probe design enables the behavior analyst to extend the operation and logic of the multiple baseline tactic to behaviors or situations in which concurrent measurement of all behaviors comprising the design is unnecessary, potentially reactive, impractical, or too costly. The delayed multiple baseline technique can be used when a planned reversal design is no longer possible or proves ineffective; it can also add additional tiers to an already operational multiple baseline design, as would be the case if new subjects were added to an ongoing study.

Multiple Probe Design

The **multiple probe design,** first described by Horner and Baer (1978), is a method of analyzing the relation between the independent variable and the acquisition of a successive approximation or task sequence. In contrast to the multiple baseline design—in which data are collected simultaneously throughout the baseline phase for each behavior, setting, or subject in the experiment—in the multiple probe design intermittent measures, or probes, provide the basis for determining whether behavior change has occurred prior to intervention. According to Horner and Baer, when applied to a chain or sequence of related behaviors to be learned, the multiple probe design provides answers to four questions: (a) What is the initial level of performance on each step (behavior) in the sequence? (b) What happens when sequential opportunities to perform each step in the sequence are provided prior to training on that step? (c) What happens to each step as training is applied? and (d) What happens to the performance of untrained steps in the sequence as criterion-level performance is reached on the preceding steps?

Figure 9.6 shows a graphic prototype of the multiple probe design. Although researchers have developed many variations of the multiple probe technique, the basic design has three key features: (a) An initial probe is taken to determine the subject's level of performance on each behavior in the sequence; (b) a series of baseline measures is obtained on each step prior to training on that step; and (c) after criterion-level performance is reached on any training step, a probe of each step in the sequence is obtained to determine whether performance changes have occurred in any other steps.

Thompson, Braam, and Fuqua (1982) used a multiple probe design to analyze the effects of an instructional procedure composed of prompts and token reinforcement on the acquisition of a complex chain of laundry skills by three students with developmental disabilities. Observations of people doing laundry resulted in a detailed task analysis of 74 discrete responses that were organized into seven major components (e.g., sorting, loading washer). Each student's performance was assessed via probe and baseline sessions that preceded training on each component. Probe and baseline sessions began with instructions to the student to do the laundry. When an incorrect response was emitted or when no response occurred within 5 seconds of a prompt to continue, the student was seated away from the laundry area. The trainer then performed the correct response and called the student back to the area so that assessment of the rest of the laundry sequence could continue.

> Probe sessions differed from baseline sessions in two ways. First, a probe measured each response in the entire

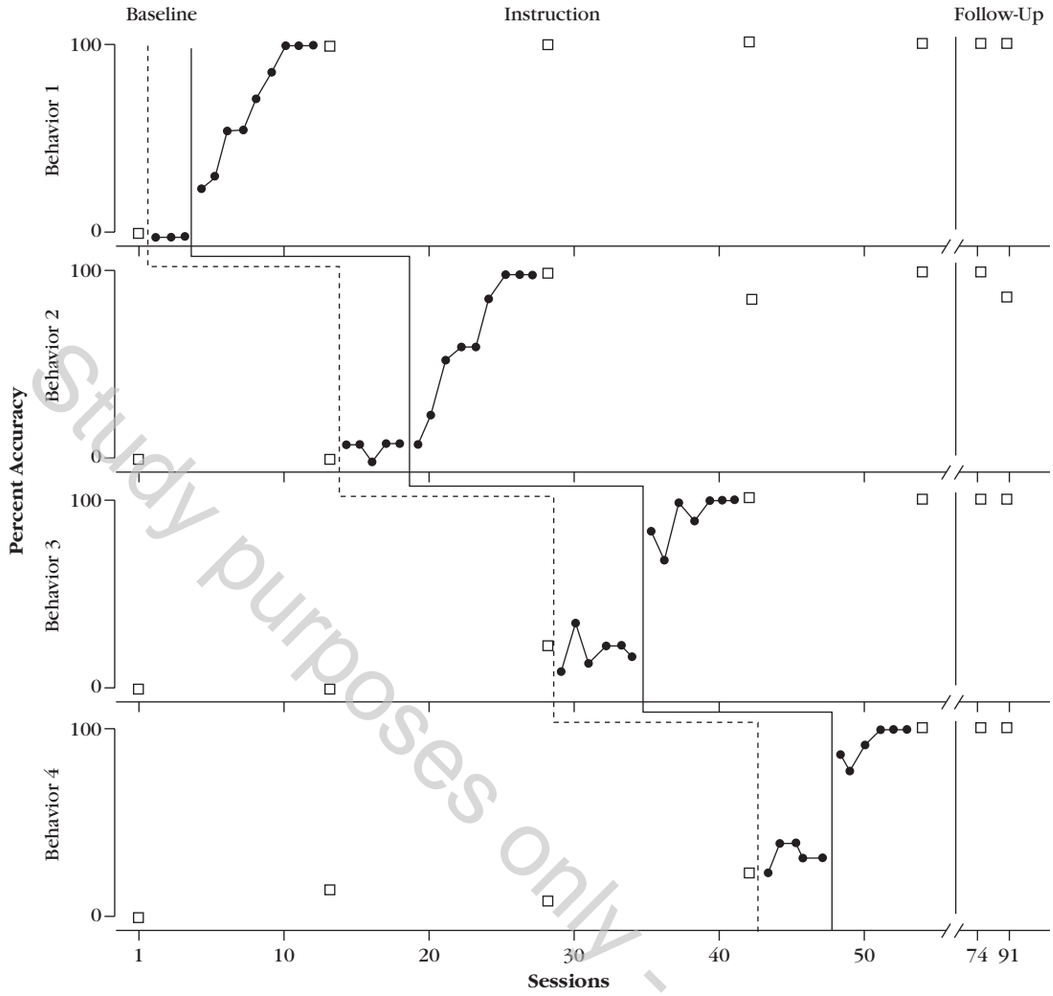

Figure 9.6 Graphic prototype of a multiple probe design. Square data points represent results of probe sessions in which the entire sequence or set of behaviors (1–4) are tested.

chain and occurred immediately prior to baseline and training for every component. Baseline sessions occurred following the probe and measured only previously trained components plus the component about to be trained. Baseline data were gathered on a variable number of consecutive sessions immediately prior to training sessions. Second, no tokens or descriptive praise were delivered during probes. During baseline, tokens were delivered for previously trained responses only. . . . Following baseline, each component was trained using a graduated 3-prompt procedure (Horner & Keilitz, 1975), consisting of verbal instruction, modeling, and graduated guidance. If one prompt level failed to produce a correct response within 5 sec, the next level was introduced. . . . When the student performed a component at 100% accuracy for two consecutive trials, he was required to perform the entire laundry chain from the beginning through the component most recently mastered. The entire chain of previously mastered components was trained (chain training condition) until it was performed without errors or prompts for two consecutive trials. (Thompson, Braam, & Fuqua, 1982, p. 179)

Figure 9.7 shows the results for Chester, one of the students. Chester performed a low percentage of correct responses during the probe and baseline sessions, but performed with 100% accuracy after training was applied to each component. During a generalization probe conducted at a community laundromat after training, Chester performed correctly 82% of the 74 total responses in the chain. Five additional training sessions were needed to retrain responses performed incorrectly during the generalization probe and to train "additional responses necessitated by the presence of coin slots and minor differences between the training and laundromat equipment" (p. 179). On two follow-up sessions conducted 10 months after training, Chester performed at 90% accu-

