CHAPTER

29

Research Methods

Behavior Analyst Certification Board 4th Edition Task List Items

A-08 Assess and interpret interobserver agreement.

B-03 Systematically arrange independent variables to demonstrate their effects on dependent variables.

B-04 Use withdrawal/reversal designs.

B-05 Use alternating treatments (i.e., multielement) designs

B-06 Use changing criterion designs.

B-07 Use multiple baseline designs.

I-04 Design and implement the full range of functional assessment procedures.

J-08 Select intervention strategies based on the social validity of the intervention.

FK-01 Lawfulness of behavior.

FUNDAMENTALS

In this chapter, we summarize a few of the issues we've dealt with and add a few more. We'll do so as we give you our answers to three questions: Why should we do behavior analysis? How should we do behavior analysis? And how should we evaluate behavior analysis?

WHY SHOULD WE DO BEHAVIOR ANALYSIS?

TO UNDERSTAND THE WORLD

In this book, we've emphasized the practical side of the science of behavior analysis—we've stressed its contributions to improving life in the universe. We've done this for several reasons: That's the goal of most science and technology of behavior analysis; that's the easiest to understand; and that's what students are initially most interested in. In fact, this is our slogan:

Save the world with behavior analysis.[1]

But there's also a theoretical side. There's also the notion that science, including behavior analysis, is of value in its own right. More precisely, there's the notion that our scientific understanding of people, the world, and the universe is of value regardless of whether it helps us save the world. According to this view, science, including behavior analysis, is like art and music. Even if it doesn't contribute much to saving the world, it makes the world a better place to live in and thus a world more worth saving. Just as we have art for art's sake, so we have scientific knowledge for knowledge's sake.

We also advocate the value of pure knowledge. One of the things we like most about behavior analysis is that it gives us insights and understanding concerning the behavior (the psychological nature) of human beings and of our cousins in the animal kingdom. So we have another slogan:

Understand the world with behavior analysis.

We should do behavior analysis for two reasons: to save the world and to understand the world.

TO BUILD A BETTER WORLD

We said the goal of humanity should be the well-being of life in the universe. **The well-being of the universe should also be the goal of behavior analysis**, as it should for every profession. So we should concentrate on designing systems that make people happy, healthy, and productive—that maximize human potential for contributing to the well-being of the universe.

We are behavior-change professionals; we analyze and deal with problems that might affect a person, group, community, ecosystem, or future generations. We ought to understand behavior and what maintains it. We ought to change behaviors that harm the universe. We ought to maintain behaviors that help the universe. We ought to use our profession to build a better world.

We don't have to look too hard to find areas where behavior analysts could make a difference. Just think close to home for a minute: your parents, your spouse, your children, your friends, your boss, your dog, yourself. Can you see anything wrong there? Can you find anything worth improving?

Would you work to improve the quality of life—physical health, psychological well-being? Would you add quality years to people's lives by helping them follow healthy diets and exercise programs? Would you reduce hassles between people—between mothers and daughters, between brothers and sisters, between husbands and wives, between workers and employers, between government and citizens, between colleagues, between friends? (We even have hassles within single people; many of us hassle ourselves more than we do anyone else.) Would you contribute to a better environment by getting rid of water and air pollution? Would you reduce conflicts between countries? Would you prevent starvation, illness, and wars (both nuclear and conventional)? Behavior analysts have developed and tested an effective behavioral change technology; that's what this book is all about. Now it's time to continue with the extension and testing of this technology in these broader areas we've just mentioned. So one way to build a better world is to use behavior analysis as one of the building tools.

QUESTION

1. What are the two main goals or values of science, including behavior analysis?

HOW SHOULD WE DO BEHAVIOR ANALYSIS?

THE LAW OF EFFECT (FK-01)

Why do we behave as we do? Behavior analysis suggests one major answer:

Definition: REVIEW PRINCIPLE

The law of effect
- The effects of our actions
- determine whether we will repeat them.

In the hands of behavior analysts, this basic principle differentiates behavior analysis from other approaches to psychology. For behavior analysts, the causes of behavior are not in the unconscious mind, not in the self, not in cognitions, and not in a supernatural power. Rather, the causes of behavior are in the environmental contingencies; the immediate effects or outcomes of behavior help determine the future occurrence of that behavior.

REINFORCEMENT

if behavior occurs often, an event, stimulus, or condition has reinforced that behavior. Behavior can be reinforced by the presentation of a reinforcer, by the removal of an aversive condition, or by either the prevention of an aversive condition or the prevention of the removal of a reinforcer. For example, Dina is an avid guitarist. What maintains Dina's guitar playing? The following table presents four contingencies that could maintain this behavior. Dina's guitar playing (1) causes an audience to enter the room to listen to the melody, (2) reduces her muscle tension and stress, (3) prevents her brother's crying, and (4) prevents the audience from leaving the room. (See the first table on p. 419.)

PUNISHMENT

If Dina isn't playing the guitar, why not? Here are four alternatives: (1) The behavior isn't in her repertoire; Dina might not play the guitar because she hasn't learned how or because her guitar playing has been extinguished. (2) There is a physical constraint; she broke her hand. 3) She doesn't have a prerequisite skill; she can't tune the guitar. (4) Something might punish guitar playing.

Guitar playing might be punished by the contingent (1) presentation of an aversive condition, (2) removal of a reinforcer, (3) prevention of the presentation of a reinforcer, or (4) prevention of the removal of an aversive condition.

The second table on p. 419 shows the four contingencies that might punish Dina's guitar playing. Her playing might (1) produce untuned, aversive sounds, (2) cause the audience to leave the room, (3) prevent her from hearing the beautiful sounds of the birds singing outside the window, or (4) prevent an unpleasant visitor from leaving her house.

The Reinforcement Contingencies

Behavioral Relationship	Outcome	Contingency	Description of the Contingency
Presentation	**Reinforcer:** attentive audience	Reinforcement	**Presents a reinforcer:** attentive audience enters the room
Removal	**Aversive condition:** tense muscles, stress	Escape	**Removes an aversive condition:** terminates her tension and stress
Prevention of presentation	**Aversive condition:** brother's crying	Avoidance	**Prevents the presentation of an aversive condition:** prevents her brother's crying
Prevention of removal	**Reinforcer:** attentive audience	Avoidance	**Prevents the removal of a reinforcer:** prevents attentive audience from leaving the room

The Punishment Contingencies

Behavioral Relationship	Outcome	Contingency	Description of the Contingency
Presentation	**Aversive condition:** untuned aversive guitar sound	Punishment	**Presents an aversive condition:** listening to an untuned guitar
Removal	**Reinforcer:** attentive audience	Penalty	**Removes a reinforcer:** attentive audience leaves the room
Prevention of presentation	**Reinforcer:** singing birds	Prevention of reinforcement	**Prevents the presentation of a reinforcer:** can't hear the birds singing
Prevention of removal	**Aversive condition:** unpleasant visitor	Prevention of escape	**Prevents the removal of an aversive condition:** unpleasant visitor stays in the house to listen to her music

The Direct-Acting Behavioral Contingencies

Behavioral	Outcome	Contingency	Description of the	Frequency of

Relationship			Contingency	behavior
Presentation	Reinforcer	Reinforcement	Presents a reinforcer	Increase
	Aversive condition	Punishment	Presents an aversive condition	Decrease
Removal	Reinforcer	Penalty	Removes a reinforcer	Decrease
	Aversive condition	Escape	Removes an aversive condition	Increase
Prevention of presentation	Reinforcer	Punishment by preventing a reinforcer	Prevents the presentation of a reinforcer	Decrease
	Aversive condition	Avoidance	Prevents the presentation of an aversive condition	Increase
Prevention of removal	Reinforcer	Avoidance of loss of a rein-forcer	Prevents the removal of a reinforcer	Increase
	Aversive condition	Punishment by preventing the removal of an aversive condition	Prevents the removal of an aversive condition	Decrease
None		Extinction	Stops the reinforcement contingency	Decrease[2]
		Recovery	Stops the punishment contingency	Increase

This third table on p. 419 summarizes all the direct-acting behavioral contingencies we've studied in this book; it's a combination of the previous two tables. Check it out.

