2

General Principles

Having some specific studies to refer to will help clarify the discussion that follows. Described next, therefore, are two examples of research in developmental psychology. Both studies have been simplified somewhat to make the points drawn from them easier to follow.

Dufresne and Kobasigawa (1989) were interested in the determinants of improvement in memory across childhood. Why do older children generally remember better than younger children? The particular determinant on which the research focused is labeled *study time.* The issue with respect to study time is what children do when they have a chance to study some set of to-be-remembered-items prior to being tested for memory. How much time, for example, do the children spend in studying the material before deciding that they are ready, and how sensibly do they distribute their effort across the various items? Perhaps one reason older children remember better than younger ones is that they make better use of their study time.

The memory task that Dufresne and Kobasigawa selected to test this hypothesis is called a *paired-associates* task. A paired-associates task consists of two phases: an initial phase during which pairs of words are presented together, followed by a second phase during which only one member of the pair is presented and the child must attempt to remember the matching item. In Dufresne and Kobasigawa's study, the pairs were of two sorts: an "easy" set in which the paired items were highly related (e.g., cat-dog, shoe-sock) and a "hard" set in which the items were not related (e.g., book-frog, skate-baby). All of the participants (children from the first, third, fifth, and seventh grades) received both sets of items, and for both they were allowed to study the material for as long as they wished before being tested.

Table 2.1 shows the average study time for each pair of items. Several conclusions are suggested by the values in the table. As would be expected, older children generally studied the items longer than did younger children. As would also be expected, hard items elicited more study time than did easy items. Finally, this easy-hard differentiation was not apparent among the youngest participants; rather, it was only the two older age groups who directed more attention to the hard items. I return to this last finding in particular later in the chapter.

Table 2.1 Mean Study Time (in Seconds) for the Child Participants in the Dufresne and Kobasigawa Study

	Type of item		
Group	*Easy*	*Hard*	*Combined*
Grade 1	5.40	5.20	5.30
Grade 3	5.53	6.96	6.25
Grade 5	4.23	8.42	6.33
Grade 7	4.45	12.48	8.47
Combined	4.90	8.27	

SOURCE: Adapted from "Children's Spontaneous Allocation of Study Time: Differential and Sufficient Aspects," by A. Dufresne and A. Kobasigawa, 1989, *Journal of Experimental Child Psychology, 47*, p. 282. Copyright © 1989, Academic Press.

The second study was also concerned with memory but of a different sort and at a different phase of the life cycle. Cherry and Park (1993) examined memory for spatial locations in samples of young (mean age = 21 years) and older (mean age = 69 years) adults. Participants first viewed a spatial array of 24 common objects. The objects were then scrambled, and the participants' task was to reproduce the original spatial arrangement.

Two presentation conditions were compared. For half of the participants, the objects were presented within the context of a colored, three-dimensional model with a number of distinctive landmarks. For the other half, the background consisted of a two-dimensional, black- and-white map of the three-dimensional model. Two questions were of interest. Would spatial memory be better when the locations to be remembered were embedded within a natural and distinctive context? And (as some studies had suggested) would any facilitative effect of a helpful context be greater for older than for younger participants?

Table 2.2 shows the results. It can be seen that the context did indeed make a difference: Performance was better with the model background than with the map background. Age also made a difference, with the younger participants outperforming the older ones. Finally, although the two age groups differed in overall performance, they did not differ in response to the context manipulation. Both the young adults and the older adults did better with the model than with the map.

Variables

I begin the discussion of general principles with some terminology. Research in psychology involves variables and the relations that hold among variables. The variables are of two sorts: dependent and independent. **Dependent variables** are outcome variables—those measures whose values constitute the results of a study. In the first example the dependent variable was the number of seconds that the child studied each pair of items; in the second example the dependent variable was the number of objects that the adult was able to place correctly. Such variables are dependent in the sense that variation in them follows from or depends on other factors. A central job for the researcher is to

Table 2.2 Mean Number of Items Correctly Placed by the Adult Participants in the Cherry and Park Study

	Context		
Group	*Map*	*Model*	*Combined*
Young	14.5	15.9	15.2
Old	11.1	14.5	12.8
Combined	12.8	15.2	

SOURCE: Adapted from "Individual Differences and Contextual Variables Influence Spatial Memory in Younger and Older Adults," by K. E. Cherry and D. C. Park, 1993, *Psychology and Aging*, 8, p. 520. Copyright © 1993, American Psychological Association.

determine what these other factors are. They are variable necessarily: If there were no possibility of variation in the dependent measure, there would be no point in doing the study.

The dependent variable is something that the researcher measures but does not directly control. **Independent variables**, in contrast, are variables that are under the control of the researcher. The object of the study is to determine whether the particular independent variables chosen do in fact relate to variations in the dependent variable. The independent variables in the Dufresne and Kobasigawa study were the age of the child and the hard-easy contrast, whereas those in the Cherry and Park study were age and type of context. Such variables are independent in the sense that their values are decided on in advance rather than following as results of the study. The "variable" part is again necessary: If there were no variation in the independent variable, there would be no possibility of determining whether that factor has an effect. Variation and comparison are intrinsic parts of all research.

The description of research as divisible into independent and dependent variables is valid for many but not for all studies. Suppose, for example that you wish to know whether there is a relation between a child's IQ and how well that child does in school. You might test a sample of grade-school children and collect two measures: performance on an IQ test and grades in school. Your interest would be in whether variations in one measure relate to variations in the other; for example, do children with high IQs tend to do well in school? A study like this does not have an independent variable whose values are under the experimenter's control; rather, IQ, grades, and the relation between them are all outcome variables in the study. "Correlational" research of this sort is discussed at length later. The point for now is simply that not all studies fit the independent variable–dependent variable mold.

The example studies can serve to illustrate a further point about independent variables. The contrasts that define an independent variable can be created in two ways. One way is through an experimental manipulation that literally creates the variable. This is what Dufresne and Kobasigawa did when they constructed their easy and hard sets of items and what Cherry and Park did when they designed their map and model backgrounds. This was not the approach, however, for the other independent variable in both studies: chronological age. Clearly, investigators cannot create an age contrast in the same way that they can create an easy-hard contrast. In the case of a variable like age, the control occurs not through manipulation but through *selection:* choosing people for

study who are at the desired levels of the variable (e.g., 20 years old or 70 years old). Because selection is the only control possible, age and other "subject variables" can present special problems of interpretation—an issue to which I return later in the chapter.

A bit more terminology is necessary before proceeding. Independent variables are also referred to as **factors**, and the particular values that the variables take are referred to as **levels**. The Dufresne and Kobasigawa study, therefore, can be described as a 4×2 factorial design—that is, an experiment with two factors, one of which (age) has four levels and one of which (condition) has two levels. Similarly, the Cherry and Park study can be described as a 2 (age) $\times$ 2 (condition) factorial design. Note that symbolizing the design in this way serves to tell us the number of distinct cells or groups in the experiment. For example, in the Cherry and Park study there are four (2×2) distinct groups: young adults in the model condition, young adults in the map condition, old adults in the model condition, and old adults in the map condition.

Validity

All research involves variables and the relations that hold among variables. When we wish to describe research, therefore, the construct of variables is central: What kinds of contrasts are being examined, and what forms do the examinations take? When we wish to move beyond description to *evaluation* of research, the central construct becomes that of **validity**. The question of validity is the question of accuracy: Has the study in fact demonstrated what it claims to demonstrate? All of the specific methodological points discussed throughout the book come down to this one basic question of the accuracy of the conclusions that we draw from research.

Various forms of validity can be distinguished (Shadish, Cook, & Campbell, 2002). In this chapter I discuss three forms: *internal, external,* and *construct.* Chapter 8 will add a fourth form: *statistical conclusion validity.*

Internal validity applies within the context of the study itself. The issue in question is whether the independent variables really relate to the dependent variables in the manner claimed. Have we drawn the correct conclusions about the causal impact (or lack of causal impact) of one set of variables on the other set? Let us take the Dufresne and Kobasigawa study as an example. Their conclusions are internally valid if the hard items really did produce longer study times than the easy items, if the average study time really did increase as a function of age, and if the ability to differentiate between easy and hard also really increased with age. If there is a plausible alternative explanation for any of these findings, then the internal validity of the study is thrown in doubt. Suppose, for example, that seventh graders in the study had been selected primarily from "gifted" classes, whereas the younger children represented more average ability levels. If so, we would have an alternative explanation for the seeming improvements with age: The differences reflect not natural changes with age but rather differences in ability level. (I discuss this problem, labeled *selection bias,* more fully later.)

The question of **external validity** is the question of generalizability. It applies, therefore, once we move outside the immediate context of the study. The question now is whether we can generalize the findings of the study to other samples, situations, and behaviors—not just any samples, situations, and behaviors, of course, but those for which we wish the study to be predictive. In this case let us take the Cherry and Park study as the example. Their findings would have external validity if young adults really do in general have better spatial memory than older adults, if distinctive contextual cues really do in general facilitate spatial memory, and if both young and old really do in general benefit equally from such cues. In each case the "in general" refers to what is found across a

variety of samples of young and old, a variety of measures of spatial memory, and a variety of contextual cues. If any one of the findings fails to generalize across these dimensions, then that finding lacks external validity. Perhaps, for example, contextual cues make a difference only for "small-scale" environments such as Cherry and Park's model and map, and there is no comparable effect in full-size, real-life settings. If this limitation actually held (other research makes clear that it does not), then the Cherry and Park study would have limited external validity.

Exactly what forms of generalizability are important varies to some extent across studies. Table 2.3 lists and briefly describes the most common dimensions that are relevant to external validity.

A satisfactory study must have both internal validity and external validity. As Campbell and Stanley (1966) observe, "internal validity is the basic minimum without which any experiment is uninterpretable" (p. 5). Logically the internal validity question is the primary one, because findings can hardly be generalized if there are no valid findings in the first place. External validity is also critical, however. Internally valid conclusions do not mean much if they cannot be generalized beyond the study in which they occur.

Internal validity is also a prerequisite for the third form of validity: **construct validity**. Construct validity has to do with theoretical accuracy: Have we arrived at the correct explanation for any cause-and-effect relations that the study has demonstrated? We assume, in other words, that we have internally valid conclusions; the question now is whether we know *why* the results have occurred.