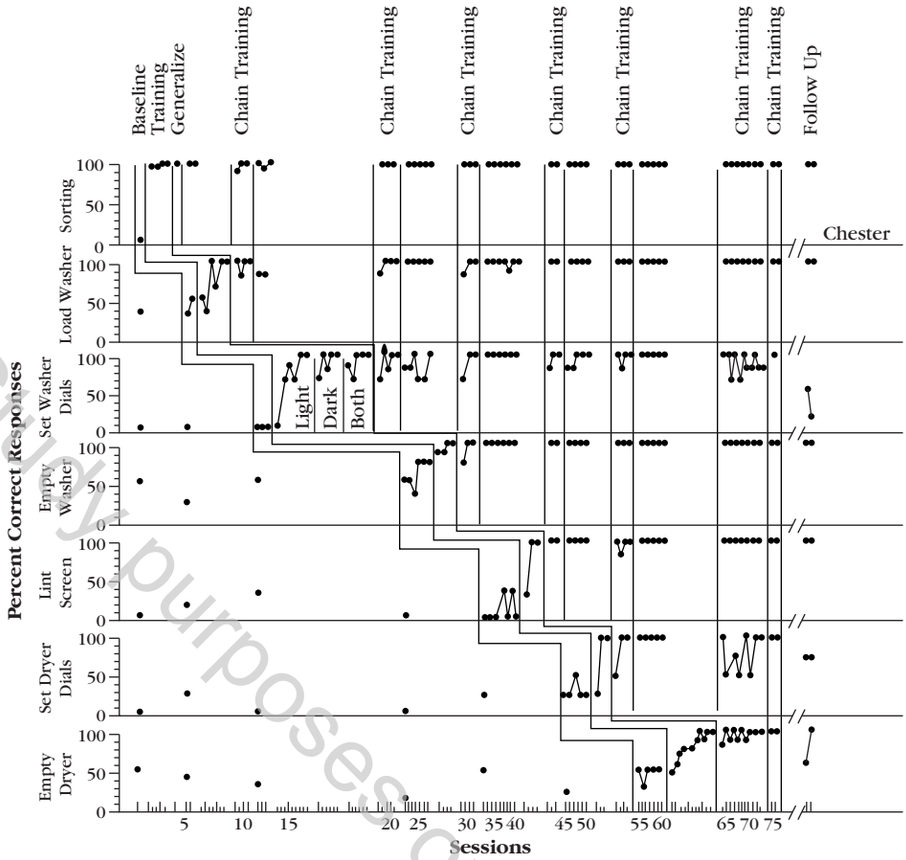

Figure 9.7 A multiple probe design showing the percentage of correct responses for each trial on each component of a laundry task by a young adult male with mental retardation. Heavy vertical lines on the horizontal axis represent successive training sessions; lighter and shorter vertical lines indicate trials within a session.

From "Training and Generalization of Laundry Skills: A Multiple-Probe Evaluation with Handicapped Persons" by T. J. Thompson, S. J. Braam, and R. W. Fuqua, 1982, *Journal of Applied Behavior Analysis, 15,* p. 180. Copyright 1982 by the Society for the Experimental Analysis of Behavior, Inc. Reprinted by permission.

racy even though he had not performed the laundry task for the past 2 months. Similar results were obtained for the other two students who participated in the study.

Thompson and colleagues (1982) added the chain training condition to their study because they believed that components trained as independent skills were unlikely to be emitted in correct sequence without such practice. It should be noted that the experimenters did not begin training a new component until stable responding had been achieved during baseline observations (see the baseline data for the bottom four tiers in Figure 9.7). Delaying the training in this manner enabled a clear demonstration of a functional relation between training and skill acquisition.

The multiple probe design is particularly appropriate for evaluating the effects of instruction on skill sequences in which it is highly unlikely that the subject can improve performance on later steps in the sequence without acquiring the prior steps. For example, the repeated measurement of the accuracy in solving division problems of a student who possesses no skills in addition, subtraction, and multiplication would add little to an analysis. Horner and Baer (1978) made this point exceedingly well:

> The inevitable zero scores on the division baseline have no real meaning: division could be nothing else than zero (or chance, depending on the test format), and there is no real point in measuring it. Such measures are *pro forma:* they fill out the picture of a multiple baseline, true, but in an illusory way. They do not so much represent zero behavior as zero opportunity for the behavior to occur, and there is no need to document at the level of well-measured data that behavior does not occur when it cannot. (p. 190)

Thus, the multiple probe design avoids the necessity of collecting ritualistic baseline data when the performance of any component of a chain or sequence is

impossible or unlikely before acquisition of its preceding components. In addition to the two uses already mentioned—analysis of the effects of instruction on complex skill sequences and reduction in the amount of baseline measurement for behaviors that have no plausible opportunity to occur—the multiple probe technique is also an effective experimental strategy for situations in which extended baseline measurement may prove reactive, impractical, or costly. The repeated measurement of a skill under nontreatment conditions can prove aversive to some students; and extinction, boredom, or other undesirable responses can occur. In his discussion of multiple baseline designs, Cuvo (1979) suggested that researchers should recognize that "there is a trade-off between repeatedly administering the dependent measure to establish a stable baseline on one hand and risking impaired performance by subjecting participants to a potentially punishing experience on the other hand" (pp. 222–223). Furthermore, complete assessment of all skills in a sequence may require too much time that could otherwise be spent on instruction.

Other examples of the multiple probe design can be found in Arntzen, Halstadtr, and Halstadtr (2003); Coleman-Martin & Wolff Heller (2004); O'Reilly, Green, and Braunling-McMorrow, (1990); and Werts, Caldwell and Wolery (1996, see Figure 20.6).

Delayed Multiple Baseline Design

The **delayed multiple baseline design** is an experimental tactic in which an initial baseline and intervention are begun, and subsequent baselines are added in a staggered or delayed fashion (Heward, 1978). Figure 9.8 shows a graphic prototype of the delayed multiple baseline design. The design employs the same experimental reasoning as a full-scale multiple baseline design with the exception that data from baselines begun after the independent variable has been applied to previous behaviors, settings, or subjects cannot be used to verify predictions based on earlier tiers of the design. In Figure 9.8 baseline measurement of Behaviors 2 and 3 was begun early

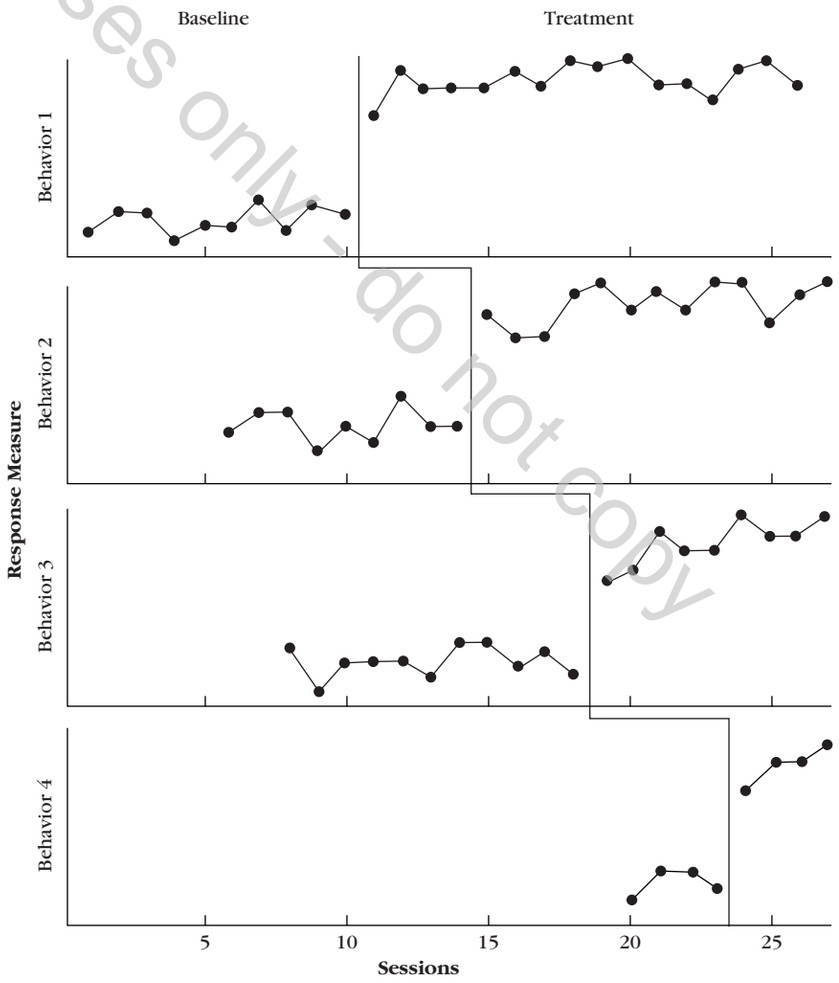

Figure 9.8 Graphic prototype of a delayed multiple baseline design.

enough for those data to be used to verify the prediction made for Behavior 1. The final four baseline data points for Behavior 3 also verify the prediction for Behavior 2. However, baseline measurement of Behavior 4 began after the independent variable had been applied to each of the previous behaviors, thus limiting its role in the design to an additional demonstration of replication.