DIRECT-AC TING AND INDIRECT-AC TING CONTINGENCIES

Look at the immediate consequences of behavior: directacting contingencies of reinforcement and punishment. **Indirect-acting contingencies** are those contingencies involving delayed outcomes. These contingencies do not themselves reinforce or punish the behavior. Only **direct-acting contingencies** reinforce and punish behavior. **Direct-acting contingencies** involve immediate, sizable, and probable outcomes.

So what controls behavior when the contingency of interest involves delayed, small, or improbable outcomes? As we saw before, lurking around somewhere there must be a direct-acting contingency controlling behavior. This is where the concept of *rule* comes into play. A rule is a description of a behavioral contingency; we think it acts as a verbal analog to a pairing procedure. The rule statement establishes noncompliance with the rule as an aversive condition. Compliance with the rule is reinforced by a reduction of an aversive condition. This direct-acting escape contingency is what controls behavior and not the outcome specified in the indirect-acting contingency.

For example, what controls writing your term paper? A stimulus in the environment evokes stating a rule, such as, "If I don't start preparing my term paper now, I'll fail the course." This statement establishes goofing off as an aversive condition, like "fear" or "anxiety." Can you relate to this aversive condition? Working on your term paper escapes or reduces that aversive condition and therefore reinforces your working. It is this direct-acting escape contingency that reinforces working on your term paper, not the long-term outcome of success in the course.

HOW TO DO A FUNCTIONAL ASSESSMENT (I-04)

As we mentioned in Chapter 3, behavior analysts nowadays ask what maintains an inappropriate behavior before trying to eliminate it, or what prevents an appropriate behavior before trying to reinforce it. They ask, "What

function does the problem behavior serve for the client?" Perhaps a less purposive way of putting it would be, "What is the functional (causal) relation between the behavior and its reinforcing outcomes?" We can also ask such similar questions when appropriate behavior doesn't occur often enough; for instance, "What is the functional relation between the behavior and its punishing outcome?" Another way to put it is, "What are the effective natural contingencies maintaining inappropriate behavior and suppressing appropriate behavior?" The systematic, empirical search for an answer to such questions, before implementing a behavioral intervention, is called a **functional assessment.**[3] (Incidentally, functional assessments may have the drawback of focusing too much attention on effective, natural competing contingencies and away from the possibility that the natural contingencies for appropriate behavior may simply be ineffective, even in the absence of competing contingencies.)

Definition: REVIEW CONCEPT

Functional Assessment
- An analysis
- of the contingencies responsible for
- behavioral problems.

Note that you usually do this assessment before the intervention.

The table on p. 421 titled *The Steps of a Functional Assessment shows how we recommend doing a functional assessment.* The answers to the questions in each step provide information about the contingencies responsible for behavioral problems. This table shows the relations among the steps of the function analysis listed in the preceding table.

To illustrate the six steps of a functional assessment, we'll use a semi-hypothetical example from our old autism practicum in the Pre-primary Impaired Classroom of Kalamazoo's former Croyden Avenue School. Because most functional assessments are of the contingencies controlling inappropriate behavior, this example will deal with the disruptive behavior of Jason, a preschool autistic child doing discrete-trial training.

1. **Behavior**. Jason has a wide variety of disruptive behavior (e.g., getting out of his seat, grabbing instructional materials, lying on the floor, taking his clothes off, leaving the study booth). Appropriate behavior consists of doing the discrete-trail training tasks. He has such a high rate of disruptive behavior that the practicum students are sometimes reduced to tears; he often spends more time disrupting than doing his learning tasks.
2. **Before Condition**. Jason is doing a difficult task and he is not getting nearly as much attention, nor nearly such dynamic attention as when he's disrupting.
3. **After Condition**. What is the immediate outcome? Usually, there are two: He escapes the discrete trial task, which may be difficult and thus aversive. And the tutor chases him, grabs him, and coaxes him to sit down and do his tasks, which may be forms of social reinforcement.
4. **Motivating Operation.** What is the motivating operation? For escape contingencies, we don't distinguish between the motivating operation and the before condition, so it's just his doing the difficult tasks, which may be aversive. And for the social reinforcement contingency, we usually don't deal with the motivating operation for learned reinforcers (e.g., social reinforcers like attention); so we can bypass the motivating operation.
5. **S^D and S^Δ** What are the S^D and S^Δ We notice that Jason disrupts much more frequently with inexperienced techs than with experienced techs. So, we suspect the inexperienced techs might be the S^D and the experienced techs might be the S^Δ. In other words, whether the relevant contingency is escape from the difficult task or inadvertent social reinforcement, the experienced tech might be better at extinguishing the disruptive behavior, at being an effective S^Δ for disruption.
6. **Contingency**. What contingency maintains the disruptive behavior? As we've said, it might be escape from an aversively difficult task and/or social reinforcement with attention.

The Steps of a Functional Assessment—What to Identify	
1. Behavior	**What is the behavior of interest?**

	What are some instances? What are some noninstances? (other behaviors that occur instead) How often does the behavior occur during baseline?
2. Before condition	**What is the before condition?** Is it a reinforcer, aversive condition, or neutral condition? (may be more than one)
3. After condition	**What is the after condition?** Is it a reinforcer, aversive condition, or neutral condition? (may be more than one) Is the after condition sizable? Is the after condition probable? Is the after condition immediate? (if not, the maintaining contingency isn't direct acting) What is the long-term outcome?
4. Motivating operation	**What is the motivating operation?**
5. Discriminative stimulus	**What is the discriminative stimulus, if there is one?**
6. Contingency	**If there is a direct-acting contingency, what is it? (it may help to diagram it)** Reinforcement (presentation of a reinforcer) Punishment (presentation of an aversive condition) Penalty (removal of a reinforcer) Escape (removal of an aversive condition) Avoidance (prevention of the presentation of an aversive condition) Punishment by the prevention of the presentation of a reinforcer Avoidance of the removal of a reinforcer Punishment by the prevention of the removal of an aversive condition Or is there no direct-acting contingency (i.e., extinction)?
	If any, what type of analog contingency exists between the behavior and the after condition? (What type of indirect-acting contingency is in effect?)

Without doing functional assessments, we had typically assumed the disruptive behavior was escape maintained and had not been very sensitive to the possibility of the accidental social reinforcement we might be providing as we attempted to prevent the escape. With the consulting of Pamela Osnes, we began to complete functional assessments showing that the disruptive behavior of our various little Jasons was often a result of social reinforcement.

Also, as a result of Osnes' consulting, we are attempting to increase the density of social and physical reinforcers our Jasons receive when they are working on their tasks or at least not disrupting. Our goal is to have the children almost swimming in reinforcers, as long as they're cool. And this seems effective, regardless of whether the disruptive behavior is escape maintained or attention maintained. If it's attention maintained, Jason is getting so much attention while he's on task, that it's hardly worth his time to disrupt and go off task. If it's escape maintained, the addition of all those reinforcers while he's doing the aversively difficult tasks, will counteract that aversiveness and reduce the frequency of his disruptive escape behavior.

SHOULD WE DO FUNCTIONAL ASSESSMENT?

Many behaviorists argue we need to analyze behavior to select the intervention. But is this really the case? Why do we have to do this analysis? Couldn't we use a penalty contingency when the undesirable behavior occurs too often, or a reinforcement contingency when the behavior occurs too rarely? Do we need to analyze the contingencies maintaining behavior? Do we need to do a functional assessment?