Suppose, for example, that we are confident that the context manipulation in the Cherry and Park study really did cause variations in memory performance. Why did the context make a difference? Probably the most obvious explanation—and the one that has guided most such research—is that it is the distinctiveness of the visual information that is important: Locations are easiest to remember when they are embedded within a well-differentiated spatial surround. But perhaps there is a different basis. Maybe the model was more interesting and engaging than the map, resulting in

Table 2.3 Dimensions of External Validity

Dimension	*Issue*
Sample	Do the results generalize beyond the sample tested to some broader population of interest?
Setting	Do the results generalize beyond the setting used in the research (e.g., a structured laboratory environment) to the real-life settings of interest (e.g., behavior at home or at school)?
Researcher	Are the results specific to the research team that collects the data, or would the same results be obtained by any team of investigators?
Materials	Are the results specific to the particular materials used to represent the constructs of interest, or would the same results be obtained with any appropriate set of materials?
Time	Are the results specific to the particular time period during which the data were collected, in either a short-term sense (e.g., a measure administered in late afternoon) or a long-term sense (e.g., a measure affected by historical events)?

closer attention and hence better memory. By this view, any manipulation that heightens attention should improve performance, quite apart from the spatial distinctiveness of the cues. Or maybe the participants were more confident when confronted with the relatively familiar model than when confronted with the abstract map, and it was this heightened confidence that led to better memory. By this view, any manipulation that increases confidence should improve performance. If plausible competing explanations for the results cannot be ruled out, then the study lacks construct validity.

The preceding discussion has been just a first pass at constructs that will recur in various contexts throughout the book. For now, let us settle for one more point with respect to validity. It concerns the difficulty of simultaneously achieving the various forms of validity in the same study. This difficulty exists because often research decisions that maximize one form of validity work against another form. The trade-off is most obvious with regard to internal and external validity. In general, the more tightly controlled an experiment is, the greater its internal validity—that is, the more certain the experimenter can be that the variables really do relate in the manner hypothesized. At the same time, the artificiality of a tightly controlled experiment may make generalization to the nonlaboratory world hazardous. Conversely, research conducted in natural settings with naturally occurring behaviors may pose little problem of generalizability, because the situations to which the researcher wishes to generalize are precisely those under study. The lack of experimental control, however, may make the establishment of valid relationships very difficult.

Sampling

Decisions about variables have to do with the *what* of research: What independent variables am I going to manipulate, and what potential outcomes of these variables am I going to measure? Also important are decisions about *who:* With what sorts of participants am I going to explore these independent variable–dependent variable links?

The selection of participants for research is referred to as **sampling**. Sampling is important because of the constraints on the scope of research. With very rare exceptions, psychologists are not able to study all of the people in whom they are interested. The researcher of infancy, for example, is not going to test all of the world's babies, or even all those in the United States, or (probably) even all those in one specific geographical community. Instead, what researchers do is to test **samples**, from which they hope to generalize to the larger **population** of interest. The generalization is legitimate if the sample is *representative* of the larger population. This, clearly, is an issue of external validity.

How can researchers ensure that a sample is representative of the population to which they wish to generalize? A logical first step is to define what the population of interest is. It need not be as broad as all of the world's infants; more likely, perhaps is something like "all full-term, healthy 3-month-olds growing up in the United States." Once the desired population has been defined, the next step is **random sampling** from that population. As the term implies, random sampling means that every member of the population has an equal chance of being selected for the research. If all members of the population really are equally likely to be selected, then the most probable outcome of the sampling process is that the characteristics of the sample will mirror those of the population. Note, however, that the likelihood that this desired outcome will in fact be achieved varies directly with the size of the sample. A random sample of 100 is a good deal more likely to be representative than a random sample of 10. This principle is just one of a number of arguments (we will encounter some others in chapter 8) for using large rather than small sample sizes.

In some instances researchers may use modified forms of random sampling, especially when the intended sample size is limited and pure random selection might therefore not produce the desired outcome. In **stratified sampling** researchers first identify the subgroups within the population that they want to be sure are represented in their correct proportions in the final sample. A researcher might want to be sure that males and females are represented equally, for example, or that different ethnic groups appear in proportions that match their numbers in the general population, or that freshmen are just as common as seniors in a college student sample. Samples are then drawn in the desired proportions from the identified subgroups—thus, equal numbers of males and females, 25% of the participants from each year in college, and so forth.

The goal of stratified sampling is to ensure that different members of the population are represented in their actual proportions in the sample selected. In contrast, with **oversampling** the researcher deliberately samples one or more subgroups at rates *greater* than their proportion in the target population, the goal being to achieve a sufficiently large sample of the subgroup to permit conclusions. Suppose, for example, that we plan to conduct a survey of high school students in which comparisons among ethnic groups are one of the issues of interest, and suppose also that Asian Americans constitute 3% of the high school population in the city in which we are working. Even with a total sample of 1,000 students, a random sampling approach will give us only about 30 Asian American participants, which may not be enough to draw conclusions. If we deliberately oversample Asian Americans, however (say at a 6% rather than a 3% rate, thus giving 60 students total), we can end up with a sufficient subsample for analysis, while still achieving adequate numbers in the other groups of interest.

How often do psychologists in fact draw their samples in the textbook-perfect fashion just described? The answer is: not very often. Random sampling and its variants are occasionally found in psychological research—perhaps most commonly in large survey projects in which it is important that the sample match some target population. More generally, most researchers undoubtedly start with at least an implicit notion of the population to which they wish to generalize, and most would certainly avoid selecting a sample that is clearly nonrepresentative of this population. Nevertheless, true random sampling from some target population is rare. The most obvious and frequent deviation from randomness is geographical. Researchers tend to draw samples from the communities in which they themselves live and work. Often, moreover, they may sample from only one or a few of the available hospitals, day care centers, or schools within the community. Such selection of samples primarily on the basis of availability or cooperation is referred to as **convenience sampling**. Samples obtained in this way may not be representative of the broader population with respect to variables such as social class and race, and they *cannot* be completely representative with respect to variables like region of the country or size of the community.

How important are these deviations from random sampling? There is no simple answer to this question; among the dimensions that are relevant are the topic under study; what the researcher wishes to conclude about the topic; and, of course, just how nonrandom and potentially nonrepresentative the sample is. We will revisit issues of sampling throughout the book in the context of particular kinds of research. For now, I settle for two pieces of advice, one directed to the reader of research reports and the other to the author of such reports.

The advice for the reader is to make a careful reading of the Participants section an important part of the critical evaluation of any research project. However satisfactory the other elements of a study may be, the results do not mean much if the sample is not representative of some larger population of interest. One question concerns the standing of the sample on the

demographic characteristics that may affect response. At the least, these characteristics will include age, sex, and race; for particular studies additional dimensions (e.g., income level, geographical region, health status) may also be important. Another question concerns the method of recruitment. What was the initial pool from which participants were drawn, how many of these potential participants actually made it into the study, and (if there was any dropout) how many stayed in the study until the end? Finding a representative pool of potential participants is a good starting point for research, but it is not sufficient; the real question is how well the final sample reflects the starting point.

The advice for the author follows from the points just made. Readers cannot critically evaluate the samples for research if Participants sections do not tell them enough about the samples. It is the author's responsibility to make sure that all of the necessary information is conveyed to the reader. Helpful further sources with respect to what sorts of information to convey include the *APA Publication Manual* (APA, 2001), Hartmann (2005), and Rosnow and Rosnow (2006).

Control

The notion of control was touched on in each of the preceding sections. Recall that the independent variable is defined as a variable that is under the control of the researcher. Control is central to the establishment of validity, especially internal validity. And selection of the right participants is one sort of control that a researcher must exercise. The purpose of the present section is to discuss the further sorts of control that become important once participants are in hand.

As Table 2.4 indicates, three forms of control are important in the execution of studies. The table summarizes the forms and gives examples of how each type applies or might apply to the illustrative studies. Both the forms of control and the examples are elaborated and should become clearer as we go. The table is intended simply as a guide to help keep track of the distinctions to be made.

One type of control concerns the exact form of the independent variable. If the interest, for example, is in the effects of a certain kind of reinforcement, then the researcher must be able to deliver exactly this kind of reinforcement to the participants. If any unintended deviations occur—in form, timing, consistency, or whatever—the researcher can no longer be certain what the independent variable is. Or consider again the Dufresne and Kobasigawa examination of study time. Because the researchers' interest was in possible effects of item difficulty, it was critical that they present the same easy-hard contrast to all of the children.

The point being made about this first form of control is hardly an esoteric one. It is simply that if one wants to study the possible effects of something, one must first be able to produce that something. Note, however, that doing so is not always as easy as in the two example studies, in which the levels of the independent variables were defined simply by the different stimulus materials that were presented. When the experimental manipulation is more complicated, delivering the variable in the same form to all participants can become a challenge. The challenges, moreover, are often multiplied when children are the participants, a point to which I return later.

A second form of control has to do with factors in the experimental setting other than the independent variable. Independent variables do not occur in a vacuum; there must always be a context for them, and it is the job of the researcher to determine exactly what this context will be. In giving a memory test, for example (as in the two example studies), the researcher must decide not only what test to use but also what the immediate environment for the testing will be like. One easy decision in this particular case is to make the environment as quiet as

Table 2.4 Forms of Control in Experimental Research

Type of control	*Methods of achieving*	*Examples from illustrative studies*
Over the independent variable	• Make the critical elements of the experimental manipulation the same for all participants	• In Dufresne and Kobasigawa, present the same sets of easy and hard items in the same way to all the children
Over other potentially important factors in the experimental setting	• Hold the factors constant for all participants • Disperse variations in the other factors randomly across participants	• In Cherry and Park, use the same quiet testing room for all participants • In Dufresne and Kobasigawa, vary the time of testing randomly across children
Over preexisting differences among the participants	• Randomly assign participants to experimental conditions • Match participants on potentially important attributes prior to experimental conditions • Test each participant under every experimental condition	• In Cherry and Park, randomly assign half of the participants at each age to the model condition and half to the map condition • In Cherry and Park, measure the participants' IQs and assign equal-IQ participants to the different conditions (not actually done) • In Dufresne and Kobasigawa, test every child with both the hard and the easy items

possible, in order to minimize distractions. Once the experimenter has made this decision, it is then his or her job to ensure that each participant receives the same quiet environment.

Let us introduce some further terminology at this point. Differences in scores on the dependent variable are referred to as the *variance* of the study. Those differences that can be attributed to the independent variables are called **primary variance**; those that result from other factors are called *secondary variance* or *error variance*. By controlling the level of other potential variables, experimenters attempt to maximize the proportion of primary variance in the study. Perhaps even more important, they attempt to make sure that other sources of variance are not systematically associated with any of the independent variables. Suppose, for example, that Cherry and Park had tested all of their young adult participants in a quiet laboratory on campus but all of their older participants in a noisy room at a senior citizens' center. Clearly, in this case there would have been two independent variables—age and testing environment—when only one had been intended. Any such unintended conjunction of two potentially important variables is referred to as **confounding**. A major goal of good research design is to rule out confounding.