A delayed multiple baseline design may allow the behavior analyst to conduct research in certain environments in which other experimental tactics cannot be implemented. Heward (1978) suggested three such situations.

- *A reversal design is no longer desirable or possible.* In applied settings the research environment may shift, negating the use of a previously planned reversal design. Such shifts may involve changes in the subject's environment that make the target behavior no longer likely to reverse to baseline levels, or changes in the behavior of parents, teachers, administrators, the subject/client, or the behavior analyst that, for any number of reasons, make a previously planned reversal design no longer desirable or possible. . . . If there are other behaviors, settings, or subjects appropriate for application of the independent variable, the behavior analyst could use a delayed multiple baseline technique and still pursue evidence of a functional relation.
- *Limited resources, ethical concerns, or practical difficulties preclude a full-scale multiple baseline design.* This situation occurs when the behavior analyst only controls resources sufficient to initially record and intervene with one behavior, setting, or subject, and another research strategy is inappropriate. It may be that as a result of the first intervention, more resources become available for gathering additional baselines. This might occur following the improvement of certain behaviors whose pretreatment topography and/or rate required an inordinate expenditure of staff resources. Or, it could be that a reluctant administrator, after seeing the successful results of the first intervention, provides the resources necessary for additional analysis. Ethical concerns may preclude extended baseline measurement of some behaviors (e.g., Linscheid, Iwata, Ricketts, Williams, & Griffin, 1990). Also under this heading would fall the "practical difficulties" cited by Hobbs and Holt (1976) as a reason for delaying baseline measurement in one of three settings.
- *A "new" behavior, setting, or subject becomes available.* A delayed multiple baseline technique might be employed when another research design was originally planned but a multiple baseline analysis becomes the preferred approach due to changes in the environment (e.g., the subject begins to emit another behavior appropriate for intervention with the experimental variable, the subject begins to emit the original target behavior in another setting, or additional subjects displaying the same target behavior become available.) (adapted from pp. 5–6)

Researchers have used the delayed multiple baseline technique to evaluate the effects of a wide variety of interventions (e.g., Baer, Williams, Osnes, & Stokes, 1984; Copeland, Brown, & Hall, 1974; Hobbs & Holt, 1976; Jones, Fremouw, & Carples, 1977; Linscheid et al., 1990; Risley & Hart, 1968; Schepis, Reid, Behrmann, & Sutton, 1998; White & Bailey, 1990 [Figure 15.1]). Poche, Brouwer, and Swearingen (1981) used a delayed multiple baseline design to evaluate the effects of a training program designed to prevent children from being abducted by adults. Three typically developing preschool children were selected as subjects because, during a screening test, each readily agreed to leave with an adult stranger. The dependent variable was the level of appropriateness of self-protective responses emitted by each child when an adult suspect approached the child and attempted to lure her away with a simple lure ("Would you like to go for a walk?"), an authoritative lure ("Your teacher said it was alright for you to come with me."), or an incentive lure ("I've got a nice surprise in my car. Would you like to come with me and see it?").

Each session began with the child's teacher bringing the child outdoors, then pretending to have to return to the building for some reason. The adult suspect (a confederate of the experimenters but unknown to the child) then approached the child and offered one of the lures. The confederate also served as observer, scoring the child's response on a 0 to 6 scale, with a score of 6 representing the desired response (saying, "No, I have to go ask my teacher" and moving at least 20 feet away from the suspect within 3 seconds) and a score of 0 indicating that the child moved some distance away from the school building with the suspect. Training consisted of modeling, behavioral rehearsal, and social reinforcement for correct responses.

Figure 9.9 shows the results of the training program. During baseline, all three children responded to the lures with safety ratings of 0 or 1. All three children mastered correct responses to the incentive lure in one to three training sessions, with one or two more sessions required for each child to master correct responses to the other two lures. Overall, training took approximately 90 minutes per child distributed over five or six sessions. All three children responded correctly when the lures were administered in generalization probes on sidewalk locations 150 to 400 feet from the school.

Although each baseline in this study was of equal length (i.e., had an equal number of data points), contradicting the general rule that the baselines in a multiple baseline design should vary significantly in length, there are two good reasons that Poche and colleagues began training when they did with each subject. First, the nearly total stability of the baseline performance of each child provided an ample basis for evaluating the training program

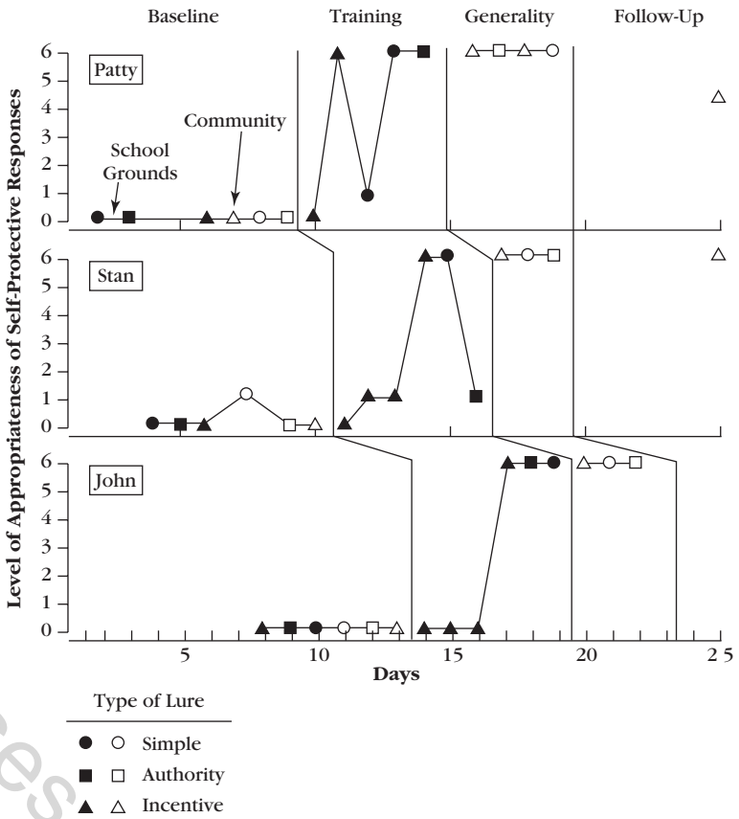

Figure 9.9 A delayed multiple baseline design showing the level of appropriateness of self-protective responses during baseline, training, and generality probes in school and community settings. Closed symbols indicate data gathered near the school; open symbols, in a location away from the school.

From "Teaching Self-Protection to Young Children" by C. Poche, R. Brouwer, and M. Swearingen, 1981, *Journal of Applied Behavior Analysis, 14,* p. 174. Copyright 1981 by the Society for the Experimental Analysis of Behavior, Inc. Reprinted by permission.

(the only exception to complete susceptibility to the adult suspect's lures occurred when Stan stayed near the suspect instead of actually going away with him on his fourth baseline observation). Second, and more important, the nature of the target behavior required that it be taught to each child as soon as possible. Although continuing baseline measurement for varying lengths across the different tiers of any multiple baseline design is good practice from a purely experimental viewpoint, the ethics of such a practice in this instance would be highly questionable, given the potential danger of exposing the children to adult lures repeatedly while withholding training.

The delayed multiple baseline design presents several limitations (Heward, 1978). First, from an applied standpoint the design is not a good one if it requires the behavior analyst to wait too long to modify important behaviors, although this problem is inherent in all multiple baseline designs. Second, in a delayed multiple baseline design there is a tendency for the delayed baseline phases to contain fewer data points than are found in a standard multiple baseline design, in which all baselines are begun simultaneously, resulting in baseline phases of considerable and varying length. Long baselines, if stable, provide the predictive power that permits convincing demonstrations of experimental control. Behavior analysts using any type of multiple baseline design must be sure that all baselines, regardless of when they are begun, are of sufficient and varied length to provide a believable basis for comparing experimental effects. A third limitation of the delayed multiple baseline design is that it can mask the interdependence of dependent variables.