Doug's mother, Shirley, shouts at him every time he turns up the volume on the TV because then she can't concentrate on her studies. His undesirable behavior maintains at a high frequency in spite of her shouts. Without doing a functional assessment, Shirley uses an analog to a penalty contingency. She follows the rule that for every time Doug turns up the volume on the TV, he'll miss 5 minutes of *The Simpsons*, his favorite TV program. This will probably work. Shirley will get rid of Doug's disruptive behavior. And she wouldn't need to do a functional assessment.

However, we might come up with several intervention options if we did do a functional assessment. For example, Doug might have a hearing impairment. So a hearing aid might be all he needs. Or earphones for Doug

and earplugs for Shirley might be ideal. Or a simple extinction procedure where Shirley doesn't say anything when Doug raises the volume on the TV might work, if Shirley's shouting is what reinforces Doug's turning up the volume. (It might sound unlikely to most people not familiar with behavior analysis that what at first appears to be an aversive condition [shouting] is really a reinforcer under some conditions.)

A functional analysis might suggest a more humane and equally effective intervention than the analog to the penalty contingency; and it might suggest an intervention more likely to transfer across settings (a hearing aid) or easier to maintain (stereo earphones) than a penalty contingency (Shirley is likely to let a punishment contingency fall through the cracks as she gets preoccupied with other problems). Generally, functional analyses might suggest interventions more cost-effective, less intrusive, more humane, more transferable, and more maintainable than interventions not planned from such analyses.

QUESTIONS

1. What is the law of effect?
2. Define functional assessment.
3. List the six questions you should ask when doing a functional assessment. Give an example of each question.
4. Should you do a functional assessment before or after designing an intervention? Explain.

HOW SHOULD WE EVALUATE BEHAVIOR ANALYSIS? (PRELIMINARIES)

So far, we've reviewed our recommendation for the goal of humanity—the well-being of life in the universe. And we've suggested this should be also the goal of behavior analysts. More generally, we said the goals or values of science, including behavior analysis, are to save the world and to understand how the world works. Also, we've studied how to do a functional assessment so that we can design better interventions to help save the world. Now we'll review techniques to determine whether the problems we're tackling are socially significant, whether they will really contribute to the saving of the world. In addition, we will review techniques to determine whether our interventions are having any impact at all on the world, or at least on small parts of the world.

Concept

SOCIAL VALIDITY (J-08)

Social validity tells us whether we have selected what our clients consider socially significant **target behaviors** —*those worthy of improving*. **Social validity** also tells us whether we have an acceptable intervention, to some extent regardless of its outcome.

Definition: CONCEPT

Social validity
- The goals,
- procedures, and
- results of an intervention
- are socially acceptable to
- the client,
- the behavior analyst, and
- society.[4]

Target behavior
- The behavior being measured,
- the dependent variable.

Van Houten pointed out that we should use social validity procedures to determine the target behavior, optimal levels of the target behavior, and intervention results.[5] We'll fail if the intervention stops below the lower limit the community expects for competent performance. A behavioral intervention might improve Jim's grades from *Es* to *Ds*, but the community might not consider *Ds* good enough. Though we, behavior analysts, might feel our intervention had succeeded, social validity might show the contrary.

Two methods commonly used to estimate social validity are social comparison and subjective evaluation of experts. **Social comparison** consists of comparing the performance of clients exposed to the intervention with a comparable or "normal" group. The intervention is successful if the subjects' performance during intervention is like that of the comparison group. Another method, **subjective evaluation of experts**, involves evaluators who have expertise or familiarity with the client.

Definition: CONCEPTS

Social comparison
- A comparison of the performance of clients
- exposed to the intervention
- with an equivalent or "normal" group.

Subjective evaluation of experts
- Experts' evaluation
- of the significance of the target behavior and the outcome.

Social comparison and subjective evaluation of experts have limitations.[6] In social comparisons, normative data might have little relevance for the functioning of the client; and subjective evaluation of experts offers no guarantee that the experts' judgments are the best criteria to determine the success of the intervention. Using both methods at the same time may overcome these problems.

A complete social validity will also tell us how acceptable our intervention is. For example, even though Jim may have gotten straight As, we've failed if our intervention required more work from the teachers than they believed appropriate.

QUESTIONS

1. What is social validity?
2. Describe two social validity methods and their limitations.

Concept

OBTRUSIVE AND UNOBTRUSIVE OBSERVATION

How can we discover the problem behavior? *Observation* is the answer. But we must avoid contaminating our observations by using obtrusive assessment procedures. Kazdin pointed out that most behavioral assessment procedures are obtrusive, and they affect the results independent of the intervention.[7]

Definition: CONCEPT

Obtrusive assessment
- Measuring performance
- when the person is aware

- of the ongoing observation.

Unobtrusive assessment

- Measuring performance
- when the person is not aware
- of the ongoing observation.

Behaviors assessed under obtrusive and unobtrusive conditions sometimes bear little relation to each other. For instance, observe the behavior of students in a grade-school classroom when the teacher is in front of it with wide-open, intimidating eyes. You'll notice a quiet and dedicated team. However, chaos reigns when the teacher leaves.

You could use unobtrusive direct observations by delegating the observation to a regular participant in the client's environment, observing behavior behind a twoway mirror, videotaping with a hidden camera, or evaluating products of behavior.

Products of behavior are usually unobtrusive assessment methods. Examples are records of behavior and physical proofs. **Records of behavior** could include number of arrests, attendance lists, or number of times a client was readmitted to the hospital. **Physical proof** is direct evidence that the behavior has occurred—the direct result of the behavior. For instance, to study academic behavior, we could review written exercises and essays. To study cigarette smoking, we could count cigarettes left in a client's cigarette pack, after agreeing that cigarettes smoked during the day would come only from that pack. To study dieting, we could weigh the client once a week to see if generally he'd been following the diet.

This example shows the difference between record and physical proof. You might have a performance contract to write two pages every day. At the end of each day, you would record your pages written on a graph; and you'd show your cumulative graph once a week to your performance manager. That's a record of behavior. But you might know that you have a bit of the sleazeball tendency; you don't trust yourself. So you also commit to showing the actual, dated pages. That's physical proof. But both are products of behavior.

Products of behavior are not direct measures of behavior; they involve the result of the behavior rather the behavior itself. However, we use products of behavior for purpose of convenience or unobtrusive assessment.[8]

Definition: CONCEPTS

Products of behavior

- Record or evidence
- that the behavior has occurred.

QUESTIONS

1. Obtrusive assessment procedures—define them and give an example.
2. Unobtrusive assessment procedures—define them and give an example.
3. Products of behavior—define them and give an example.

Concept

UNITS OF MEASUREMENT

The basic units of measurement are frequency, duration, and force. A violation of your diet won't make you fat, but a violation once a day, or even once a week, might. Smoking a cigarette won't kill you, but 20 a day for 20 years might. You can survive a hassle with a friend once, but a hassle every time you open your mouth might end the relationship. Throughout this book, we've studied many examples that involve measuring frequency—for instance, Todd's frequency of bowel movements per week, Jimmy's percentage of disruptive escape responses in 10-second intervals, Dawn's frequency of nail biting per day, Bob's frequency of aggressive episodes.

Duration consists of the length of time the response lasts, like the number of minutes Rod's crying lasted. **Force** means intensity of response. We studied a few examples: increasing Melanie's voice intensity and the force (in grams) with which Rudolph the Rat pressed a lever.

Definition: CONCEPTS

Duration
- The time from
- the beginning
- to the end
- of a response.

Force
- Intensity of a response.

QUESTION

1. Define and give an example of each of the following behavior measures:
a. frequency or rate
b. duration
c. force

Example of Measurement Organizational Behavior Management

SAMUEL DEE'S TOY FACTORY

In this semi-hypothetical case study, Samuel Dee has been CEO for 3 years of PlasToy, Inc., a small plastic toy manufacturing company. Orders had increased during the last few months, and the sales forecast for the rest of the year was higher than ever. Though the sales department celebrated, the production department was not coping. Many shipments were late, and customers were looking for other suppliers who could deliver orders on the due date.