As Table 2.4 indicates, control of unwanted variables can take a couple of forms. Often it is possible to control the variable by making it the same for all participants. This is the case in the memory example, in which the noise level of the testing environment is held constant for all participants. Sometimes, however, such literal

equating is not practical. We can return to the Dufresne and Kobasigawa study for an example. In research with school-aged children, a plausible contributor to how the children respond is the time of day at which the testing occurs. Cooperation and attentiveness are not necessarily the same late in the school day as they are first thing in the morning, or immediately before recess as compared with immediately after, or on a Friday as compared with a Monday. Clearly, Dufresne and Kobasigawa would have introduced a potentially important confounding if they had tested all of their first-graders early in the day and all of their seventh-graders in the afternoon. One way to avoid this problem would be to test all of the children at the same point in the day, say at one o'clock on Wednesday. With this approach, however, most studies would take months to complete, and even then only time of day and not time of year (which also can be important) would be held constant across participants. A sensible alternative would be to let the time of testing vary across children but to make sure that the variations are the same for the different groups being compared—in this case, first-, third-, fifth-, and seventh-graders. In this case the control of the time-of-testing variable would lie not in its equation but in its randomization—that is, by dispersing differences in it equally across the groups of interest.

Shorn of certain specifics, the discussion thus far should have a familiar sound to it. What has been presented here is simply the classic scientific method: to determine the effects of some factor, systematically vary that factor (the first form of control) while holding other potentially important factors constant (the second form of control).

There is still a third form of control that is essential. Thus far, the "other potentially important factors" that have been discussed have been factors within the experimental setting—for example, the noise level of the testing room. Another important source of variance in any experiment stems from individual differences among the participants. Participants are not identical at the start of an experiment, and differences among them contribute error variance to the final results. Because there is no way to rule out such differences, the method of control must again be through dispersion rather than equation. What the experimenter must ensure is that the differences are spread equally across the different treatment groups—or, to make the same point in different words, that the groups are equivalent prior to the application of the treatment. Doing so requires that the experimenter have control not only over the form of the treatment but also over who gets what treatment.

How can the experimenter assign people to groups in a way that will ensure that the groups all are initially equivalent? The answer is that although there is no way literally to ensure equivalence, there are ways to come as close as can reasonably be expected. The most common method is through **random assignment** of participants to the different groups. Random assignment means that each participant has an equal chance of being assigned to each group. If each participant has an equal chance of being assigned to each group, then the characteristics associated with each participant (IQ, sex, relevant past experience—whatever might affect the results) have an equal chance of falling in each group. It follows that the most probable outcome of the assignment process is that these characteristics will end up equally distributed in the different groups, a result that is, of course, the researcher's goal. The logic of random assignment is clearly the same as the logic of random sampling, and the success of the process shows a similar dependence on sample size. One could not randomly divide 8 participants into two groups and conclude with any confidence that the randomization had produced equivalent groups. With a sample of 80 participants, the odds are much better.

Random assignment is a much more frequent component of research than is true

random sampling. Indeed, random assignment has been referred to as "the key defining attribute of the experimental method" (McCall & Green, 2004, p. 4).

Powerful though random assignment is, it does have a limitation. At best, random assignment makes it *probable* that the groups being compared are equivalent; it cannot guarantee this outcome. An obvious question follows: Why settle for probability? Why not identify the dimensions on which we wish the groups to be equal (e.g., intelligence, SES, health status—the list will vary across studies) and then assign participants based on these dimensions—thus, the same proportion of high-intelligence participants in each group, the same proportion of middle-class participants in each group, and so forth? Why, in short, not do the assignment in a way that *ensures* equivalence?

The general answer to this question is that such matching is more difficult than might at first appear and that the attempt to achieve it can sometimes create more problems than it solves. A more specific answer is given in chapter 3, when we return to the issue of selecting and assigning participants. Also discussed in chapter 3 is the third general technique for achieving equivalence: testing every participant under each experimental condition.

Subject Variables

Manipulable Versus Nonmanipulable Variables

Thus far the discussion of experimental control has focused on the ideal situation for research: the case in which the researcher can systematically manipulate the independent variables of interest while holding all other variables constant, and can assign participants to the different treatment groups either randomly or randomly within certain desired constraints. With many variables such control is not only desirable but quite feasible. We saw examples of this kind of control in both of the cited studies: the easy-hard contrast in the Dufresne and Kobasigawa study, and the model-map contrast in the Cherry and Park study.

The developmental psychologist's life is complicated, however, by the fact that not all variables of interest lend themselves to the kind of manipulation that good research design demands. Again, both of the cited studies provide examples, and in this case it is the same example: chronological age. Clearly, age is not something that the researcher randomly assigns to people; rather it is a characteristic that people bring to the experimental setting. Age is just one example of what are called **subject** (or *classification* or *attribute)* **variables**: intrinsic properties of individuals that cannot be experimentally manipulated but must be taken as they naturally are. Other common examples are race and sex. The researcher who wishes to work with such characteristics as independent variables forgoes the possibility of control through manipulation. The only control possible in such cases is control through selection of people who already possess the characteristic.

A number of other variables of interest, although not literally nonmanipulable, are never in fact the subject of controlled experiments with humans. From a theoretical perspective, for example, it would be very interesting to know whether infants deprived of mothers develop in the same way as infants who have mothers. Despite the early work of Frederick II (noted in chapter 1), we do not have manipulative studies of this issue. Yet there has long been a literature on "maternal deprivation" and its effects on the child. What researchers have done is to identify situations in which infants have already been left motherless (usually in orphanages) and then take advantage of these "natural experiments" by studying how the infants develop. And there are numerous similar examples of psychologists' ability to capitalize upon naturally occurring events—studies of malnutrition in infancy, of father absence during childhood, of social isolation in old age, and so forth. In each case the

independent variable is created through selection rather than experimental manipulation.

Research with nonmanipulated variables does not attain the status of the "true experiment," because the controlled manipulation that constitutes the heart of an experiment is not possible. For this reason such research is labeled as *preexperimental* in Campbell and Stanley's (1966) influential discussion of experimental design. Because of the lack of control, such studies can never establish cause-and-effect conclusions with the certainty that is possible in a manipulative experiment.

What exactly are the limitations of such research? The problems are of two main sorts. First, it is impossible to assign participants randomly to groups. Because random assignment is impossible, there is no way to be sure that the groups under study are equivalent except for the variable of interest (e.g., presence or absence of mother), and therefore no way to be sure that any differences between groups are caused by that variable. This, in fact, was one criticism of the early maternal deprivation studies. Perhaps babies who grow up in orphanages are a nonrandom subset of the general population of babies, a subset that includes an unusually high proportion of genetic or organic problems. If so, then differences between orphanage babies and other babies could not be attributed with any confidence to the effects of the orphanage rearing. In a well-designed experiment, such confounding would be ruled out by random assignment. This, it should be clear, is a problem with internal validity: We cannot be certain that our independent variable is really the causal factor.

The other problem concerns the broad-scale and longstanding nature of most subject variables. Orphanage rearing, father absence, social isolation, growing up Black (or White), and growing up male (or female) all encompass a host of factors that can affect an individual's development. Thus, even if we find a significant effect associated with a particular subject variable, we still do not know what the specific causal factors are. This, too, has been a problem in research on maternal deprivation. Although the damaging effects of certain kinds of orphanage rearing are not in dispute, there has long been debate about whether the effects result from lack of normal mothering or from more general cognitive-perceptual deprivation. Even if we could conclude that mothering per se is important, we still would not know which of the many things that mothers normally do with infants are critical to the effect. Again, there is a confounding of factors that a well-designed experiment would keep separate. A researcher with control over variables is unlikely to set up an independent variable that is so global that its effects cannot be interpreted. This, it should be clear, is a problem with construct validity: We do not know whether we have arrived at the correct theoretical interpretation of the results.

This discussion is not meant to suggest that there is no value in demonstrating that a variable like maternal deprivation or sex or age is associated with important outcomes in the child. But it should be realized that such a demonstration is merely the first step in a research program.

Age as a Variable

Because of its importance in developmental research, the variable of chronological age deserves a somewhat fuller consideration. Much research in developmental psychology has as one of its points a demonstration that participants of different ages either are or are not similar on the dependent variables being studied. Even studies with a single age group may have age comparisons at their core, for often the comparison is implicit rather than explicit. A researcher of neonates, for example, may not include a comparison group of older children in the study, but findings about how neonates function can nevertheless be interpreted in light of a large body of information about the functioning of older children. To take a very simple example, one would hardly do

research to determine whether young infants have color vision (e.g., Adams, 1995) unless one already knew that color vision is eventually part of the human competence.

Developmental psychologists are sometimes apologetic about the "merely age differences" nature of much research in developmental psychology. But the identification of genuine changes with age is clearly a valid part of a science of development. Not only is description a legitimate part of any science, but accurate description provides the phenomena to which explanatory models must speak. It is only when we know, for example, that young children do not understand conservation (Piaget & Szeminska, 1952) that we can begin to build a model of why this is so and of where eventual understanding comes from.

Although we may agree that the study of age changes is legitimate, it is important to be clear about exactly what is meant by a "genuine change with age." What is *not* meant, certainly, is that chronological age in any direct sense causes the change. What *is* meant is that variables that are regularly and naturally associated with age produce the change. It is then the job of the researcher to determine which of the potentially important variables are in fact important.

The earlier discussion stressed that a primary goal of experimental control is the creation of groups that are equivalent in every way except for the independent variable being examined. This goal takes on special meaning in the case of a broad subject variable like age. Imagine that you are interested in comparing 7-year-olds and 12-year-olds. If you wish to make the groups equivalent in every way except age, then you will have to find 7- and 12-year-olds whose levels of biological maturation are the same, who have been going to school for the same number of years, whose general experiences in the world are equivalent, and so forth. Clearly, such a goal is not only impossible but quite misguided. Variables like biological maturation, years of schooling, and general experience are among the variables that are "regularly and naturally associated with age." As such, they are factors to be studied, not ruled out through experimental control.

On the other hand, there are other potentially important factors that must not be allowed to confound the age comparison. An obvious kind of confounding would occur if all of the 7-year-olds were boys and all of the 12-year-olds girls. Maleness is not an intrinsic part of being 7, nor is femaleness an intrinsic part of being 12; hence, this factor must not be allowed to covary with age. A somewhat less obvious confounding might occur if all of the 7-year-olds were drawn from one school and all of the 12-year-olds from another school. The mere fact of attending different schools is probably not important, and in any case this difference may be unavoidable for the particular age range studied. Nevertheless, it will be important for the researcher to select schools that are as comparable as possible on dimensions such as educational philosophy, geographical location, and socioeconomic status of the population served. If this criterion is not met, then an apparent age change may not in fact be genuine.

As these examples suggest, decisions about what to match and what not to match when comparing different ages are generally straightforward. As we will see, however, such decisions are not always straightforward, nor is it always easy to achieve whatever matching one has decided on. We will return to the issue of age comparisons in chapter 3.