> The strength of any multiple baseline design is that little or no change is noticed in the other, as yet untreated, behaviors until, and only until, the experimenter applies the independent variable. In a delayed multiple baseline design, the "delayed baseline" data gathered for subsequent behaviors may represent changed performance due to the experimental manipulation of other behaviors in the design and, therefore, may not be representative of the true, preexperimental operant level. . . . In such instances, the delayed multiple baseline might result in a "false negative," and the researcher may erroneously conclude that the intervention was not effective on the subsequent behavior(s), when in reality the lack of simultaneous baseline data did not permit the discovery that the behaviors covaried. This is a major weakness of the delayed multiple baseline design and makes it a research tactic of second choice whenever a full-scale multiple baseline can be employed. However, this limitation can and should be combated whenever possible by beginning subsequent baselines at least several sessions prior to intervention on previous baselines. (Heward, 1978, pp. 8–9)

Both the multiple probe design and the delayed multiple baseline design offer the applied behavior analyst alternative tactics for pursuing a multiple baseline analysis when extended baseline measurement is unnecessary, impractical, too costly, or unavailable. Perhaps the most useful application of the delayed multiple baseline technique is in adding tiers to an already operational multiple baseline design. Whenever a delayed baseline can be supplemented by probes taken earlier in the course of the study, experimental control is strengthened. As a general rule, the more baseline data, the better.

Assumptions and Guidelines for Using Multiple Baseline Designs

Like all experimental tactics, the multiple baseline design requires the researcher to make certain assumptions about how the behavior–environment relations under investigation function, even though discovering the existence and operation of those relations is the very reason for conducting the research. In this sense, the design of behavioral experiments resembles an empirical guessing game—the experimenter guesses; the data answer. The investigator makes assumptions, hypotheses in the informal sense, about behavior and its relation to controlling variables and then constructs experiments designed to produce data capable of verifying or refuting those conjectures.[3]

Because verification and replication in the multiple baseline design depends on what happens, or does not happen, to other behaviors as a result of the sequential application of the independent variable, the experimenter must be particularly careful to plan and carry out the design in a manner that will afford the greatest degree of confidence in any relations suggested by the data. Although the multiple baseline design appears deceptively simple, its successful application entails much more than selecting two or more behaviors, settings, or subjects, collecting some baseline data, and then introducing a treatment condition to one behavior after the other. We suggest the following guidelines for designing and conducting experiments using multiple baseline designs.

Select Independent, yet Functionally Similar, Baselines

Demonstration of a functional relation in a multiple baseline design depends on two occurrences: (a) the behavior(s) still in baseline showing no change in level, variability, or trend while the behavior(s) in contact with the independent variable changes; and (b) each behavior changes when, and only when, the independent variable has been applied to it. Thus, the experimenter must make two, at times seemingly contradictory, assumptions about the behaviors targeted for analysis in a multiple baseline design. The assumptions are that the behaviors are functionally independent of one another (the behaviors will not covary with one another), and yet the behaviors share enough similarity that each will change when the same independent variable is applied to it (Tawney & Gast, 1984). An error in either assumption can result in a failure to demonstrate a functional relation.

For example, let us suppose that the independent variable is introduced with the first behavior, and changes in level and/or trend are noted, but the other behaviors still in baseline also change. Do the changes in the still-in-baseline behaviors mean that an uncontrolled variable is responsible for the changes in all of the behaviors and that the independent variable is an effective treatment? Or do the simultaneous changes in the untreated behaviors mean that the changes in the first behavior were affected by the independent variable and have generalized to the other behaviors? Or, let us suppose instead that the first behavior changes when the independent variable is introduced, but subsequent behaviors do not change when the independent variable is applied. Does this failure to replicate mean that a factor other than the independent variable was responsible for the change observed in the first behavior? Or does it mean only that the subsequent behaviors do not operate as a function of the experimental variable, leaving open the possibility that the change noted in the first behavior was affected by the independent variable?

Answers to these questions can be pursued only by further experimental manipulations. In both kinds of failure to demonstrate experimental control, the multiple baseline design does not rule out the possibility of a functional relation between the independent variable and the behavior(s) that did change when the variable was applied. In the first instance, the failure to demonstrate experimental control with the originally planned design is offset by the opportunity to investigate and possibly isolate the variable robust enough to change multiple behaviors simultaneously. Discovery of variables that reliably produce

[3] *Hypothesis,* as we are using the term here, should not be confused with the formal hypothesis testing models that use inferential statistics to confirm or reject a hypothesis deduced from a theory. As Johnston and Pennypacker (1993a) pointed out, "Researchers do not need to state hypotheses if they are asking a question about nature. When the experimental question simply asks about the relation between independent and dependent variables, there is no scientific reason to make a prediction about what will be learned from the data" (p. 48). However, Johnston and Pennypacker (1980) also recognized that "more modest hypotheses are constantly being subjected to experimental tests, if only to establish greater confidence in the details of the suspected controlling relations. Whenever an experimenter arranges to affirm the consequent of a particular proposition, he or she is testing a hypothesis, although it is rare to encounter the actual use of such language [in behavior analysis]. Hypothesis testing in this relatively informal sense guides the construction of experiments without blinding the researcher to the importance of unexpected results" (pp. 38–39).

generalized changes across behaviors, settings, and/or subjects is a major goal of applied behavior analysis; and if the experimenter is confident that all other relevant variables were held constant before, during, and after the observed behavior changes, the original independent variable is the first candidate for further investigation.

In the second situation, with its failure to replicate changes from one behavior to another, the experimenter can pursue the possibility of a functional relation between the independent variable and the first behavior, perhaps using a reversal technique, and seek to discover later an effective intervention for the behavior(s) that did not change. Another possibility is to drop the original independent variable altogether and search for another treatment that might be effective with all of the targeted behaviors.

Select Concurrent and Plausibly Related Multiple Baselines

In an effort to ensure the functional independence of behaviors in a multiple baseline design, experimenters should not select response classes or settings so unrelated to one another as to offer no plausible means of comparison. For the ongoing baseline measurement of one behavior to provide the strongest basis for verifying the prediction of another behavior that has been exposed to an independent variable, two conditions must be met: (a) The two behaviors must be measured concurrently, and (b) all of the relevant variables that influence one behavior must have an opportunity to influence the other behavior. Studies that employ a multiple baseline approach across subjects and settings often stretch the logic of the design beyond its capabilities. For example, using the stable baseline measures of one child's compliance with parental requests as the basis for verifying the effect of intervention on the compliance behavior of another child living with another family is questionable practice. The sets of variables influencing the two children are surely differentiated by more than the presence or absence of the experimental variable.

> There are some important limits to designating multiple behavior/setting combinations that are intended to function as part of the same experiment. In order for the use of multiple behaviors and settings to be part of the same design and thus augment experimental reasoning, the general experimental conditions under which the two responses (whether two from one subject or one from each of two subjects) are emitted and measured must be ongoing concurrently. . . . Exposure [to the independent variable] does not have to be simultaneous for the different behavior/setting combinations, [but] it must be the identical treatment conditions along with the associated extraneous variables that impinge on the two responses and/or settings. This is because the conditions imposed on one behavior/setting combination must have the *opportunity* of influencing the other behavior/setting combination at the same time, regardless of the condition that actually prevails for the second. . . . It follows that using responses of two subjects each responding in different settings would not meet the requirement that there be a coincident opportunity for detecting the treatment effect. A treatment condition [as well as the myriad other variables possibly responsible for changes in the behavior of one subject] could not then come into contact with the responding of the other subject, because the second subject's responding would be occurring in an entirely different location. . . . Generally, the greater the plausibility that the two responses would be affected by the single treatment [and all other relevant variables], the more powerful is the demonstration of experimental control evidenced by data showing a change in only one behavior. (Johnston and Pennypacker, 1980, pp. 276–278)

The requirements of concurrency and plausible influence must be met for the verification element of baseline logic to operate in a multiple baseline design. However, replication of effect is demonstrated each time a baseline steady state is changed by the introduction of the independent variable, more or less regardless of where or when the variable is applied. Such nonconcurrent and/or unrelated baselines can provide valuable data on the generality of a treatment's effectiveness.[4]

This discussion should not be interpreted to mean that a valid (i.e., logically complete) multiple baseline design cannot be conducted across different subjects, each responding in different settings. Numerous studies using mixed multiple baselines across subjects, responses classes, and/or settings have contributed to the development of an effective technology of behavior change (e.g., Dixon et al., 1998; Durand, 1999 [see Figure 23.4]; Ryan, Ormond, Imwold, & Rotunda, 2002).