PlasToy had 10 molding presses and three 8-hour shifts. The pressroom was scheduled to work 24 hours a day, 5 days a week. S. Dee wasn't sure what to do. Everyone told him that PlasToy needed more presses, but S. Dee couldn't afford to buy more presses. S. Dee wanted to assess how much his presses were used before deciding what to do. He called the quality control inspectors of each shift and asked them to walk through the pressroom every hour and record whether or not each press was working.

After a week of data collection, S. Dee found that all 10 presses combined were working only 72 hours a day on the average, including three shifts. The daily averages were 12 hours for first shift, 36 for the second, and 24 for the third. But 72 hours were too few, considering that there were 240 possible production hours a day (8 hours per shift × 10 presses × 3 shifts = 240 hours). Seventy-two hours was only 30% of PlasToy's capacity. S. Dee thought, this must be an unusual week. So he asked the inspectors to continue collecting data for the next several weeks; but the data consistently showed that presses were running only 30% of the time. And, surprisingly, the second shift ran more press hours than the other two shifts; this made no sense. S. Dee could think of nothing that would make the second shift more productive than the other two shifts.

Suspicious of the data, he checked it out himself. For a few days he came to work during second shift, from 3:00 to 11:00 p.m. He went to the control office, his preferred place in the plant. He had designed that office in his plant manager's days. From there, he could observe the whole pressroom without being noticed. At the sound of the hourly beep on his watch, he would look through the window and record how many presses were running.

The next day, S. Dee compared records. The second-shift inspector counted 30 hours running time, but S. Dee counted only 20. So he analyzed the two sets of data like this: If for a given hour both agreed the press was running, he wrote **RUN**; if both agreed the press was down, he wrote *Down*; if they disagreed on whether the press was running or not, he wrote *Dis*. He recorded agreements and disagreements as shown in the table on p. 425.

In the table, you can see that S. Dee and the secondshift inspector agreed Press 1 was running at 3:00 p.m. and Press 2 was down at 8:00 p.m. They disagreed as to whether Press 4 was running or down at 6 p.m. S. Dee found that only 48% of his observations agreed with the second-shift inspector's.

Why was agreement between observation records so low? In reviewing his notes, S. Dee noticed that the second-shift inspector wrote **RUN** when a production worker was getting a press ready to mold. For instance, at 6:00 p.m. the materials handler was cleaning the hopper of Press 2 before filling it up with plastic; the second-shift inspector wrote **RUN** at that time, even though the press was off. S. Dee wrote *Down*. He wrote **RUN** only if the press was producing toys. He also found that, on some days, the second-shift inspector walked the pressroom every half hour during the first 4 hours, and she would stay in her office the last 4 hours of the shift studying for school. But she identified each record as an hourly record. Of course, this would also create disagreements between her observations and S. Dee's.

Press	3 P.M.	4 P.M.	5 P.M.	6 P.M.	7 P.M.	8 P.M.	9 P.M.	10 P.M.
1	**RUN**	**RUN**	Dis	Dis	**RUN**	**RUN**	**RUN**	**RUN**
2	Dis	Dis	Dis	Dis	Dis	*Down*	*Down*	*Down*
3	**RUN**	**RUN**	Dis	Dis	Dis	Dis	**RUN**	**RUN**
4	*Down*	*Down*	*Down*	Dis	Dis	Dis	**RUN**	**RUN**
5	*Down*	*Down*	Dis	Dis	*Down*	Dis	Dis	Dis
6	Dis	Dis	Dis	**RUN**	**RUN**	Dis	Dis	Dis
7	*Down*	*Down*	*Down*	*Down*	Dis	Dis	**RUN**	Dis
8	Dis	Dis	Dis	Dis	Dis	Dis	Dis	Dis
9	*Down*	*Down*	*Down*	*Down*	*Down*	*Down*	*Down*	*Down*
10	Dis	Dis	Dis	Dis	**RUN**	Dis	Dis	**RUN**

So, S. Dee called the inspectors and explained what "press running" meant. He asked them to write **RUN** only when the press was producing parts, not when someone was working near it, removing a mold, washing the press, or trimming parts molded another day. He also reviewed carefully the observation procedures. Every hourly check must start on the hour, and it should take no more than 10 minutes. He also said that someone else would be recording press running time independently for comparison.

S. Dee trained the maintenance department inspectors to track press-running time, exactly in the same way as the quality control inspectors. Maintenance and quality control inspectors were never to see each other's observation records. Then S. Dee compared records every day by shift, and 99% of the observations agreed; he trusted the data.

He learned that presses were active about 25% of the time, and in general, there were no differences in performance between the three shifts. He also learned that the production department was very inefficient. He was positive that with performance management he could make a significant improvement in production. There would be no need to buy more presses. What a relief!

Concept

INTEROBSERVER AGREEMENT (A-08)

S. Dee found that their initial measure of press running was not reliable. He found this by using a method called **interobserver agreement.**

Definition: CONCEPT

Interobserver agreement

- Agreement between
- observations of
- two or more independent observers.

Interobserver agreement forces us to define the critical behavior or products of behavior so clearly that two or more observers can get the same observational records independently. The second-shift inspector had a different definition of press running. Of course, her records and S. Dee's differed. But this difference reflected only a poor definition of "press running."

Sometimes, observers are affected by personal biases. For instance, press-running data might be affected by how much the second-shift inspector liked the second-shift supervisor. If she thought a high record would benefit the supervisor, she might have reported higher press-running times than a neutral observer would have.

Reviewing the criteria for recording observations only once is not enough. Observers tend to change their definition of the behavior over time. So variation in the data may be caused by changes in the observer's definition of the observation criterion rather than actual changes of the dependent variable. That's why it's important to review the observation criteria periodically during data collection.

Interobserver agreement also forces us to clarify procedures. For instance, S. Dee had to tell the secondshift inspector that recording was to be done at the beginning of each hour, because it had turned out that the inspector was making her observations at any convenient time within the hour. Also, S. Dee was measuring once during each of the 8 hours of the shift, while the inspector was only measuring during the first 4 hours. So, of course, their interobserver agreement was low.

By the way, S. Dee made sure that he and the supervisor made independent observations, even though they were observing the same thing at the same time. In other words, neither of them could see what the other was writing on their recording forms, otherwise, the supervisor might be tempted, either consciously or unconsciously to kiss up to S. Dee, by always writing down the same thing S. Dee wrote, even though she sometimes disagreed with his observations. Interobserver agreement is meaningless if the observers aren't independent.

To compute the 48% interobserver-agreement ratio, S. Dee used the **point-by-point agreement ratio**, one of the most commonly used methods in applied behavior analysis research to estimate interobserver agreement. This ratio indicates whether we can trust observation records. In the table, you can see that S. Dee agreed with the inspector that presses were running 17 hours (add all the **RUN** records) and presses were down 21 hours (add all the *Down* records). They disagreed on 42 records (add all the *Dis* records). With this information, S. Dee computed the point-by-point ratio by following these steps:

1. He added all the agreements; 17 RUN hours plus 21 Down hours equaled 38 hours.
2. Then he computed all possible instances; he added all 38 agreement records to 42 disagreement and got 80 possible observations for second shift.
3. Finally, he computed the point-by-point agreement ratio; he divided 38 total agreement records by 80 possible records and multiplied by 100. This was 47.5% or, to round it up, 48%.