Outcomes

Researchers manipulate independent variables in order to examine effects on dependent variables. But what are the possible effects? In a factorial study—that is, a study with two or more independent variables—the possible effects are of two sorts: main effects and interactions.

Main Effects

A main effect is a direct effect of an independent variable on a dependent variable. It is what researchers examine when they compare the levels of a single independent variable independent of (or summed across) the other independent variables in the study. Both of the illustrative studies provide examples of main effects. In the Cherry and Park study there was a main effect of age: Young participants performed better than older ones. The means for this effect are shown in the rightmost column of Table 2.2; they are the values for all the young participants and all the old participants in the study, summed across the levels of the other independent variable (the model-map contrast). Similarly, there was a main effect of experimental condition, and the values for this effect are shown at the bottom of the table: the means for all participants in the model and all those in the map condition, summed across the two levels of age.

The Dufresne and Kobasigawa study also produced main effects of age and experimental condition. The means for these effects appear in the "combined" portions of Table 2.1. In both studies, therefore, we can say that both independent variables had an effect: Scores on the dependent variable varied as a function of age and experimental condition. Note, however, that the effect of age in the Dufresne and Kobasigawa study is more complicated than the other main effects, because in this case there are four levels of the independent variable rather than simply two. Main effects for variables with more than two levels pose some special statistical and interpretive complexities—an issue to which I return in chapter 8.

Interactions

A main effect is an effect of a single independent variable considered in isolation. An **interaction**, in contrast, becomes possible when we consider two or more independent variables simultaneously. An interaction occurs whenever the effect of one independent variable varies with the level of another independent variable.

The Dufresne and Kobasigawa study produced an interaction in addition to its two main effects. In this study, the effects of the easy-hard comparison varied with the level of age—little effect at the two younger grade levels, a marked effect at the two older grade levels. As with any interaction, the results can also be stated with the opposite emphasis: The effects of grade varied with the level of item difficulty—no differences with the easy items, strong differences with the hard items. This two-way ("two-way" because two independent variables are involved) interaction is graphed in Figure 2.1. The data are the same as those presented in Table 2.1; the graphical presentation, however, makes the nature of the interaction more visible. Note, in particular, the nonparallel nature of the lines. Graphically, an interaction is always signaled by some deviation from parallelism—some spreading apart or crossing over of lines that reflects the differential effects of one variable across the levels of the other. Note also, however, that graphs are not sufficient to determine that an interaction has occurred; rather, there must be a statistical test of the data to identify both main effects and interactions. I discuss the most common such test, the analysis of variance or ANOVA, in chapter 8.

What would the graph look like if there were no interaction? Figure 2.2, which plots the means from the Cherry and Park study, provides an answer. Recall that their study found equivalent benefits from the model condition for both the younger and the older participants—thus, two main effects (age and condition) but no interaction. This situation is reflected in the essentially parallel lines of Figure 2.2. (That the lines are not perfectly parallel reflects the fact that there was a slight trend toward an interaction—a slight

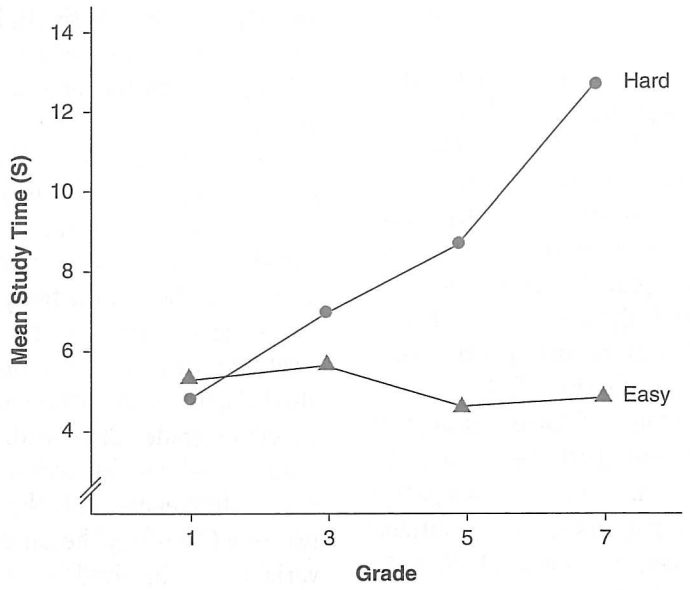

Figure 2.1 Interaction of age and experimental condition in the Dufresne and Kobasigawa study

SOURCE: Adapted from "Children's Spontaneous Allocation of Study Time: Differential and Sufficient Aspects," by A. Dufresne and A. Kobasigawa, 1989, *Journal of Experimental Child Psychology, 47*, 274–296. Copyright © 1989, Academic Press.

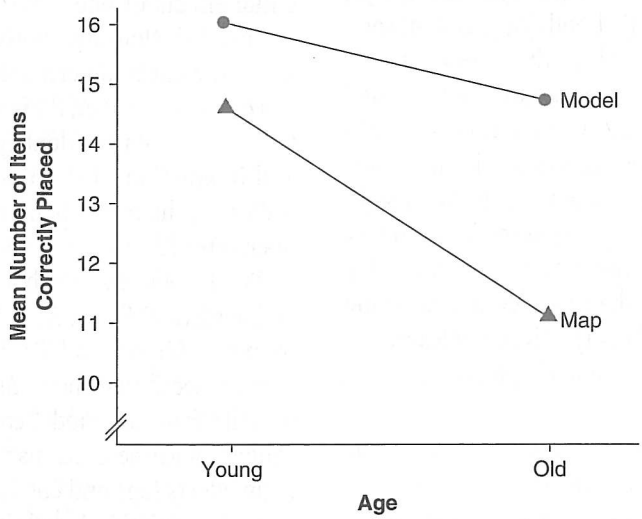

Figure 2.2 Main effects in the Cherry and Park study

SOURCE: Adapted from "Individual Differences and Contextual Variables Influence Spatial Memory in Younger and Older Adults," by K. E. Cherry and D. C. Park, 1993, *Psychology and Aging, 8*, 517–526. Copyright © 1993, American Psychological Association.

tendency for older participants to benefit more from the model.)

The interaction in the Dufresne and Kobasigawa study was between a subject variable and an experimentally manipulated variable. Interactions are not limited to such designs, however; rather, they can occur between independent variables of any sort. Interactions are possible, therefore, in any multiple-factor experiment. Figure 2.3 shows an interaction between two experimentally manipulated variables, and Figure 2.4 shows an interaction between two subject variables. The main finding of the Patterson and Carter (1979) study, pictured in Figure 2.3, was that the presence of a desired reward lessened children's self-control when they were simply waiting for the reward but enhanced self-control when they were working to complete a task rather than simply waiting. One finding of the Underwood, Coie, and Herbsman (1992) study, pictured in Figure 2.4, was that children's tendency to use "display rules" to mask feelings of sadness varied as a function of both age and sex. At the two younger grade levels, girls were slightly more likely than boys to report that they would attempt to disguise the fact that they were sad; by seventh grade, however, a clear difference had emerged in favor of boys.

As a comparison of Figures 2.1, 2.3, and 2.4 suggests, interactions can take a variety of forms. They can also become exceedingly complicated when more than two independent variables are involved. Although some researchers try, it is seldom possible to make sense of a four- or five-way interaction.

Interpreting any sort of interaction can be a complex matter, both statistically and theoretically (Levin, 1985; Rosnow & Rosenthal, 1995). I settle here for one basic point about interactions. The most general implication of a significant interaction between two variables is that interpretations of main effects involving those variables must be made with caution. In the Dufresne and Kobasigawa study, for example, there were main effects of both age and item difficulty; as Figure 2.1 reveals, however, the age effect was limited to the hard items, and the item effect was limited to the older children. In the Patterson and Carter study, in contrast, the main effect of the reward

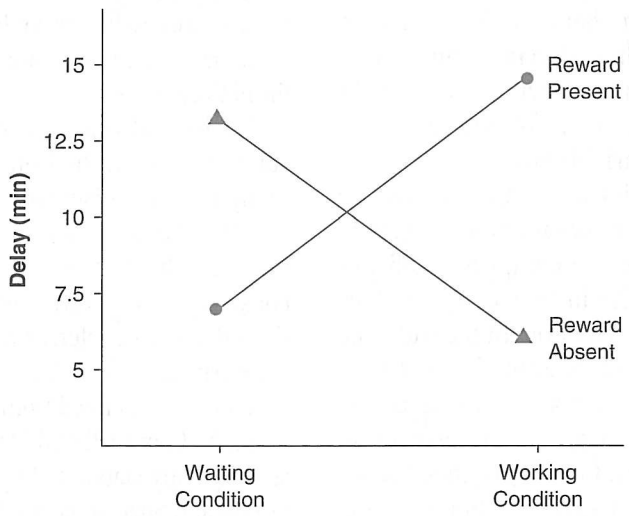

Figure 2.3 Interaction of experimental conditions in the Patterson and Carter study

SOURCE: Adapted from "Attentional Determinants of Children's Self-Control in Waiting and Working Situations," by C. J. Patterson and D. B. Carter, 1979, *Child Development, 50*, 272–275. Copyright © 1979, Blackwell Publishing, Inc.

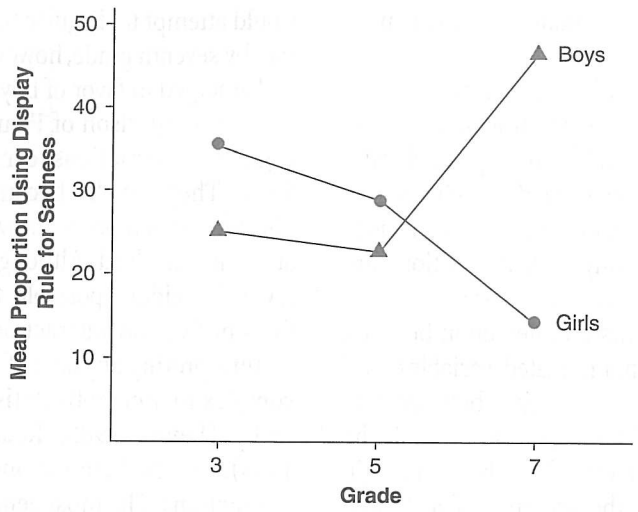

Figure 2.4 Interaction of age and sex in the Underwood, Coie, and Herbsman study

SOURCE: Adapted from "Display Rules for Anger and Aggression in School-Age Children," by M. K. Underwood, J. D. Coie, and C. R Herbsman, 1992, *Child Development, 63*, 366–380. Copyright © 1992, Blackwell Publishing, Inc.

present–reward absent manipulation was *not* significant, a finding that would suggest that this variable had no effect. Such a conclusion, however, is clearly contradicted by a separate analysis of the working and waiting conditions. An interaction, then, is a signal that the world is more complicated than we might have expected. Studying an independent variable in isolation cannot give us a full picture of the way in which that variable operates.