Let us consider an experiment designed to analyze the effects of a particular teacher training intervention, perhaps a workshop on using tactics to increase each student's opportunity to respond during group instruction. Concurrent measurement is begun on the frequency of student response opportunities in the classrooms of the teachers who are participating in the study. After stable

[4]A related series of A-B designs across different behaviors, settings, and/or participants in which each A-B sequence is conducted at a different point in time is sometimes called a *nonconcurrent multiple baseline design* (Watson & Workman, 1981). The absence of concurrent measurement, however, violates and effectively neuters the experimental logic of the multiple baseline design. Putting the graphs of three A-B designs on the same page and tying them together with a dogleg dashed line might produce something that "looks like" a multiple baseline design, but doing so is of questionable value and is likely to mislead readers by suggesting a greater degree of experimental control than is warranted. We recommend describing such a study as a series or collection of A-B designs and graphing the results in a manner that clearly depicts the actual time frame in which each A-B sequence occurred with respect to the others (e.g., Harvey, May, & Kennedy, 2004, Figure 2).

baselines have been established, the workshop is presented first to one teacher (or group of teachers) and eventually, in staggered multiple baseline fashion, to all of the teachers.

In this example, even though the different subjects (teachers) are all behaving in different environments (different classrooms), comparison of their baseline conditions is experimentally sound because the variables likely to influence their teaching styles operate in the larger, shared environment in which they all behave (the school and teaching community). Nevertheless, whenever experiments are proposed or published that involve different subjects responding in different settings, researchers and consumers should view the baseline comparisons with a critical eye toward their logical relation to one other.

Do Not Apply the Independent Variable to the Next Behavior Too Soon

To reiterate, for verification to occur in a multiple baseline design, it must be established clearly that as the independent variable is applied to one behavior and change is noted, little or no change is observed in the other, as-yet-untreated behaviors. The potential for a powerful demonstration of experimental control has been destroyed in many studies because the independent variable was applied to subsequent behaviors too soon. Although the operational requirement of sequential application in the multiple baseline tactic is met by introduction of the independent variable even in adjacent time intervals, the experimental reasoning afforded by such closely spaced manipulations is minimal.

> The influence of unknown, concomitant, extraneous variables that might be present could still be substantial, even a day or two later. This problem can be avoided by demonstrating continued stability in responding for the second behavior/setting combination during and after the introduction of the treatment for the first combination until a sufficient period of time has elapsed to detect any effect on the second combination that might appear. (Johnston & Pennypacker, 1980, p. 283)

Vary Significantly the Lengths of Multiple Baselines

Generally, the more the baseline phases in a multiple baseline design differ in length from one another, the stronger the design will be. Baselines of significantly different lengths allow the unambiguous conclusion (assuming an effective treatment variable) that each behavior not only changes when the independent variable is applied, but also that each behavior does not change until the independent variable has been applied. If the different baselines are of the same or similar length, the possibility exists that changes noted when the independent variable is introduced are the result of a confounding variable, such as practice or reactivity to observation and measurement, and not a function of the experimental variable.

> Those effects . . . called practice, adaptation, warm-up, self-analysis, etc.; whatever they may be and whatever they may be called, the multiple baseline design controls for them by systematically varying the length of time (sessions, days, weeks) in which they occur prior to the introduction of the training package. . . . Such control is essential, and when the design consists of only two baselines, then the number of data points in each prior to experimental intervention should differ as radically as possible, at least by a factor of 2. I cannot see not systematically varying lengths of baselines prior to intervention, and varying them as much as possible/practical. Failure to do that . . . weakens the design too much for credibility. (D. M. Baer, personal communication, June 2, 1978)

Intervene on the Most Stable Baseline First

In the ideal multiple baseline design, the independent variable is not applied to any of the behaviors until steady state responding has been achieved for each. However, the applied behavior analyst is sometimes denied the option of delaying treatment just to increase the strength of an experimental analysis. When intervention must begin before stability is evident across each tier of the design, the independent variable should be applied to the behavior, setting, or subject that shows the most stable level of baseline responding. For example, if a study is designed to evaluate the effects of a teaching procedure on the rate of math computation of four students and there is no a priori reason to teach the students in any particular sequence, instruction should begin with the student showing the most stable baseline. However, this recommendation should be followed only when the majority of the baselines in the design show reasonable stability.

Sequential application of the independent variable should be made in the order of greatest stability at the time of each subsequent application. Again, however, the realities of the applied world must be heeded. The social significance of changing a particular behavior must sometimes take precedence over the desire to meet the requirements of experimental design.

Considering the Appropriateness of Multiple Baseline Designs

The multiple baseline design offers significant advantages, which no doubt have accounted for its widespread use by researchers and practitioners. Those advantages,

however, must be weighed against the limitations and weaknesses of the design to determine its appropriateness in any given situation.

Advantages of the Multiple Baseline Design

Probably the most important advantage of the multiple baseline design is that it does not require withdrawing a seemingly effective treatment to demonstrate experimental control. This is a critical consideration for target behaviors that are self-injurious or dangerous to others. This feature of the multiple baseline design also makes it an appropriate method for evaluating the effects of independent variables that cannot, by their nature, be withdrawn and for investigating target behaviors that are likely or that prove to be irreversible (e.g., Duker & van Lent, 1991). Additionally, because the multiple baseline design does not necessitate a reversal of treatment gains to baseline levels, parents, teachers, or administrators may accept it more readily as a method of demonstrating the effects of an intervention.

The requirement of the multiple baseline design to sequentially apply the independent variable across multiple behaviors, settings, or subjects complements the usual practice of many practitioners whose goal is to develop multiple behavior changes. Teachers are charged with helping multiple students learn multiple skills to be used in multiple settings. Likewise, clinicians typically need to help their clients improve more than one response class and emit more adaptive behavior in several settings. The multiple baseline design is ideally suited to the evaluation of the progressive, multiple behavior changes sought by many practitioners in applied settings.

Because the multiple baseline design entails concurrent measurement of two or more behaviors, settings, or subjects, it is useful in assessing the occurrence of generalization of behavior change. The simultaneous monitoring of several behaviors gives the behavior analyst the opportunity to determine their covariation as a result of manipulations of the independent variable (Hersen & Barlow, 1976). Although changes in behaviors still under baseline conditions eliminate the ability of the multiple baseline design to demonstrate experimental control, such changes reveal the possibility that the independent variable is capable of producing behavioral improvements with desirable generality, thereby suggesting an additional set of research questions and analytic tactics (e.g., Odom, Hoyson, Jamieson, & Strain, 1985).

Finally, the multiple baseline design has the advantage of being relatively easy to conceptualize, thereby offering an effective experimental tactic for teachers and parents who are not trained formally in research methodology (Hall et al., 1970).

Limitations of the Multiple Baseline Design

The multiple baseline design presents at least three scientific limitations or considerations. First, a multiple baseline design may not allow a demonstration of experimental control even though a functional relation exists between the independent variable and the behaviors to which it is applied. Changes in behaviors still under baseline conditions and similar to concurrent changes in a behavior in the treatment condition preclude the demonstration of a functional relation within the original design. Second, from one perspective, the multiple baseline design is a weaker method for showing experimental control than the reversal design. This is because verification of the baseline prediction made for each behavior within a multiple baseline design is not directly demonstrated with that behavior, but must be inferred from the lack of change in other behaviors. This weakness of the multiple baseline design, however, should be weighed against the design's advantage of providing multiple replications across different behaviors, settings, or subjects. Third, the multiple baseline design provides more information about the effectiveness of the treatment variable than it does about the function of any particular target behavior.