The second-shift inspector and S. Dee's initial observations agreed only 48% of the time. In other words, their observation records were not reliable; he could not trust the data. Normally, behavior analysts consider data reliable only when the agreement is over 80%. Furthermore, you can end this section assured that S. Dee and the supervisor eventually achieved interobserver agreement well above the 80% goal, as a result of his efforts to get agreement on the measurement criteria and the measurement procedures. Getting good interobserver reliability isn't easy, but it is essential, before you intervene; otherwise, you can't evaluate the success of your intervention. And you can also end this section assured that once he had good objective measures of the press operator's performance, S. Dee was able to reliably measure that performance and implement a pay-for-performance intervention that eliminated most late shipments, increased customer satisfaction, fattened the worker's take-home checks, and removed the necessity of buying expensive new presses, the commonsense solution everyone had originally suggested. Win, win, win.

Objectivity vs. Subjectivity in Technology

In Chapter 13, we introduced the concepts of subjective and objective measurement, arguing that objective measurement is essential for our science to achieve a reliable understanding of the physical and psychological

world. Furthermore, objectivity is also essential for applied technology, such as S. Dee's organizational-behavior-management efforts. If S. Dee had just trusted his subjective opinion of whether or not the presses were being fully utilized, he might have bankrupted the company by unnecessarily buying expensive, new presses. Unfortunately, many CEOs have too much confidence in their own subjective evaluations, with the possibility of disastrous results for their companies; instead, they need to use objective measures with high interobserver agreement, before making important decisions.

QUESTIONS

1. Interobserver agreement—define it and give an example.
2. Explain how you would compute an interobserveragreement ratio using the point-by-point agreement procedure and be able to do the computation, given a set of interobserver data.

HOW SHOULD WE EVALUATE BEHAVIOR ANALYSIS? (RESEARCH DESIGN)

Example

THE SHAMAN AND THE BACKACHE

Vitoria Sousa's husband ran a small grocery store in Santa Cruz das Flores, a Portuguese town of 500 people. Vitoria and her family lived above the store in a twostory Portuguese-colonial building.

Vitoria was 35 but felt like 55. Her weight had increased so much through the years she didn't look like the girl in her wedding picture on her vanity table. At the time of her wedding, she dreamed about how great the future would be. Now she dreamed about how great the past had been. Life hadn't worked out as she had hoped.

A few days after her 35th birthday she had given birth to Orlando, her ninth child. Vitoria hoped Orlando would be the last. She had also had a few miscarriages. Vitoria felt guilty because she hadn't looked forward to the birth of Orlando as she had Carlos, her oldest son. She no longer dreamed about the newborn's future during the 9 months. Pregnancy was just another burden. She just got pregnant and had babies. Her beautiful fantasies became worries: "Can we pay for school, clothes, shoes, the doctor? Will we manage with another child under this roof?"

One day, soon after Orlando was born, Vitoria bent over to pick him up. As she lifted the infant, she let out a horrible scream that brought her husband running up the stairs from the grocery store. He arrived to see her bent double, barely holding Orlando, her mouth twisted, her brow perspiring.

"What's the matter?" he asked.

After a minute, Vitoria muttered, "It's my back. It hurts so much … and I can't stand up."

Vitoria's husband took Orlando and then helped Vitoria hobble to the bed.

The next day, Matilde, her sister-in-law came with the news: "Everybody in town is talking about a shaman who has supernatural powers. He's a priest who uses magic to cure any illness. He looks at you and can tell what's wrong. He cures you in one visit. Let's fix your back. It only costs 10 escudos." Matilde continued with a long list of people who had been cured of blindness, cancer, ulcers, unhappiness—you name it. A back pain had to be a cinch!

Matilde and Vitoria drove to Vila do Conde, a 3-hour trip through the mountains. Vitoria was in pain all the way. In spite of Matilde's long, wrinkled page of directions, they got lost more than once. But they knew when they had arrived: A line of waiting people started in the road and ended in a humble house with a wretched garden. Vitoria and Matilde went to the end of the line, waited for hours, and shared experiences. Everybody told the history of their pains and illness. The shaman was their last hope, and all but the most skeptical knew it would work.

Matilde went first; she wanted the shaman to cure her recurring migraines. She returned with a smile from ear to ear. "Your turn," Matilde said.

The room was 15 feet by 15 feet with no windows, no door (just an open entrance), no chairs, and a dirt floor. An altar was against the south wall. It was a large, wooden table holding crude statues of the Virgin Mary, Jesus Christ, and Saint Joseph. The flowers, incense, and lighted candles filling the rest of the table reminded Vitoria of the familiar fragrance of her church in Santa Cruz das Flores. The walls were covered with

overlapping posters and religious paintings. The 70-year-old shaman, with torn, dirty clothes, sat on the dirt floor in front of the altar. Vitoria was shaking and still bent with pain.

The shaman asked how she was feeling. "I've had a back pain for the last few days that's killing me; I can't stand it any more," Vitoria said.

The shaman replied, "Don't worry; after today, you'll feel no more pain." He asked her to kneel down and close her eyes.

He touched her forehead and said, "Relax, you will be cured within the week." With his hand still on her forehead, he whispered a prayer. Vitoria felt relaxed and transported. Then the shaman stood up and walked around her continuing with his whispered prayer.

He touched her lower back and asked, "Is the pain here?"

After Vitoria nodded yes, he said, "I can feel your pain."

After more prayer, he said, "Soak in a hot tub when you get home, and do so once a week. Also walk at least 15 minutes every day and you will no longer suffer pain." He paused a minute and added, "It's 10 escudos."

Vitoria was slow in her reaction. She took the money from her purse. The shaman placed it in a 2-liter jar, nearly full of cash—today's fees. Vitoria hobbled out, a smile of confidence on her face.

The shaman was right. Vitoria stopped having pain and became another one of his supporters.

Concept

CASE STUDY

This was quite a deal for Vitoria, as you can imagine if you've ever been bent over with such lower severe back pain that you had to rest your elbows on the edge of the sink when you washed your hands. Vitoria was amazed at the cure. Yet you will never read of this shamanistic spiritual/medical intervention in the *New England Journal of Medicine*. Why not? For many reasons, as you can no doubt see. But even if we did have a careful independent measure of Vitoria's back pain, we still would have troubles with confounding variables.

Definition: REVIEW CONCEPT

Confounded variables

- Two or more possible independent variables have changed at the same time;
- so it is not possible to determine which of those variables caused the changes in the dependent variable.

For example, Vitoria's back pain might have gone away in a few days anyhow. This means the wise man's spiritual intervention is confounded with the normal relaxing of the muscle spasms correlated with the passage of time. So this normal relaxing of the muscles is an independent variable (possible cause) that confounds the spiritual intervention. This confounding makes it impossible to be sure the spiritual cure caused the relief from back pain.

That's the trouble with this sort of pre-experimental, anecdotal **case study**. You can never be sure what caused what.

Definition: CONCEPT

Case study

- The evaluation of the results of
- an applied intervention or
- a naturally changing condition
- that involves confounded variables.

A case study is not a poorly designed experiment. Instead, it may be an intervention done to produce particular results. For example, the shaman did not perform his magical intervention to prove that his magic works. Instead, he intervened to remove Victoria's back pain. His intervention combined with the removal of Victoria's back pain becomes a case study when we offer it (perhaps tentatively) as evidence for the power of shamanism.

Similarly, as people in our culture age, they usually get that middle-age midriff. This naturally changing condition does not occur to prove that aging causes weight gain. However, specific instances of this correlation between weight and age becomes a case study when we offer it as evidence of the effect of age on weight. Because it is a simple case study, there is a confounding of variables, for example, age may be correlated with, and thus confounded with, other possible causes such as an increase in calorie consumption or a decrease in calorie-burning exercise.

Incidentally, by *cause*, we typically mean the *independent variable*; and by *effect*, we typically mean the *dependent variable*.

The case study is the most primitive within-subject research design. In the case study, the intervention is often introduced before collecting baseline data, making it practically impossible to rule out confounding variables and to test the effects of the independent variable on the dependent variable. For that reason, we consider the case study to be a pre-experimental research design. Another way researchers talk about this weakness of the study is to say it cannot rule out threats to **internal validity**.