Note that the point just made can also be put in the context of external validity. An interaction implies a limitation in the generality of conclusions about the independent variables that enter into the interaction. In the Dufresne and Kobasigawa study, for example, the effects of item difficulty did not generalize across age, and the effects of age did not generalize across item difficulty. Conversely, the absence of an interaction is evidence in support of the external validity of conclusions regarding the variables in question—at least across the particular dimensions and levels that are sampled.

Threats to Validity

As we have seen, the ultimate goal in designing research is always to arrive at valid conclusions about the phenomena being studied. The converse to successful research design comes when there are threats to validity—uncertainties or limitations in what can be concluded that the design has failed to rule out. Several threats to validity were touched on in this chapter, and many more are discussed in the coming chapters. It will be helpful for the coming discussion to have a brief overview of the factors to be considered—an overall list and a set of definitions that can be referred to as necessary. This is the purpose of Table 2.5.

Table 2.5 is derived from an influential monograph by Campbell and Stanley (1966) that was subsequently elaborated by Cook and Campbell (1979) and Shadish et al. (2002). It does not provide an exhaustive list of things that can go wrong in research (Shadish et al. discuss 37 threats to validity!); it does, however, include many of the problems that are discussed later in

Table 2.5 Threats to Validity

Source	*Description*
Selection bias	Assignment of initially nonequivalent participants to the groups being compared
Selective drop-out	Nonrandom, systematically biased loss of participants in the course of the study
History	Potentially important events occurring between early and later measurements in addition to the independent variables being studied
Maturation	Naturally occurring changes in the participants as a function of the passage of time during the study
Testing	Effects of taking a test upon performance on a later test
Reactivity	Unintended effects of the experimental arrangements upon participants' responses
Instrumentation	Unintended changes in experimenters, observers, or measuring instruments in the course of the study
Statistical regression	Tendency of initially extreme scores to move toward the group mean upon retesting
Low reliability	Errors of measurement in the assessment of the dependent variable
Low statistical power	Low probability of detecting genuine effects because of characteristics of the design and statistical tests
Mono-operation bias	Use of a single operationalization of either the independent or dependent variable
Mono-method bias	Use of a single experimental method for examining possible relations between the independent and dependent variables

the text. Again, there is no expectation that the table is completely self-explanatory; its purpose, rather, is as a preliminary guide to concepts that will receive further attention as we go along.

Summary

This chapter begins with some basic terms and concepts. All research involves variables. *Dependent variables* are the outcome variables in research—for example, the number of aggressive acts in a study of aggression. *Independent variables* are potential causal factors that are controlled by the researcher—for example, reinforcement for aggression. The goal of most research is to determine whether variations in the independent variable relate to variations in the dependent variable—for example, does aggression increase following reinforcement?

The basic issue with respect to all research is validity. *Validity* refers to the accuracy with which conclusions can be drawn from research. Three forms are discussed in this chapter: *internal validity,* which concerns the accuracy of cause-and-effect conclusions

within the context of the study; *external validity,* which concerns the generalizability of the conclusions beyond the study; and *construct validity,* which concerns the accuracy of the theoretical interpretation of the conclusions.

An important decision that the researcher must make concerns the participants for research. The goal in sampling participants is to obtain a *sample* that is representative of the larger *population* to which the researcher wishes to generalize. The common prescription for achieving representativeness is to do *random sampling* from the target population. In fact, most research in developmental psychology employs sampling procedures that are less than totally random. In some instances the deviations are intentional and systematic, the goal being to ensure that the sample possesses certain characteristics; *stratified sampling* and *oversampling* are examples. More commonly, the deviations reflect the use of samples that are readily available, an approach known as *convenience sampling.* How important such departures from randomness are varies across different topics. Nevertheless, representativeness and external validity remain important questions to examine for any study.

The discussion turns next to the construct of control. Three kinds of control are important if clear cause-and-effect conclusions are to be drawn. A first is over the exact form of the independent variable. A second is over other potentially important factors in the situation. Two methods of achieving this second form of control are discussed: holding the other factors constant and randomly dispersing variations in them across participants. The third kind of control is over preexisting differences among participants. One method of achieving this form of control, random assignment, is discussed in the present chapter; two others (matching and within-subject testing) are deferred for later consideration.

In some kinds of research, the degree of control is limited by the nature of the variables. The term *subject variable* refers to preexisting differences among people that are not experimentally manipulable; examples include age, sex, and race. The only control possible with such variables is through selection, a point that applies also to situations (e.g., maternal deprivation) whose experimental induction would be unethical. Although such variables are often of great interest to the developmental psychologist, cause-and-effect conclusions are difficult to establish in the absence of experimental manipulation. Specifying the exact basis for an effect can be a problem with a broad and multifaceted variable; ruling out other possible causal factors can also be difficult.

Subject variables are often of special interest when they enter into interactions. An *interaction* occurs whenever the effects of one independent variable depend on the level of another variable. In contrast, a *main effect* refers to an effect of an independent variable that is independent of the other factors in the study. Interactions can occur with independent variables of any sort, and they can take a variety of forms. Their most general message is that relations are complicated and that conclusions about any one variable must be made with caution.

The chapter concludes with a brief return to the concept of validity and an overview of some of the major threats to validity that are considered throughout the book.

Exercises

1. Find at least three recent summaries of developmental psychology research in the popular press (newspapers, magazines). For each, generate a list of possible threats to the validity of the research. If the description of the research is not complete enough for you to evaluate some forms of validity, specify what further information you would need.

2. Consider the task of recruiting research participants of the following ages: 6 months, 4 years, 12 years, 70 years. For each age group, generate a list of ways in which you might recruit prospective participants. For each method of sampling, discuss the likely representativeness of your final sample.

3. A particular construct can serve as either an independent or a dependent variable, depending on the way it is used in research. Consider the following constructs: anxiety, activity level, academic readiness. For each, generate a study in which the construct serves as (a) a dependent variable, (b) an experimentally manipulated independent variable, (c) a subject variable, and (d) a correlational variable.

3

Design

We saw in chapter 2 that all research involves comparison. In most cases, the comparison is between different levels of an independent variable. If the independent variable is a nonmanipulable subject characteristic such as age, then the researcher must select participants who already possess different levels of the characteristic. If the independent variable is an experimentally manipulable factor, then the researcher must assign participants to conditions that embody the desired levels of the factor. In either case, the researcher must do the selecting and assigning in a way that will allow a clear, nonconfounded comparison of the different levels being studied (the internal validity question), that will permit generalization to other samples and situations of interest (the external validity question), and that will allow identification of the causal bases for any relations that are found (the construct validity question).

The steps and the goals just sketched are issues of *experimental design.* Design, in the words of Kerlinger and Lee (2000, p. 449), is "the plan and structure of investigation"—the way in which studies are put together. Although the overall goal—valid conclusions—is always the same, studies can in fact be put together in a variety of ways. This chapter considers some of the most important dimensions along which research designs vary.

Once again, the sample studies described in chapter 2 can serve to illustrate some general points and standard terminology. Both the Dufresne and Kobasigawa and the Cherry and Park studies included two levels of an experimentally manipulated variable: easy versus hard items in Dufresne and Kobasigawa, and the model versus map context in Cherry and Park. Cherry and Park assigned separate participants to their two experimental conditions; hence their approach can be labeled a **between-subject design**. Dufresne and Kobasigawa tested all of their participants in both the easy and hard conditions; hence their approach can be labeled a **within-subject design**. One basic decision that a researcher must make is whether to use the same or different participants when comparing the effects of two or more experimental treatments. Strengths and weaknesses of both approaches are discussed later in this chapter.

Both of the sample studies also included the nonmanipulable variable of chronological age. In this case the methodological decision was

the same: Both sets of researchers tested separate participants at the different ages. The strategy of testing different groups of people at different ages is referred to as a **cross-sectional design.** It is not the only possible approach to studying differences with age. Dufresne and Kobasigawa, for example, might have tested a sample of first-graders, waited 2 years and tested the children again as third-graders, waited another 2 years and tested the children in fifth grade, and finally tested the now seventh-graders after one last 2-year wait. The strategy of repeatedly testing the same sample of participants across the ages of interest is referred to as a **longitudinal design.**

It should be clear that there is a basic similarity between the between-versus-within contrast and the cross-sectional versus longitudinal contrast. In both cases the central issue is whether to examine effects within the same people or across different people. The relative merits of cross-sectional and longitudinal approaches are also discussed shortly.

Although the Dufresne and Kobasigawa and the Cherry and Park studies differed on the between-versus-within dimension, they were similar in another, perhaps more basic, respect. The similarity is that both studies did include an experimentally manipulated independent variable: easy-hard in Dufresne and Kobasigawa, and model-map in Cherry and Park. As we saw in chapter 2, not all studies include true independent variables of this sort. In so-called correlational or nonexperimental designs, the variables are simply measured, not controlled, and the researcher then searches for relations among the measures. Correlational designs are the third major topic considered in this chapter.

Because age comparisons are central to research in developmental psychology, the chapter begins with a consideration of designs for studying age. The discussion then moves to methods for comparing experimental conditions, and the chapter concludes with a consideration of the strengths and limitations of correlational research.

Age Comparisons

As noted earlier, age is just one of a number of subject variables that can be examined in research. Because the focus here is on age, it is worth noting an important difference between age and most other subject variables—a difference with implications for choice of research design. The investigator of a variable like sex or race does not have the option of waiting for the participants to change from one level of the variable to another; rather, studies of these variables must necessarily involve separate groups of people. In the case of age, however, today's 6-year-old is tomorrow's 8- or 10- or 20-year-old. It is because of this natural change along the age dimension that the researcher of age differences has the option of adopting either a within-subjects or a between-subjects approach.

There is a further point here as well. If we do a study to compare boys and girls, then our interest clearly is in differences (or, of course, lack of differences) between boys and girls. If we do a study to compare 6- and 10-year-olds, our interest may be partly in differences between 6- and 10-year olds, but it is likely to go deeper as well. What we may really be interested in is the possibility that the 6-year-old *will become like* the 10-year-old, or, equivalently, that the 10-year-old *was once like* the 6-year-old. Our interest, in short, may be not just in age *differences* but in age *changes.* As we will see, one of the thorny problems for developmental research is to determine when differences between age groups really reflect natural changes with age as people develop.

Longitudinal Designs

A *longitudinal study* tests the same sample at least twice across some period of time.

Although there are no clear-cut rules for deciding when a study with repeated testing becomes "longitudinal," at least two rough criteria seem to govern use of the label. First, the reference is usually to the study of naturally occurring rather than experimentally induced changes. Thus, the use of delayed follow-up tests in intervention or training research is not usually classified as longitudinal, even though the same children may be tested several times. Second, the reference is typically to repeated tests that span an appreciable period of time. Thus, simply testing the same people several times at 1-week intervals is not likely to earn a study the designation "longitudinal." Note, however, that what constitutes "an appreciable period of time" will vary with the developmental level of the sample. A series of 1-week retests probably *would* be considered longitudinal if the participants were only a few days old at the time of the initial testing.