> Consistently [the] multiple baseline is less an experimental analysis of the response than of the technique used to alter the response. In the reversal design, the response is made to work again and again; in the multiple-baseline designs, it is primarily the technique that works again and again, and the responses either work once each [if different responses are used] or else a single response works once each per setting or once each per subject. Repetitive working of the same response in the same subject or the same setting is not displayed. But, while repetitive working of the response is foregone, repetitive and diverse working of the experimental technique is maximized, as it would not be in the reversal design. (Baer, 1975, p. 22)

Two important applied considerations that must be evaluated in determining the appropriateness of the multiple baseline design are the time and resources required for its implementation. Because the treatment variable cannot be applied to subsequent behaviors, settings, or subjects until its effects have been observed on previous behaviors, settings, or subjects, the multiple baseline design requires that intervention be withheld for some behaviors, settings, or subjects, perhaps for a long time. This delay raises practical and ethical concerns. Treatment cannot be delayed for some behaviors; their importance makes delaying treatment impractical. And as Stolz (1978) pointed out, "If the intervention is generally acknowledged to be effective, denying it simply to achieve a multiple-baseline design might be unethical" (p. 33). Second, the resources needed for the concurrent measure-

ment of multiple behaviors must be considered. Use of a multiple baseline design can be particularly costly when behavior must be observed and measured in several settings. However, when the use of intermittent probes during baseline can be justified in lieu of continuous measurement (Horner & Baer, 1978), the cost of concurrently measuring multiple behaviors can be reduced.

Changing Criterion Design

The changing criterion design can be used to evaluate the effects of a treatment that is applied in a graduated or stepwise fashion to a single target behavior. The changing criterion design was first described in the applied behavior analysis literature in two papers coauthored by Vance Hall (Hall & Fox, 1977; Hartmann & Hall, 1976).

Operation and Logic of the Changing Criterion Design

The reader can refer to Figure 9.10 before and after reading Hartmann and Hall's (1976) description of the **changing criterion design.**

> The design requires initial baseline observations on a single target behavior. This baseline phase is followed by implementation of a treatment program in each of a series of treatment phases. Each treatment phase is associated with a step-wise change in criterion rate for the target behavior. Thus, each phase of the design provides a baseline for the following phase. When the rate of the target behavior changes with each stepwise change in the criterion, therapeutic change is replicated and experimental control is demonstrated. (p. 527)

The operation of two elements of baseline logic—prediction and replication—is clear in the changing criterion design. When stable responding is attained within each phase of the design, a prediction of future responding is made. Replication occurs each time the level of behavior changes in a systematic way when the criterion is changed. Verification of the predictions based on each phase is not so obvious in this design but can be approached in two ways. First, varying the lengths of phases systematically enables a form of self-evident verification. The prediction is made that the level of responding will not change if the criterion is not changed. When the criterion is not changed and stable responding continues, the prediction is verified. When it can be shown within the design that levels of responding do not change unless the criterion is changed, regardless of the varied lengths of phases, experimental control is evident. Hall and Fox (1977) suggested another possibility for verification: "The experimenter may return to a former criterion and if the behavior conforms to this criterion level there is also a cogent argument for a high degree of behavioral control" (p. 154). Such a reversed criterion is shown in the next-to-last phase of Figure 9.10. Although returning to an earlier criterion level requires a brief interruption of the steady improvement in behavior, the reversal tactic strengthens the analysis considerably and should be included in the changing criterion design unless other factors indicate its inappropriateness.

One way to conceptualize the changing criterion design is as a variation of the multiple baseline design. Both Hartmann and Hall (1976, p. 530) and Hall and Fox (1977, p. 164) replotted data from changing criterion design experiments in a multiple baseline format with each tier of the multiple baseline showing the occurrence or nonoccurrence of the target behavior at one of the criterion levels used in the experiment. A vertical condition change line doglegs through the tiers indicating when the criterion for reinforcement was raised to the level

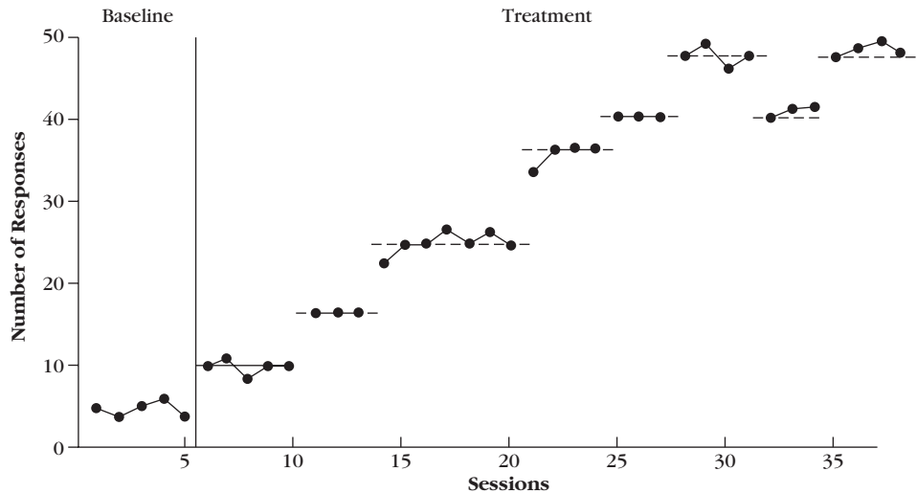

Figure 9.10 Graphic prototype of a changing criterion design.

represented by each tier. By graphing whether the target behavior was emitted during each session at or above the level represented on each tier both before and after the change in criterion to that level, a kind of multiple baseline analysis is revealed. However, the strength of the multiple baseline argument is not quite so convincing because the "different" behaviors represented by each tier are not independent of one another. For example, if a target behavior is emitted 10 times in a given session, all of the tiers representing criteria below 10 responses would have to show that the behavior occurred, and all of the tiers representing criteria of 11 or more would have to show no occurrence of the behavior, or zero responding. The majority of the tiers that would appear to show verification and replication of effect, in fact, could only show these results because of the events plotted on another tier. A multiple baseline design provides its convincing demonstration of experimental control because the measures obtained for each behavior in the design are a function of the controlling variables for that behavior, not artifacts of the measurement of another behavior. Thus, recasting the data from a changing criterion design into a many-tiered multiple baseline format will often result in a biased picture in favor of experimental control.

Even though the multiple baseline design is not completely analogous, the changing criterion design can be conceptualized as a method of analyzing the development of new behaviors. As Sidman (1960) pointed out, "It is possible to make reinforcement contingent upon a specified value of some aspect of behavior, and to treat that value as a response class in its own right" (p. 391). The changing criterion design can be an effective tactic for showing the repeated production of new rates of behavior as a function of manipulations of the independent variable (i.e., criterion changes).

Other than the experiments included in the Hartmann and Hall (1976) and Hall and Fox (1977) papers, there have been relatively few examples of pure changing criterion designs published in the applied behavior analysis literature (e.g., DeLuca & Holborn, 1992 [see Figure 13.2]; Foxx & Rubinoff, 1979; Johnston & McLaughlin, 1982). Some researchers have employed a changing criterion tactic as an analytic element within a larger design (e.g., Martella, Leonard, Marchand-Martella, & Agran, 1993; Schleien, Wehman, & Kiernan, 1981).

Allen and Evans (2001) used a changing criterion design to evaluate the effects of an intervention to reduce the excessive checking of blood sugar levels by Amy, a 15-year-old girl diagnosed with insulin-dependent diabetes about 2 years prior to the study. Persons with this form of diabetes must guard against hypoglycemia (i.e., low blood sugar), a condition that produces a cluster of symptoms such as headaches, dizziness, shaking, impaired vision, and increased heart rate, and can lead to seizures and loss of consciousness. Because hypoglycemic episodes are physically unpleasant and can be a source of social embarrassment, some patients become hypervigilent in avoiding them, checking for low blood sugar more often than is necessary and deliberately maintaining high blood glucose levels. This leads to poor metabolic control and increased risk of complications such as blindness, renal failure, and heart disease.