Definition: CONCEPT

Internal validity
- The extent to which a research design
- eliminates confounding variables.

An internally valid study is one in which no more than one independent variable is presented at the same time (not confounded). If you eliminate confounding of independent variables, you are better able to determine what independent variables are responsible for changes in the dependent variable. This should be the goal of all research designs.

Definition: CONCEPT

Research design
- The arrangement of the various conditions of an experiment or intervention
- to reduce the confounding of independent variables.

As we said, the case study is a primitive research design because it does not prevent threats to the internal validity of its research design (confounding of the independent variable), but it can still be of great value. For example, Freud's career was based on his reporting and insightful analyses of clinical case studies. But now it's time to look at more powerful research designs that do a better job of eliminating confounding variables so we can get a clear picture of what's causing what.

Concept

SIMPLE BASELINE DESIGN

Frank was a young man sent to the Psychology Service. He spent many hours slapping his own face, so the staff had restrained him. Fortunately, we collected baseline data on the frequency of the response before intervening.

During eleven 30-minute observation periods without restraint, Frank's frequency of face slapping dropped from over 600 an hour to nearly zero. And we hadn't done anything! This was just baseline.

Definition: REVIEW CONCEPT

Baseline
- The phase of an experiment or intervention
- in which the behavior is measured
- in the absence of an intervention.

Suppose we had given Frank a tranquilizer every day, in the hope that this would get rid of his face slapping. And suppose we had used the drug without getting baseline data first. It would have looked as if the drug had caused the decrease in slapping. Then there would have been a good chance that Frank would have unnecessarily been on that drug for the rest of his life.

The data collected on Frank illustrate an important point in research design: We must collect baseline data before we implement the intervention.

Definition: CONCEPT

Simple baseline design
- An experimental design
- in which the baseline data are collected
- before the intervention.

The **simple baseline design** is better at ruling out confounding variables than is the case study that doesn't involve baseline. The simple baseline design works well if we have a long, stable baseline followed by an abrupt and sizable change in the dependent variable when we change the value of our independent variable. Otherwise, we need a more powerful research design to rule out the threats to internal validity that confounding variables cause.

Concept

REVERSAL DESIGN (B-04)

In Chapter 12, Mae's teachers gave the children toys only when they asked for them using color-noun combinations, for example, green car, red apple.[9] Following this intervention, the frequency of using color-noun combinations increased from 0.4 to 14.2 an hour. But this simple baseline design didn't allow Mae to be completely sure her intervention had increased behavior. Maybe the increase in color-noun combinations would have occurred anyhow. Maybe some confounding variable had really caused the increase in the color-noun frequency. How could Mae find out if her intervention had made the difference?

She asked her teachers to remove the intervention (in other words, to reverse to the original baseline condition). This is called a ***reversal design***. During this reversal to the baseline condition, the children got snacks and materials regardless of whether they used a noun alone or a color-noun combination. Mae kept this reversal to baseline conditions going for the next 18 days. And, sure enough, the frequency of using colornoun combinations decreased to 7.4 per hour.[10] Now Mae was more confident that requiring her students to use the color-noun combinations had increased their frequency. In fact, she was confident enough that she didn't keep the baseline going for longer than 18 days to see if the frequency of color-noun combinations would have eventually dropped to the original baseline frequency of 0.4 per hour.

Definition: CONCEPT

Reversal design

- An experimental design
- in which the intervention (experimental)
- and baseline conditions
- are reversed
- to determine if the dependent variable changes as
- those conditions (independent variable)
- change.

Behavior analysts often call the reversal design an ***ABA design***. *A* stands for the baseline condition and *B* for the intervention. In an ABA design, a second baseline phase follows the intervention. Usually, if the performances in both baseline conditions are similar to each other and different from the intervention, you can be fairly sure that the intervention is what changed the performance. Most likely, you've ruled out any confounding variables. Consecutive ABA replications, like ABABABA, increase the internal validity. The reversal design provides more convincing evidence than the simple baseline design that the intervention is responsible for the change.

But we can't reverse some processes—for example, training processes. If we implement a training program to teach someone to drive a car, when we stop the training program, the natural reinforcers will maintain the person's driving skills without our continued training. So we couldn't use the reversal design to demonstrate effectiveness of our training program. Also, we wouldn't want to use the reversal design when returning to baseline is too dangerous. For instance, suppose we're working with a dangerously self-abusive child. And suppose the frequency of self-abuse goes to near zero during intervention. We wouldn't want to reverse to the dangerous baseline condition to rule out the threats that confounding variables pose for our internal validity. We need some other research design.

Concept

MULTIPLE-BASELINE DESIGN (B-07)

In Chapter 23, Juke helped Red coach his young football team.[11] They worked on three plays: the option play, the power sweep, and the off-tackle counter play. They task-analyzed the three plays into five behavioral components. For example, in the option play, the components included the quarterback-center exchange and the quarterback-right halfback fake. They specified the detailed behavior sequence in each component.

Juke and Red then prepared a checklist of the behavioral components for each of the three plays. Baseline consisted of the traditional training method: The coach gave verbal description of plays and offered suggestions from the sidelines when the team was on the field. The coach also gave feedback on what the players had done correctly and what they could improve. The intervention consisted of the presentation and explanation of the appropriate task-analysis checklist and frequent contingent feedback and recognition for instances of desired play execution.

Suppose Juke had implemented the intervention package with one team at a time, after collecting baseline data for differing durations for each team. This design is known as the **multiple-baseline design** across groups.

Definition: REVIEW CONCEPT

Multiple-baseline design

- An experimental design
- in which the replications involve
- baselines of different durations
- and interventions of different starting times.

In the multiple-baseline design, the intervention could be implemented across behaviors, settings, and groups or subjects. It's like several reversal designs, where the intervention is implemented at different times

across different baselines. If performance consistently changes during the intervention, and across baselines, we can be confident it's the intervention that's causing the change.

Example

DRIVER EDUCATION[12]

About 23% of Utah's high-school students failed their driver ed. course in 1991; and in Utah that meant they couldn't get their driver's license. That's a bigger deal than just losing the chance to cruise for 'burgers. Not being able to drive is correlated with social isolation, dependence, and poverty. Most students (94%) who failed driver ed. failed the written part of the tests. So four behavior analysts at Utah State University—Bell, Young, Salzberg, and West—developed a behavioral training program to help high-risk high school students pass that part of their test.

The program involved highly structured peer-tutoring sessions of 10 minutes each day. A fellow high school student would ask questions about complex driving maneuvers involving backing, turning, and passing. The peer tutor would also model correct answers using verbal descriptions and diagram drawing. Then he or she would encourage the student to practice answering the questions with high speed and accuracy (fluency). The peer tutor would also provide immediate feedback and record the student's performance.

Bart was a 16-year-old male with a normal IQ who had already failed driver ed. once and was going down the tubes a second time. Fortunately, he got to take part in the program. In addition to the tutoring, every day his tutor also gave him a 1-minute speed test on each of the three tasks: backing right, turning left, and passing. The tutor collected several days' baseline data before starting the tutoring and continued his tests throughout the training. Figure 29.1 shows some of Bart's results. It's a bit deep, so take a big breath before you dive in.

One of the measures the tutor took was the number of errors per minute Bart made. Look at the first data point in the top graph. See that during day 1, Bart made 80 errors in his speed test on writing about and diagramming how to back right (each wrong letter and each misplaced line was an error). If you follow the graph, you can see that he made at least 40 errors per minute for the first 7 school days—baseline. Then the 10-minute daily sessions of behaviorally structured peer tutoring started. And Bart's error rate immediately dropped to 8 per minute and quickly went on down to 0 per minute, with one little flare-up on day 13.

So it looks like behavioral peer tutoring really helped Bart. But you can't be sure. Maybe the improvement was only a coincidence. Maybe his noisy sister had just gone off to college, thus allowing Bart to seriously crack the books for the first time.