Longitudinal studies are a good deal less common than are cross-sectional studies. It is not difficult to see why. Longitudinal studies are more time-consuming, more expensive, and more difficult to bring to successful completion than are cross-sectional studies. Consider as examples the two illustrative studies from chapter 2. The Dufresne and Kobasigawa experiment probably took a few weeks to complete. Had they opted for longitudinal rather than cross-sectional testing, the minimum time period for the study would have been 6 years. The contrast is, of course, even clearer for the Cherry and Park study. If these authors had decided on a longitudinal approach, they would have had to wait 40 or 50 years for their young adults to turn into elderly adults.

In itself, the extended time frame of the longitudinal approach is simply a practical problem—bothersome certainly, but not a threat to the validity of the conclusions. There are other problems associated with the extended time, however, that do threaten validity. One is the possible obsolescence of the tests and instruments being used. Because the essence of the longitudinal design is the earlier time–later time comparison, the researcher is committed to continued use of whatever measures were selected at the beginning of the project. Often, however, a test may become outmoded or lose its theoretical interest in the course of a long study; conversely, new tests and new issues will almost certainly arise. Thus, what one wants to know in 2010 may not be what one wished to know in 1980. This problem of test obsolescence is especially great in very long-term studies, such as some of the life-span studies begun in the 1920s (Kagan, 1964). It need not be a problem in more short-term longitudinal efforts.

Other problems relate to the nature of the sample in longitudinal research. Any at all long-term longitudinal study requires a substantial commitment of time and effort on the part of its participants (and, in the case of child samples, the parents of the participants as well). Samples may be selected, therefore, at least partly on the basis of factors such as belief in the value of research or probable geographical stability. If so, they may not be representative of the population to which the researcher wishes to generalize. Furthermore, any single longitudinal sample, all born at about the same time, constitutes but a single generation or **cohort**, and any findings may be at least somewhat specific to this one generation. We may be interested, for example, in how people change across the first 30 years of life. If all of our sample were born in 1940, however, then all we know with any certainty is how people born in 1940 changed as they encountered the changing world of the 1940s, 1950s, and 1960s. Had our sample been born either earlier or later, we might have obtained somewhat different results.

Although longitudinal samples may be nonrepresentative in various ways, they do at least avoid Campbell and Stanley's (1966) problem of **selection bias**—that is, the selection of initially nonequivalent groups for comparison. There can be no problem of selection bias when each participant is being compared with

himself or herself. There can, however, be **selective dropout** (also labeled *attrition* or *mortality*), and such dropout does in fact occur. People can be lost from longitudinal samples for a variety of reasons—change of residence, unwillingness to continue to participate, or (especially in elderly samples) mortality in its literal sense. If such dropouts were random, then the only problems would be the reduction in sample size and the waste of effort in collecting early measures for which there turns out to be no later counterpart. Often, however, the dropout is not random but selective—that is, participants who are lost from the study are systematically different from those who remain. In longitudinal studies of IQ, for example, participants who drop out tend to have lower scores on the initial tests than do participants who continue (e.g., Siegler & Botwinick, 1979). Because the lower-competence dropouts contribute scores at the younger but not the older ages, the result is a "positive bias" in favor of the older groups. It is possible, of course, to limit the younger-older comparison to people who remain in the study and thus contribute scores at all ages. In this case, however, the initially nonrepresentative sample becomes even more nonrepresentative.

There is still one further way in which the participants in a longitudinal study differ from the broader population to which the researcher wishes to generalize. The difference is an obvious one: The participants in a longitudinal study undergo repeated psychological testing of a kind that most of the population escapes. Two of Campbell and Stanley's (1966) threats to validity are therefore potentially relevant. One is **testing**: the effects on later test performance of having taken the same or a similar test earlier. It seems likely, for example, that taking the same IQ test repeatedly at fairly close intervals could eventually begin to affect responses, and indeed research demonstrates that practice effects do occur (e.g., Rabbitt, Diggle, Holland, & McInnes, 2004). The second problem is the more general one of **reactivity**. Knowledge that one is the subject of research can affect anyone's responses, and such knowledge is probably especially salient for the participants in long-term, frequent-measurement longitudinal studies. Responses obtained from such participants, therefore, may not be representative of the typical course of development.

The final problem to be noted is an elaboration of the earlier point about the one-generational nature of many longitudinal samples. In longitudinal research there is an inevitable confounding between the age of the participants and the historical time of testing. This confounding follows from the fact that the age comparisons are all within subject; if we want different ages, therefore, we must test at different times. Suppose, for example, that we wish to examine possible changes between age 15 and age 20. We select a sample born in 1985 whom we test at age 15 and again at age 20. Should the second measure differ from the first, we would have two possible explanations for the difference: the fact that the participants are 5 years older, or the fact that one test was given in 2000 and the other in 2005. Age can never be disentangled from time in a longitudinal design.

How likely is it that this potential problem will actually be a problem? One determinant is undoubtedly the nature of the phenomena being studied. Let us move to the elderly years for an example of this point. Imagine that your interest is in changes in visual acuity as people age. You test a sample of 65-year-olds in 2000 and the same people at age 70 in 2005. Although historical time is a logically possible explanation for any changes you find, it is not a very plausible explanation in the case of a dependent variable like acuity. What is more likely, should you find differences, is that the visual system really undergoes natural changes between age 65 and age 70. Imagine, however, that instead of visual acuity you had tested attitudes toward airport security. You find that people are more concerned about security and more accepting of strong security measures at age 70 than at age 65. A clear case of increased caution with

increasing age? Hardly, given that the events of September 11 intervened between the two measurements (I draw this example from Schaie & Caskie, 2005). In this case the historical-cultural explanation seems the more plausible one. In both cases, however, the standard longitudinal design permits conclusions that are at best plausible, not certain. The confounding of age and time can never be removed.

This catalog of the woes that can beset the longitudinal researcher raises the question of why anyone other than a confirmed masochist would ever attempt a longitudinal study. The answer, as might be expected, is that the longitudinal approach has a number of compensating virtues (Bullock, 1995; Hartmann, 2005; Jordan, 1994). It is to the more positive side of longitudinal studies that I turn next.

I noted earlier the distinction between age changes and age differences. As long as different samples are studied at different ages, the only direct measure a study can provide is of age differences; it is a further inference that any differences found reflect changes from the earlier to the later age. In longitudinal studies, however, the measure of age changes is direct rather than inferred. As we have seen, there can be questions about why the changes occur or how generalizable they are. But at least the focus is squarely on the central question of developmental psychology: that of intraindividual development over time.

The focus on intraindividual development makes the longitudinal approach uniquely suited to questions of individual consistency or individual change. Suppose that you wish to know whether a child's IQ tends to remain the same or to go up or down as the child develops. Clearly, you cannot answer this question by testing different children at different ages; rather, you must follow the same child as he or she develops. Whenever the interest is in individual consistency or change, then the longitudinal approach is not merely a nicety; it is a must.

The value of the longitudinal approach is not limited to tracing the course of a single trait or a single behavioral system over time. The value, rather, is much broader, for potentially *any* interesting cross-age patterning can be examined if we only obtain the measures of interest. In some cases the focus may be on the relation between one aspect of the child's development early in life and some other aspect later in life. We might seek to determine, for example, whether speed of skeletal maturation in the first 2 years relates to age of onset of puberty at adolescence. In other cases the interest may be in the relation between some aspect of the environment early in life and some aspect of development later in life. Thus, we might try to determine whether the parents' child-rearing practices during the child's first 2 years relate to measures of the child's personality at middle childhood or adolescence. Whenever the interest is in the relation between something early and something later, then the longitudinal approach is again a must.

Longitudinal research is also especially suited for tracing the continuous and progressive transformations that certain very general behavioral systems undergo as the child develops. This rather murky statement needs to be clarified by examples, and two examples will in fact readily occur to anyone familiar with research in developmental psychology. One is Piaget's research on the development of intelligence in infancy (Piaget, 1952). Piaget studied each of his three children longitudinally from birth to about age 2, painstakingly charting the sequences within and relations among various domains of intelligent behavior. The result was a conception of infant intelligence that in scope and insight surpassed anything that had come before and has served as a model for much that has been done since. It is possible that at least some of the same insights might have been derived from a judicious cross-sectional study of different babies at different ages; it is doubtful, however, that the full picture of infant intelligence and how it develops could ever have emerged without the intensive, almost day-to-day study of changes within a single child over time.

A similar argument can be made for research on early language development (e.g., Brown, 1973). In much the same way as Piaget, researchers of child language have used the longitudinal approach to trace gradual changes in language across the early years of language learning. What, for example, is the earliest form that negation takes in the child's speech, and how does this rudimentary form eventually turn into the complex rule system of the older child or adult? Again, the intensive longitudinal study, in which changes can be charted within a single child, has made possible a view of early language and how it evolves that probably could not have been gleaned from cross-sectional study alone.

Clearly, longitudinal research of the sort just described involves more than simply testing the same child at least twice; such research becomes, rather, an extended case history of individual development. When is such intensive longitudinal study likely to prove most fruitful? Certainly a prime rationale for such research lies in its application to new research terrain in which many of the basic phenomena still remain to be discovered. "New terrain" was certainly an accurate description of the field of infant intelligence when Piaget began his work. Once some idea of the general form and salient landmarks of development has emerged, more focused cross-sectional studies can be profitably applied. Longitudinal study is also especially suited to tracing the gradual construction of new abilities, the slow evolution of initially primitive forms through various intermediate steps to full maturity. How, for example (to add a Piagetian instance to the earlier example of negation), does the neonate's primitive grasping reflex eventually become the skilled, visually directed reaching of the older infant? Finally, the intensive study of the same children over time may be especially helpful when it comes to interpreting behavior—that is, attempting to move beyond the surface behavior itself to some conception of the underlying basis for it (a cognitive structure? linguistic rule? individually learned response? or what?). In most research the investigator sees the participants for the first and only time when they appear for testing, and the investigator's ability to make sense of their behavior is dependent on this very brief interaction. Piaget, however, had been studying the same children literally since birth, and his extensive knowledge of each child's background gave him an excellent basis for interpreting any particular behavior from the child.[1]

A last argument in support of the longitudinal approach is of a more negative sort. The main alternative to the longitudinal design is the cross-sectional design, yet the cross-sectional design is also subject to a number of criticisms. Possible problems with cross-sectional studies are the subject of the next section.