At home Amy's parents helped her monitor her blood sugar levels and insulin injections; at school Amy checked her blood glucose levels independently. Her physician recommended that Amy keep her blood sugar levels between 75 and 150 mg/dl, which required her to check her blood sugar 6 to 12 times per day. Soon after she had been diagnosed with diabetes, Amy experienced a single hypoglycemic episode in which her blood sugar fell to 40 mg/dl, and she experienced physical symptoms but no loss of consciousness. After that episode Amy began checking her glucose levels more and more often, until at the time of her referral she was conducting 80 to 90 checks per day, which cost her parent approximately $600 per week in reagent test strips. Amy was also maintaining her blood sugar level between 275 to 300 mg/dl, far above the recommended levels for good metabolic control.

Following a 5-day baseline condition, a treatment was begun in which Amy and her parents were exposed to a gradually decreasing amount of information about her blood glucose level. Over a 9-month period Amy's parents gradually reduced the number of test strips she was given each day, beginning with 60 strips during the first phase of the treatment. Allen and Evans (2001) explained the treatment condition and method for changing criteria as follows:

> The parents expressed fears, however, that regardless of the criterion level, Amy might encounter a situation in which additional checking would be necessary. Concerns about adherence to the exposure protocol by the parents resulted in a graduated protocol in which Amy could earn a small number of additional test strips above and beyond the limit set by the parents. One additional test strip could be earned for each half hour of engagement in household chores. Amy was allowed to earn a maximum of five additional tests above the criterion when the criterion was set at 20 test strips or higher. Amy was allowed two additional test strips when the criterion was set below 20. Access to test strips was reduced in graduated increments, with the parents setting criteria to levels at which they were willing to adhere. Criteria changes were contingent upon Amy successfully reducing total test strip use to below the criterion on 3 successive days. (p. 498)

Figure 9.11 shows the criterion changes and the number of times Amy monitored her blood glucose level during the last 10 days of each criterion level. The results

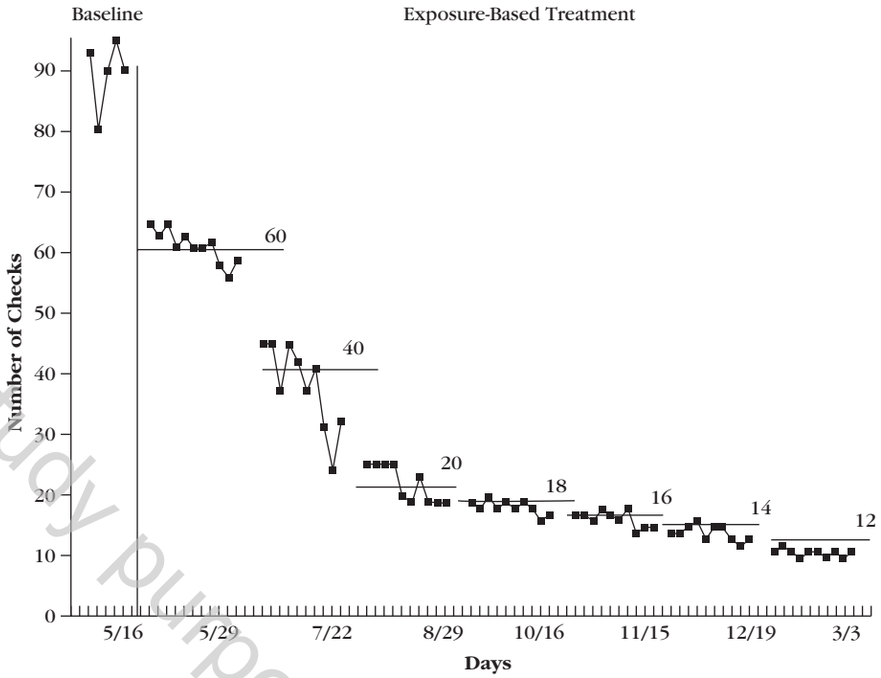

Figure 9.11 A changing criterion design showing the number of blood glucose monitoring checks conducted during the last 10 days of each criterion level. Dashed lines and corresponding numbers indicate the maximum number of test strips allotted at each level. Checks above the criterion levels were conducted with additional test strips earned by Amy.

From "Exposure-Based Treatment to Control Excessive Blood Glucose Monitoring" by K. D. Allen and J. H. Evans, 2001, *Journal of Applied Behavior Analysis, 12,* p. 499. Copyright 2001 by the Society for the Experimental Analysis of Behavior, Inc. Reprinted by permission.

clearly show that Amy responded well to the treatment and rarely exceeded the criterion. Over the course of the 9-month treatment program, Amy reduced the number of times she monitored her blood sugar from 80 to 95 times per day during baseline to fewer than 12 tests per day, a level that she maintained at a 3-month follow-up. Amy's parents indicated that they did not plan to decrease the criterion any further. A concern was that Amy might maintain high blood sugar levels during treatment. The authors reported that her blood sugar levels increased initially during treatment, but gradually decreased over the treatment program to a range of 125 to 175 mg/dl, within or near the recommended level.

Although the figure shows data only for the final 10 days of each criterion level, it is likely that the phases varied in length.[5] The study consisted of seven criterion changes of two magnitudes, 20 and 2. Although greater variation in the magnitude of criterion changes and a return to a previously attained higher criterion level may have provided a more convincing demonstration of experimental control, the practical and ethical considerations of doing so would be questionable. As always, the applied behavior analyst must balance experimental concerns with the need to improve behavior in the most effective, efficient, ethical manner.

This study illustrates very well the changing criterion design's flexibility and is a good example of behavior analysts and clients working together. "Because the parents were permitted to regulate the extent of each criterion change, the intervention was quite lengthy. However, by allowing the parents to adjust their own exposure to acceptable levels, adherence to the overall procedure may have been improved." (Allen & Evans, 2001, p. 500)

Guidelines for Using the Changing Criterion Design

Proper implementation of the changing criterion design requires the careful manipulation of three design factors: length of phases, magnitude of criterion changes, and number of criterion changes.

[5]Data on the number of checks by Amy throughout the intervention are available from Allen and Evans (2001).

Length of Phases

Because each phase in the changing criterion design serves as a baseline for comparing changes in responding measured in the next phase, each phase must be long enough to achieve stable responding. "Each treatment phase must be long enough to allow the rate of the target behavior to restabilize at a new and changed rate; it is stability after change has been achieved, and before introduction of the next change in criterion, that is crucial to producing a convincing demonstration of control" (Hartmann & Hall, 1976, p. 531). Target behaviors that are slower to change therefore require longer phases.

The length of phases in a changing criterion design should vary considerably to increase the design's validity. For experimental control to be evident in a changing criterion design, the target behavior not only must change to the level required by each new criterion in a predictable (preferably immediate) fashion, but also must conform to the new criterion for as long as it is in effect. When the target behavior closely follows successively more demanding criteria that are held in place for varied periods of time, the likelihood is reduced that the observed changes in behavior are a function of factors other than the independent variable (e.g., maturation, practice effects). In most situations, the investigator should not set a predetermined number of sessions for which each criterion level will remain in effect. It is best to let the data guide ongoing decisions whether to extend the length of a current criterion phase or introduce a new criterion.

Magnitude of Criterion Changes

Varying the size of the criterion changes enables a more convincing demonstration of experimental control. When changes in the target behavior occur not only at the time a new criterion is implemented but also to the level specified by the new criterion, the probability of a functional relation is strengthened. In general, a target behavior's immediate change to meet a large criterion change is more impressive than a behavior change in response to a small criterion change. However, two problems arise if criterion changes are too large. First, setting aside practical considerations, and speaking from a design standpoint only, large criterion changes may not permit inclusion of a sufficient number of changes in the design (the third design factor) because the terminal level of performance is reached sooner. The second problem is from an applied view: Criterion changes cannot be so large that they conflict with good instructional practice. Criterion changes must be large enough to be detectable, but not so large as to be unachievable. Therefore, the variability of the data in each phase must be considered in determining the size of criterion changes. Smaller criterion changes can be employed with very stable levels of responding, whereas larger criterion changes are required to demonstrate behavior change in the presence of variability (Hartmann & Hall, 1976).