Bell and his colleagues controlled for that possibility. Look at the middle graph. These behavior analysts kept the baseline for turning left going for two more days before they started the peer tutoring on that problem (peer tutoring for turning left started for the data just to the right of the break in the data). When you compare the two graphs, you can see that the error rate on the turning-left problem didn't go down when the rate went down for the back-right problem. The error rate for turning left didn't go down until the peer tutoring started for that specific problem, day 10.

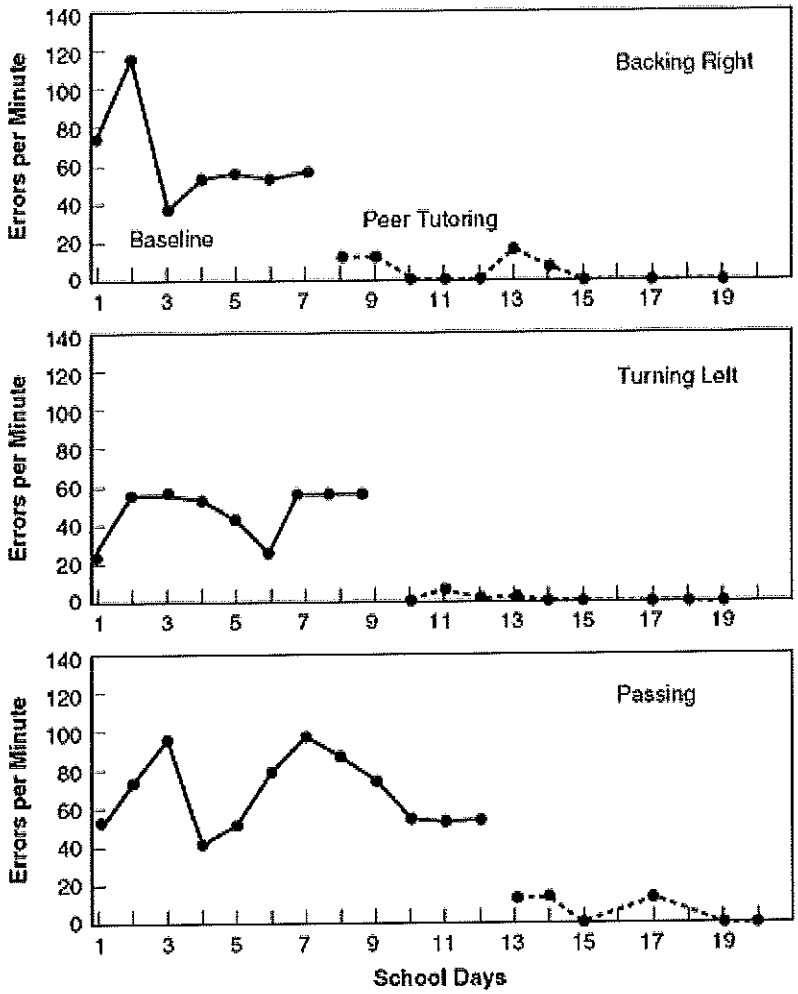

FIGURE 29.1 Behavioral Tutoring & Driver's Ed.

And it's hard to say the sister's leaving (or whatever) on day 8 caused this abrupt improvement in the turning-left graph on day 10. So the behavior analysts could be fairly confident that the peer tutoring was causing the decreased error rates because they had two different baselines that ended on two different days; and they got abrupt changes in performance each time they introduced peer tutoring the day after baseline. In other words, they successfully used a multiple-baseline experimental design.

But they wanted to be even more confident. Maybe poor Bart's second loud sister had gone off to a different college (or whatever) on day 10; not likely, but maybe just another coincidence. So the bottom graph shows their confidence-building third baseline. This baseline (for the passing problem) didn't end until the 12th day. And sure enough, the error rate dropped from 60 on the last day of baseline with no tutoring to 18 on day 13 when the tutoring intervention started. Now, no one could be so unlucky as to have three noisy big sisters, who just happened to go off to college on the 3 days the behavior analysts just happened to start each of their three interventions. Not likely. Nor is it likely that some other positive changes happened coincidentally with each of the three interventions. Surely the three improved performances resulted from the three peer-tutoring interventions.

In summary, for each of these multiple baselines, the error rate was higher on every baseline day than for any of the intervention days. And the error rate dropped way down each time the intervention was started. This excellent use of the multiple-baseline design seems pretty much to eliminate the chance that the results were due to a coincidental change in some other factor rather than due to the behavioral tutoring.

By the way, Bart passed his driver ed. course, passed his driver's test, got his driver's license, and has driven 25 miles to work every day, with no accidents or traffic tickets for the last 8 months. That's better than what a third of his classmates have done. What about you?

Concept

CHANGING-CRITERION DESIGN (B-06)

Every day for the last 2 months you've resolved to start exercising—tomorrow. Yet you've never gotten into exercise clothes, visited the gym, attempted to jog, or done any exercise, except for your regular daily activities, like pushing the buttons on the remote control of your TV. But you really do want to get in shape.

When all else fails, people turn to applied behavior analysis for help. So do you. You call Conscience International.13 With the help of the woman at the other end of the line, you set up a performance (behavior) contract based on an analog to avoidance of the loss of a reinforcer. In addition to a small charge for your daily phone calls, you deposit $200 with Conscience International for your penalty pool. You will lose $5 every weekday you fail to do some sort of exercise in the university gym and report on it to Conscience.

Because you're so out of shape, you think 10 minutes of exercise a day is good enough for openers. So for each day you don't exercise for 10 consecutive minutes, you lose $5. After a few weeks of consistent performance at around 10 minutes a day, 5 days a week, you up the minimum criterion to 20 minutes a day. Now if you continue to exercise for only 10 minutes, you'll lose $5. Every few weeks, you raise your exercise criterion by 10 minutes until you hit 60 minutes a day. And your performance rises to match whatever criterion is in effect on that day. You fail to hit criterion only 3 times, and rarely do you exceed it by more than a few minutes.

After 6 months, you start to wonder if the contracting's worth it. There's the daily charge plus the hassle. Maybe the performance contracting wasn't really crucial to your success. Maybe you should kiss Conscience good-bye and continue to exercise on your own.

But you're nervous. You're on a roll with Conscience International. And you're afraid that if you stop performance contracting and your exercise falls apart, you may need another 5 years before you get your act together enough to give it another shot. You tell your troubles to sympathetic Conscience, and she tells you you've already got a solution. It's called the **changing-criterion design**.

The changing-criterion design is a special type of experimental design in which the experimenter repeats the experiment with differing values of the independent variable and measures changes in the dependent variable. If the values of the dependent variable change systematically with the changes in the independent variable, coincidental confounding variables are probably eliminated. Then we have internal validity, and there is probably a causal or functional relation between the independent and the dependent variables. In the case of the changing criterion design, the value of the independent variable that is being changed is the criterion required to meet a behavioral contingency.

Definition: CONCEPT

Changing-criterion design
- An experimental design
- in which the replications involve
- interventions with criteria of differing values.

And in the case of your performance contract, that criterion being changed is the number of minutes you need to exercise to avoid the $5 loss. Conscience points out that your amount of exercise has varied systematically and abruptly every time you change your criterion. This means that something about the intervention package of your performance contract is crucial. You don't know whether you need all the components—the daily goal, the daily phone call to Conscience, and the daily analog to an avoidance contingency (the potential $5 loss); but you can be confident that at least some of those components are crucial. With your changing-criterion experimental design, you've shown a causal or functional relation between your contracting and your amount of exercise.