Cross-Sectional Designs

A cross-sectional study tests different people at different ages. For this reason, the cross-sectional approach cannot measure age changes directly, nor can it answer questions about individual stability over time. As we saw, these limitations of the cross-sectional approach provide a primary motivation for longitudinal study.

There are other possible problems. Because cross-sectional studies test different samples at different ages, the possibility of *selection bias* arises. Perhaps the groups being compared differ not just on the independent variable of interest (in this case age) but in other ways as well, and it is these other differences that produce differences on the dependent measures.

The issue of selection bias was discussed briefly in chapter 2 when I considered the special nature of age as an independent variable. As noted there, the goal is not to rule out all differences between groups other than chronological age, but just those differences that are not naturally associated with age. I noted too that in most cases the decision about what to match is fairly obvious—for example, sex, race, social class, IQ. What must be added now, however, is

that actually achieving the desired matching may not always be easy. Developmental researchers typically draw samples of different ages from quite different sources—newborns from a hospital nursery, infants from parents who respond to solicitations to participate, preschoolers from preschools or day care centers, children between 5 and 11 from elementary schools, adolescents from junior high schools or high schools, adults from college classes. The populations served by these different settings may differ in a number of ways. Thus, even though the researcher may realize the importance of matching, selecting groups that are in fact comparable may prove difficult.

Bias can also occur in the form of *selective dropout* from the study. An initial equivalence between groups may quickly vanish if some participants drop out before testing is completed. The problem is not simply that there may be more dropouts at one age than another. The problem, rather, is the same one identified for longitudinal studies: People who drop out may be different from those who remain in. Thus, once again it is the "selective" part of selective dropout that threatens validity.

It is not hard to imagine situations in which selective dropout might bias comparisons between different ages. Suppose that we are doing a study of preschool children. We divide our sample into younger (2½ to 4) and older (4 to 5½) children, thus giving us two groups to compare. Our procedure is a fairly demanding one, requiring the child to process a variety of instructions and to continue to respond appropriately for a lengthy period. Not all preschool children are capable of such responding, and some are therefore lost from the study. The odds are strong that more children will be lost from the younger group than from the older group. The odds are strong also that those children who are lost will be, on the average, the less competent members of the sample. If so, we will end up with two noncomparable groups: a fairly representative sample of older children, and a nonrepresentative, biased-toward-superior sample of younger children. Clearly, any such differential dropout would decrease the chances of finding an improvement in performance with age.

Let us return to the issue of initial selection of participants. I have twice stated that decisions about what to match when comparing different age groups are generally straightforward. It is time now to consider the exceptions implied by the qualifier "generally."

Uncertainties about what should be matched are most likely when there is a wide separation between the ages being compared and thus many ways in which the groups potentially differ. They loom largest, therefore, in research comparing elderly adults with younger samples. Perhaps the most obvious example of this point is the variable of educational level. The average amount of schooling completed is greater now than it was 50 or 60 years ago. Suppose, then, that we wish to do a study comparing 25-year-olds and 75-year-olds. If we sample randomly at each age, our younger sample will be more highly educated than our older sample. We would have, then, a confounding of age and educational level. If we restrict our older sample to the more highly educated individuals, we will achieve comparability in educational level, but at the cost of selecting a nonrepresentative, positively biased older group. Neither solution is very satisfactory; perhaps the best course, if the researcher has the resources, is to incorporate both approaches. The main point, however, is that age and educational level, at this point in history, are unavoidably confounded in any attempt to compare adults of different ages.

The point just made about matching is actually part of a larger point concerning cross-sectional designs. I noted earlier that the longitudinal approach to studying age differences involves an inevitable confounding of age and time of testing. I will add now that the cross-sectional approach involves an inevitable confounding of age and generation or cohort. The samples in a cross-sectional study, being

different ages, must necessarily be born at different times and grow up under at least somewhat different sets of circumstances. The disparity in educational opportunities between today's 25-year-old and today's 75-year-old is just one example of such generational differences. Many other examples could easily be cited. Today's 75-year-olds lived through the Great Depression as children, encountered a world war in adolescence and another war in young adulthood, somehow survived into adulthood without TV or computers or many other commonplaces of modern life, and so forth. Suppose, then, that we find that 25-year-olds and 75-year-olds differ on our dependent measure. Should we attribute the difference to differences in age or differences in generation?

As with the other threats to validity discussed in this chapter, the extent to which the age-cohort confound is in fact a problem depends on the particular kind of study being done. Two factors are important in assessing the likelihood of cohort effects. One factor is the dependent variable under study. If our focus is on political attitudes or IQ test performance, then cohort effects may be quite important; indeed, such effects have been clearly demonstrated in the study of IQ (e.g., Schaie, 2005). If our focus is on heart-rate change or visual acuity, then cohort effects are much less likely to be important. In general, the more "basic" and "biological" a dependent variable appears, the less likely it is to vary across cohorts. Note, however, that there can almost always be dispute about how "basic" and cohort-general a particular variable is. Perhaps, for example, visual acuity actually does vary across generations as a function of changes in factors such as adequacy of artificial lighting or presence of TV during the formative years.

The other factor to consider is the age spread of the sample. Cohort effects are most obviously a problem in studies with widely separated age groups. Indeed, the issue of cohort differences first arose in research comparing young adult and old adult samples, and it is still most often discussed in that context. At the other extreme, the child psychologist who compares 3- and 4-year-olds probably does not need to worry about the fact that one group was born in 2000 and the other in 2001. Samples within the span of childhood can usually be assumed to belong to the same generation. Even here, however, doubts may arise. What about a comparison between pre–Sesame Street and post–Sesame Street generations? What about a comparison between a computers-since-kindergarten grade-schooler and a computers-since-sixth-grade adolescent? We live in a time of rapid cultural and educational change, and these changes may affect at least some between-age comparisons even among child samples.

The final problem to be noted is that of **measurement equivalence**. If we wish to compare the level of a particular behavior or particular ability in different age groups, then we need a procedure that can accurately tap the behavior or ability at each of the ages being studied. Often, however, a test that is appropriate for one age may not be appropriate for another age. A test of classification skills, for example, may be a fine indicator of such skills among 7-year-olds but may be too verbally demanding for many 4-year-olds. If so, the test may measure different things at the two ages: classification at age 7 and vocabulary at age 4. Note that the test would still reveal a genuine and perhaps important difference between the two ages: 7-year-olds really do perform better on this measure than do 4-year-olds. But the basis for the difference might not be the one that the investigator is seeking to study.

The problem of measurement equivalence is not limited to cross-sectional studies. The issue arises in any comparison of different ages; thus it applies with equal force to the longitudinal approach. The particular form of the equivalence problem, however, is likely to be different in longitudinal than in cross-sectional studies. Consider the longitudinal study of aggression (e.g., Cairns, Cairns, Neckerman, Ferguson, & Gariepy, 1989). The investigator who studies

aggression in a group of children at age 4 and again at age 12 is unlikely to be interested simply in comparing levels of aggression at the two ages. If levels of aggression *were* the focus, then serious problems would arise from the fact that the forms that aggression takes and the circumstances under which it occurs are quite different at age 12 than at age 4. The fact that the same children are being studied over time, however, probably means that the real interest of the longitudinal investigator is the *stability of individual differences* in aggression as children develop. The question, in other words, is whether children who are relatively high or low in aggression at age 4 are also relatively high or low in aggression at age 12. A child may be high in aggression at both 4 and 12 even though the frequency and forms of the behavior have changed greatly. This focus on relative standing within a group, rather than absolute level of response, provides a partial solution to the measurement-equivalence problem. Note, however, that it is still necessary to have valid measures of aggression at both ages.

More Complicated Designs

A clear message from the preceding discussion is that both the longitudinal approach and the cross-sectional approach suffer from various limitations. Table 3.1 summarizes the problems that have been discussed. Some of these problems are at least in principle avoidable—for example, the possibility of selection bias in cross-sectional research. Some of the problems, however, are intrinsic to the longitudinal and cross-sectional designs and hence can never be ruled out. Specifically, it is impossible ever to avoid the confounding of age with generation in the cross-sectional approach or the confounding of age with time of measurement in the longitudinal approach.

These limitations of the traditional longitudinal and cross-sectional designs have been much discussed in recent years, and they have motivated the development of several new procedures for studying changes with age. Because these new procedures have thus far been applied most often in studies of old age, I defer the main discussion of them until the chapter on aging (chapter 14). A brief introduction is possible here, however.

Figure 3.1 provides a schematic summary of the two designs discussed thus far. The body of the figure shows the ages that would be obtained from the various combinations of date of birth and year of measurement. A longitudinal design would be represented by any of the rows in the figure. In this case, a sample of people born at the same time is studied

Table 3.1 Problems With Longitudinal and Cross-Sectional Designs

Longitudinal	*Cross-Sectional*
• Practical difficulties (expensive, time-consuming) • Possible obsolescence of measures • Possible nonrepresentative samples • Limitation to a single cohort • Possible selective drop-out • Effects of repeated testing • Difficulty in establishing equivalent measures • Confounding of age and time of measurement	• No direct measure of age changes • Inapplicable to issues of individual stability • Possible selection bias • Possible selective drop-out • Difficulty in establishing equivalent measures • Confounding of age and time of birth (cohort)

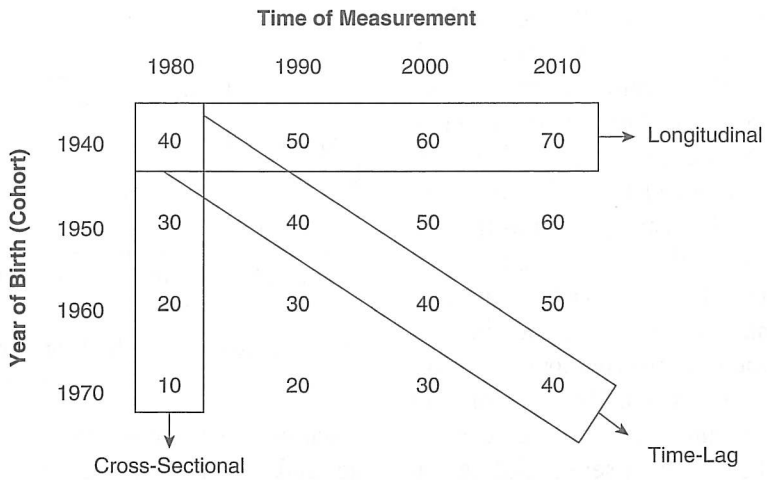

Figure 3.1 Examples of longitudinal, cross-sectional, and time-lag designs

NOTE: Numbers in the body of the figure indicate ages.

repeatedly across a span of ages. A cross-sectional design would be represented by any of the columns in the figure. In this case, separate samples born in different years are studied at the same time.