When using a changing criterion design, behavior analysts must guard against imposing artificial ceilings (or floors) on the levels of responding that are possible in each phase. An obvious mistake of this sort would be to give a student only five math problems to complete when the criterion for reinforcement is five. Although the student could complete fewer than five problems, the possibility of exceeding the criterion has been eliminated, resulting perhaps in an impressive-looking graph, but one that is badly affected by poor experimental procedure.

Number of Criterion Changes

In general, the more times the target behavior changes to meet new criteria, the more convincing the demonstration of experimental control is. For example, eight criterion changes, one of which was a reversal to a previous level, were implemented in the changing design illustrated in Figure 9.10, and Allen and Evans (2001) conducted seven criterion changes (Figure 9.11). In both of these cases, a sufficient number of criterion changes occurred to demonstrate experimental control. The experimenter cannot, however, simply add any desired number of criterion changes to the design. The number of criterion changes that are possible within a changing criterion design is interrelated with the length of phases and the magnitude of criterion changes. Longer phases mean that the time necessary to complete the analysis increases; with a limited time to complete the study, the greater the number of phases, the shorter each phase can be.

Considering the Appropriateness of the Changing Criterion Design

The changing criterion design is a useful addition to the behavior analyst's set of tactics for evaluating systematic behavior change. Like the multiple baseline design, the changing criterion design does not require that improvement in behavior be reversed. However, partial reversals to earlier levels of performance enhance the design's capability to demonstrate experimental control. Unlike the multiple baseline design, only one target behavior is required.

Several characteristics of the changing criterion design limit its effective range of applications. The design can be used only with target behaviors that are already in the subject's repertoire and that lend themselves to stepwise modification. However, this is not as severe a limitation as it might seem. For example, students perform

many academic skills to some degree, but not at a useful rate. Many of these skills (e.g., solving math problems, reading) are appropriate for analysis with a changing criterion design. Allowing students to progress as efficiently as possible while meeting the design requirements of changing criterion analysis can be especially difficult. Tawney and Gast (1984) noted that "the challenge of identifying criterion levels that will permit the demonstration of experimental control without impeding optimal learning rates" is problematic with all changing criterion designs (p. 298).

Although the changing criterion design is sometimes suggested as an experimental tactic for analyzing the effects of shaping programs, it is not appropriate for this purpose. In shaping, a new behavior that initially is not in the person's repertoire is developed by reinforcing responses that meet a gradually changing criterion, called successive approximations, toward the terminal behavior (see Chapter 19). However, the changing response criteria employed in shaping are topographical in nature, requiring different forms of behavior at each new level. The multiple probe design (Horner & Baer, 1978), however, is an appropriate design for analyzing a shaping program because each new response criterion (successive approximation) represents a different response class whose frequency of occurrence is not wholly dependent on the frequency of behaviors meeting other criteria in the shaping program. Conversely, the changing criterion design is best suited for evaluating the effects of instructional techniques on stepwise changes in the rate, frequency, accuracy, duration, or latency of a single target behavior.

Summary

Multiple Baseline Design

1. In a multiple baseline design, simultaneous baseline measurement is begun on two or more behaviors. After stable baseline responding has been achieved, the independent variable is applied to one of the behaviors while baseline conditions remain in effect for the other behavior(s). After maximum change has been noted in the first behavior, the independent variable is then applied in sequential fashion to the other behaviors in the design.
2. Experimental control is demonstrated in a multiple baseline design by each behavior changing when, and only when, the independent variable is applied.
3. The multiple baseline design takes three basic forms: (a) a multiple baseline across behaviors design consisting of two or more different behaviors of the same subject; (b) a multiple baseline across settings design consisting of the same behavior of the same subject in two or more different settings; and (c) a multiple baseline across subjects design consisting of the same behavior of two or more different participants.

Variations of the Multiple Baseline Design

4. The multiple probe design is effective for evaluating the effects of instruction on skill sequences in which it is highly unlikely that the subject's performance on later steps in the sequence can improve without instruction or mastery of the earlier steps in the chain. The multiple probe design is also appropriate for situations in which prolonged baseline measurement may prove reactive, impractical, or too costly.
5. In a multiple probe design, intermittent measurements, or probes, are taken on all of the behaviors in the design at the outset of the experiment. Thereafter, probes are taken each time the subject has achieved mastery of one of the behaviors or skills in the sequence. Just prior to instruction on each behavior, a series of true baseline measures are taken until stability is achieved.
6. The delayed multiple baseline design provides an analytic tactic in situations in which (a) a planned reversal design is no longer desirable or possible; (b) limited resources preclude a full-scale multiple baseline design; or (c) a new behavior, setting, or subject appropriate for a multiple baseline analysis becomes available.
7. In a delayed multiple baseline design, baseline measurement of subsequent behaviors is begun sometime after baseline measurement was begun on earlier behaviors in the design. Only baselines begun while earlier behaviors in the design are still under baseline conditions can be used to verify predictions made for the earlier behaviors.
8. Limitations of the delayed multiple baseline design include (a) having to wait too long to modify certain behaviors, (b) a tendency for baseline phases to contain too few data points, and (c) the fact that baselines begun after the independent variable has been applied to earlier behaviors in the design can mask the interdependence (covariation) of behaviors.

Assumptions and Guidelines for Using Multiple Baseline Designs

9. Behaviors comprising multiple baseline designs should be functionally independent of one another (i.e., they do not covary) and should share a reasonable likelihood that each will change when the independent variable is applied to it.

10. Behaviors selected for a multiple baseline design must be measured concurrently and must have an equal opportunity of being influenced by the same set of relevant variables.

11. In a multiple baseline design, the independent variable should not be applied to the next behavior until the previous behavior has changed maximally and a sufficient period of time has elapsed to detect any effects on behaviors still in baseline conditions.

12. The length of the baseline phases for the different behaviors comprising a multiple baseline design should vary significantly.

13. All other things being equal, the independent variable should be applied first to the behavior showing the most stable level of baseline responding.

14. Conducting a reversal phase in one or more tiers of a multiple baseline design can strengthen the demonstration of a functional relation.

Considering the Appropriateness of Multiple Baseline Designs

15. Advantages of the multiple baseline design include the fact that (a) it does not require withdrawing a seemingly effective treatment, (b) sequential implementation of the independent variable parallels the practice of many teachers and clinicians whose task is to change multiple behaviors in different settings and/or subjects, (c) the concurrent measurement of multiple behaviors allows direct monitoring of generalization of behavior change, and (d) the design is relatively easy to conceptualize and implement.

16. Limitations of the multiple baseline design include the fact that (a) if two or more behaviors in the design covary, the multiple baseline design may not demonstrate a functional relation even though one exists; (b) because verification must be inferred from the lack of change in other behaviors, the multiple baseline design is inherently weaker than the reversal design in showing experimental control between the independent variable and a given behavior; (c) the multiple baseline design is more an evaluation of the independent variable's general effectiveness than an analysis of the behaviors involved in the design; and (d) conducting a multiple baseline design experiment requires considerable time and resources.

Changing Criterion Design

17. The changing criterion design can be used to evaluate the effects of a treatment on the gradual or stepwise improvement of a behavior already in the subject's repertoire.

18. After stable baseline responding has been achieved, the first treatment phase is begun, in which reinforcement (or punishment) is usually contingent on the subject's performing at a specified level (criterion). The design entails a series of treatment phases, each requiring an improved level of performance over the previous phase. Experimental control is demonstrated in the changing criterion design when the subject's behavior closely conforms to the gradually changing criteria.

19. Three features combine to determine the potential of a changing criterion design to demonstrate experimental control: (a) the length of phases, (b) the magnitude of criterion changes, and (c) the number of criterion changes. The believability of the changing criterion design is enhanced if a previous criterion is reinstated and the subject's behavior reverses to the level previously observed under that criterion.

Considering the Appropriateness of the Changing Criterion Design

20. The primary advantages of the changing criterion design are that (a) it does not require a withdrawal or reversal of a seemingly effective treatment, and (b) it enables an experimental analysis within the context of a gradually improving behavior, thus complementing the practice of many teachers.

21. Limitations of the changing criterion design are that the target behavior must already be in the subject's repertoire, and that incorporating the necessary features of the design may impede optimal learning rates.