Let me add some personal data to these hypothetical data. In my younger days, when I was writing this section for *EPB2e*, I was training to run a marathon and was contracting with Conscience to run either 6, 12, or 18 miles each day. Otherwise, I, too, paid $5. I never ran 18 miles when I could run 12; and I never ran 12 when I could run 6. And I almost never had to pay the $5. This informal changing-criterion design convinced me that my performance in the marathon that spring hung by the fragile thread of my performance contract with Conscience.

All experimental designs involve some sort of comparison between two different conditions, between an experimental condition and a control condition or between various experimental conditions. We just saw that the changing-criterion design involves comparisons of performance between differing response requirements (e.g., differing amounts of exercise required to avoid paying a fine, differing forces of a lever press to get the water reinforcer, or differing percentages of completed homework problems to get an A). If the performance tends to match the different criteria, we are safe in concluding that the contingency we're using is controlling behavior. This is an excellent research design, but it is limited to evaluating the effectiveness of behavioral contingencies. The next design is more general.

Concept

ALTERNATING-TREATMENTS DESIGN (B-05)

Remember David in Chapter 4? He was 21 years old and had lived in an institution for the mentally handicapped for the past 9 years. He showed a high frequency of stereotypic behaviors, such as weaving his head, staring at his hands, and repeatedly manipulating objects. These stereotyped behaviors prevented him from taking part in vocational placement, from learning new skills that could increase his quality of life, and they embarrassed his family.

Jordan, Singh, and Repp (in the research described in Chapter 4) used an **alternating-treatments design** to compare the effectiveness of visual screening, gentle teaching, and baseline conditions. Visual screening consisted of a punishment contingency, in which the trainer covered David's eyes with one hand and held the back of his head with another hand for 5 seconds contingent on the occurrence of a stereotyped behavior. Gentle teaching consisted of using almost no vocal instructions, only gestures and signals. This intervention included a combination of physical guidance, reinforcement of the desirable behavior, and extinction of the undesirable behavior, but no punishment. During baseline, no procedure was in effect.

The experimenters did their research by implementing the three different procedures, in three different 30-minute training sessions, all three on each day of training. In other words, within each day, they alternated between the three experimental conditions—visual screening, gentle teaching, and baseline. Behavior analytic experimenters use such an alternating-treatments design to compare two or more interventions using the same subject.

Definition: CONCEPT

Alternating-treatments design

- An experimental design
- in which the replications involve
- presenting different values of the independent variable
- in an alternating sequence
- under the same general conditions
- or in the same experimental phase,
- while measuring the same dependent variables.

In other words, you go back and forth between two or more specific treatments (values of the independent variable).

To appreciate the value of this experimental design, let's compare it with a couple of others. The experimenters might have used a **between-subjects** or **between-groups design**. They could have used gentle teaching with one client or group of clients, visual screening with a second, and baseline conditions with a third. Then they could have compared the amount of stereotyped self-stimulation among the three clients or groups.

The experimenters couldn't have been sure the differences in self-stimulation were due to their three different experimental conditions (interventions) if they had used three different clients. The problem is that it's too hard to be sure you've really got three equal clients. Maybe the three clients had different rates of self-stimulation to begin with. Or maybe they weren't equally influenced by the interventions. This raises the

possibility that the differences between the results obtained with the three interventions might have been attributed to the differences between the three clients. So it's too hard to rule out the confounding variable of the different clients.

There are a couple of problems with a betweengroups design in which you would randomly assign a large group of clients to each of the three interventions and compare their performances. First, it's hard to get large groups of clients with similar behavioral problems. And second, it's easy to lose sight of the details of the effects of the interventions when your data consist of group averages rather than the performance of individual clients.

Another approach would have been to work only with David, but to use the three conditions successively, perhaps a week of baseline, a week of gentle teaching, followed by a week of visual screening. But then it's not easy to be sure that the 3 weeks were comparable. Maybe 1 week was hotter than the others. Or maybe David would have done less self-stimulation during the 3rd week no matter which condition was in effect. Perhaps his self-stimulation would have decreased as a function of his becoming more familiar with the trainers. In other words, maybe the differences between the three conditions of intervention had nothing to do with the interventions. Such a design would have made it difficult to rule out these and other confounding variables that would have threatened the internal validity of the experiment.

The alternating-treatments design elegantly dealt with all these problems. It ensured that the subjects in the three experimental conditions were alike, by using the same subject— David. And it ensured that the days of exposure to the three conditions were alike, by using all three conditions on the same day (but in a different order each day).

However, the experimenters had to pay a price for using this design. The price was the potential for an **experimental interaction**. Experimental interaction is a risk you always run when you expose an experimental subject to more than one experimental condition. The risk of interaction is that exposure to one condition may have influenced the effects of another condition. For example, the advocates of gentle teaching might argue that gentle teaching would have been much more effective if the experimenters had not also exposed David to the punishment contingency involving visual screening. The risk of this sort of interaction is higher because the experimenters used the same trainers, tasks, and setting with David.

Definition: CONCEPT

Experimental Interaction
- One experimental condition
- affects the results of another.

So, here are advantages of the alternating-treatments design:

- The subject is the same during all treatments.
- The conditions of the treatment are essentially the same.
- Single-subject data can be analyzed, rather than averaged data.

And here is a disadvantage of the alternating-treatments design:

- Experimental interactions cannot be ruled out.

Rarely does a single experiment decide complex theoretical and practical issues such as gentle teaching versus punishment contingencies. We will need many more such careful experiments, conducted by both the advocates of gentle teaching and the advocates of punishment contingencies. But each experiment, with its design strengths and design compromises, moves us closer to an understanding of how the world works and how to help the world work better.

QUESTIONS

1. Define and give an example of the following concepts:
   a. research design
   b. internal validity
   c. experimental interaction

d. interobserver agreement

2. Define and give an example of the following research designs:
   a. case study
   b. simple baseline
   c. reversal
   d. multiple baseline
   e. changing criterion
   f. alternating treatments
3. Why is the case study a pre-experimental design?
4. Give an example of the importance of collecting baseline data. Explain what could happen if we don't.
5. What advantage does the multiple-baseline design have over the reversal design?
6. Give an example of the use of a multiple-baseline design to demonstrate the value of behaviorally based peer tutoring in driver education.
7. What are two advantages and disadvantages of the alternating-treatments design?

GENERALITY OF RESULTS

In this chapter, we've emphasized the importance of doing research that will contribute to the well-being of the universe. And we've pointed out that science in general, and behavior analysis in particular, can contribute in two ways. First, it can contribute as a means toward an end, as a means toward better physical and behavioral health. But science can also contribute as an end in its own right: Knowing how things in the physical, biological, and behavioral worlds work makes it more reinforcing to be alive, just as art, music, and sports do, even if they turn out to be of little utilitarian value.

We've discussed the role of functional analyses in improving the contribution of behavior analysis to universal well-being. And we've discussed the role of research methods in making sure we discover valid cause-effect or functional relationships—internal validity. Finally, let's address the notion of **external validity** or *generality of results*.

External validity means the extent to which the cause-effect relation or functional relation you've shown in your experiment is valid under conditions external to your experiment. External validity means the generality of your results. For example, imagine this:

In the 1930s, a man named Skinner put a small number of simple rats in a small number of simple Skinner boxes, along with a small number of simple behavioral contingencies. He discovered a small number of simple behavioral principles and concepts. No big deal, perhaps. But over the next 70-some years, a few thousand behavior analysts, both scientists and practitioners, discovered that those simple principles and concepts applied to essentially all endeavors of the human and the nonhuman animal. Furthermore, they discovered that the applications of the principles of behavior could improve those endeavors, no matter what those endeavors were. Skinner's results with the rats in his experimental box generalized to mentally handicapped children learning to tie their shoes, to parents trying to coexist peacefully with their children, to doctoral students trying to complete their dissertations, and to corporate presidents trying to make a profit and benefit their employees at the same time. Now that's **external validity**.

Definition: CONCEPT

External validity

- The extent to which the conclusions of an experiment
- apply to a wide variety of conditions.