Figure 3.1 also includes a third design not yet discussed: the **time-lag design**. A time-lag design would be represented by any of the diagonals in the figure. Thus, we might study a sample of 40-year-olds in 1980, another sample of 40-year-olds in 1990, another sample of 40-year-olds in 2000, and another sample in 2010. Clearly, the time-lag design cannot give us direct information about age changes or age differences, because only one age group is studied. What it can do, however, is provide information about factors that may confound the age comparisons in longitudinal or cross-sectional designs. Specifically, if we find differences among our samples of 40-year-olds, then we know that these differences must reflect either generational factors (the main confound in the cross-sectional design) or time-of-measurement factors (the main confound in the longitudinal design) or, of course, some combination of the two factors. The fact that we cannot be certain which factor is important indicates that the time-lag design suffers from its own brand of confounding: a confound between generation and time of measurement.

Time-lag designs are not used very often. Occasionally, however, time-lag comparisons become available simply through the natural historical course of research. Piaget, for example, first studied young children's understanding of conservation during the 1930s and 1940s. When the Piagetian approach became popular decades later, the result was a second wave of conservation studies during the 1960s and 1970s. The conjunction of original research and newer research constitutes a time-lag comparison: two groups of children of the same age but born at different times and tested at different times. In this case neither cohort nor time of measurement appears important, since children of the 1970s responded to conservation tasks in essentially the same way as had children of the 1930s.

IQ tests provide a second, and contrasting, example. As I discuss more fully in chapter 12, when a child takes an IQ test, the child's performance is compared with that of children of the same age who formed the normative sample when the test was first developed. Thus, a

10-year-old (say) taking an IQ test in 2005 might be compared with 10-year-olds who took the same test in 1985. A common finding from such comparisons is a slight improvement in average performance over time, a phenomenon known as the *Flynn Effect* (Flynn, 1998). Because age is held constant, we know that the improvement must reflect effects of either time of measurement or (more probably) cohort.

Longitudinal, cross-sectional, and time-lag designs are sometimes referred to as the "simple" developmental designs. They are simple in comparison to the alternative, the decidedly *not* simple sequential design. A **sequential design** consists of a combination of longitudinal, cross-sectional, and time-lag components within a single research design, the goal being to tease apart effects of age, cohort, and time of measurement. The components can be combined and analyzed in different ways, and thus there are several different kinds of sequential design. In this chapter I discuss, briefly and generally, two of the kinds. In chapter 14 I return to the topic of sequential designs, but in this case in concrete rather than hypothetical form with a discussion of the major research program to date to employ such designs.

A word first about the logic of sequential designs. Ideally, what we would like to be able to do is to examine the contribution of all three potentially important factors—age, cohort, and time of measurement—within a single analysis. Unfortunately, doing so is precluded by the interdependencies among the three factors; as soon as any two of them are set, the levels of the third follow automatically. Once we have decided, for example, that we wish to study particular ages and particular cohorts, the times of measurement are necessarily fixed; they are whatever values we need to get the desired conjunction of ages and cohorts. The result of these interdependencies is that only two of the three factors can function as independent variables within the same analysis. The various sequential designs vary in which of the factors they concentrate on. In the first of the examples discussed next, the independent variables are age and cohort; in the second, the independent variables are age and time of measurement.

Year of Birth (Cohort)	Time of Measurement			
	1980	1990	2000	2010
1940	40	50	60	
1950		40	50	60

Figure 3.2 Example of a cohort-sequential design

NOTE: Numbers in the body of the figure indicate ages.

Figure 3.2 presents a *cohort-sequential design.* A cohort-sequential design selects samples from different cohorts (i.e., years of birth) and tests them repeatedly across the same span of ages. It consists, therefore, of two (or more) overlapping longitudinal studies. In the example shown in the figure, groups born in 1940 and 1950 are tested three times across a 20-year span. Such a design offers several advantages in comparison to a standard longitudinal or cross-sectional approach: (a) Because different times of measurement are used, the age variable is not confounded with the cohort variable (the prime confound in cross-sectional studies). (b) Because samples are drawn from different years of birth, the longitudinal comparisons are not limited to a single generation or cohort. (c) Because different age groups are tested at each time of measurement, there is a cross-sectional as well as a longitudinal dimension. (d) Because the same age group is represented at different times of measurement, there is a time-lag dimension as well. There is, in short, more information than in a standard design, and thus more chance to disentangle the contributions of various factors.

Figure 3.3 shows a *time-sequential design.* A time-sequential design consists of two (or more) cross-sectional studies carried out at different times of measurement. In the example, samples of 40-, 50-, and 60-year-olds are compared in 1990,

Year of Birth (Cohort)	Time of Measurement 1990	2000	2010
1930	60		
1940	50	60	
1950	40	50	60
1960		40	50
1970			40

Figure 3.3 Example of a time-sequential design

NOTE: Numbers in the body of the figure indicate ages.

in 2000, and in 2010. The samples at the different times may be either independent (i.e., different people at the three test occasions) or the same (if the original participants are followed longitudinally). This design has the same general virtue as the cohort-sequential design: It provides considerably more information than the simpler designs. A specific strength is that it unconfounds the variables of age and time of measurement (the prime confound in longitudinal studies). If independent samples are studied at the different times, this method also avoids some of the problems found with longitudinal designs (selective dropout, effects of repeated testing).

As noted, I consider these designs more fully in chapter 14. Two points can be made here, however. First, it is obvious that sequential designs, though more informative, are also considerably more costly—in time, effort, and money—than the simpler cross-sectional and longitudinal designs. Execution of the design pictured in Figure 3.3, for example, would require 20 years and (in the independent-samples version) nine groups of participants. In any research project there are a large number of things that would be desirable to do, only a subset of which it is actually possible to do. The best designs are always those that can actually be carried out.

The second point concerns the threats to validity that apply in the traditional designs—namely, the confound of age with cohort in the cross-sectional design and the confound of age with time in the longitudinal design. It is a point that applies to threats to validity in general. Although it is true that the goal of good research design is always to minimize threats to validity, the fact is that it is never possible to rule out every conceivable alternative explanation for one's findings. The question then becomes how plausible the alternative explanations are. And for much developmental research, especially within the span of childhood, the possibility of cohort or time-of-measurement effects is simply not plausible. In such cases, cross-sectional or longitudinal methods may produce—and indeed have produced—data about changes with age that are of considerable validity and use.

FOCUS ON

Box 3.1. The Microgenetic Method

As we have seen, one argument in support of longitudinal studies is that they provide a direct measure of change that is lacking in the cross-sectional approach. Most longitudinal studies, however, are limited to documenting the results or products of change. That is, they tell us what the individual is like at time 1 and time 2 and time 3. But they do not tell us *how* the changes from 1 to 2 to 3 come about, and they do not tell us about any intermediate states between 1 and 3 that are not represented in the times of measurement. In Robert Siegler's words, longitudinal studies provide snapshots of development (Siegler, 1996).

The microgenetic method, in contrast, is intended to provide something more akin to a movie of development. The **microgenetic method** refers to repeated, high-density observations of the

(Continued)

(Continued)

behaviors being studied across a period when change is occurring. It is therefore a form of longitudinal research, in that repeated observations are made of the same individuals across the time period of interest. In contrast to a standard longitudinal study, however, the observations are both more frequent and more closely spaced, and there is an emphasis on capturing not just levels of performance but processes of change.

Let us consider an example. Siegler and Jenkins (1989) were interested in how children develop strategies to solve simple arithmetic problems, such as 3 plus 5. Table 3.2 shows some of the strategies that children might use. To test these possibilities, Siegler and Jenkins performed a microgenetic study with 10 four- and five-year-olds who had not yet developed any of the more advanced strategies. The children participated in three experimental sessions per week across a period of 11 weeks. During each session they attempted to solve seven problems, and across sessions the problems gradually increased in complexity. Videotapes were made of the children's performance, and they were also directly questioned at various points about the strategies they were using. Through this approach, Siegler and Jenkins were able not only to document the gradual and often halting emergence of new strategies but also to identify precursors to and conditions for strategy change.

Table 3.2 Children's Strategies for Solving Simple Addition Problems

Strategy	*Typical use of strategy to solve 3 + 5*
Sum	Put up 3 fingers, put up 5 fingers, count fingers by saying "1, 2, 3, 4, 5, 6, 7, 8."
Finger recognition	Put up 3 fingers, put up 5 fingers, say "8" without counting.
Short-cut sum	Say "1, 2, 3, 4, 5, 6, 7, 8," perhaps simultaneously putting up one finger on each count.
Count-from-first-addend	Say "3, 4, 5, 6, 7, 8" or "4, 5, 6, 7, 8," perhaps simultaneously putting up one finger on each count.
Min (count-from-larger-addend)	Say "5, 6, 7, 8," or "6, 7, 8," perhaps simultaneously putting up one finger on each count beyond 5.
Retrieval	Say an answer and explain it by saying "I just knew it."
Guessing	Say an answer and explain it by saying "I guessed."
Decomposition	Say "3 + 5 is like 4 + 4, so it's 8."

SOURCE: From *How Children Discover New Strategies* (p. 59), by R. S. Siegler and E. Jenkins, 1989, Hillsdale, NJ: Erlbaum. Copyright © 1989 by Lawrence Erlbaum. Adapted with permission.

In discussing the results of this and other microgenetic studies, Siegler (1996) identifies five issues related to cognitive change for which microgenetic techniques can provide valuable data. Such techniques can inform us about the *path* of cognitive change: the sequences and levels through which children move in acquiring new knowledge. They can provide information about the *rate* of change: how quickly or slowly children master different forms of knowledge. They speak to the issue of *breadth* of change: when children acquire a new competency (such as a particular

arithmetical strategy), how narrowly or broadly they apply it. They are relevant to the question of possible *variability* in the pattern of change: Do all children follow the same route in mastering a new concept? Finally, microgenetic methods can provide information about the *sources* of change: the experiences and processes through which children construct new knowledge.

The discussion to this point may have given the impression that the microgenetic approach is specific to the study of cognitive development. In fact, most applications of the approach to date have been in the cognitive realm. Among the other topics that have been studied are mnemonic strategies (e.g., Coyle & Bjorklund, 1997) scientific reasoning (e.g., Kuhn, 1995), problem solving (e.g., Chen & Siegler, 2000), and language (e.g., Ruhland & van Geert, 1998). The approach is not limited to cognitive outcomes, however; it has been applied, for example, to the study of early mother-infant interaction (Lavelli, Pantoja, Hsu, Messinger, & Fogel, 2005). Nor, as this last example illustrates, is it limited to older children.

As with all methods, the microgenetic approach is subject to possible criticisms and threats to validity (Miller & Coyle, 1999; Pressley, 1992). Perhaps the major concern is that the frequent, high-density observations may in themselves change the phenomenon being studied. Most 4-year-olds, after all, do not spend dozen of hours solving arithmetic problems and responding to explicit questions about what they are doing. Perhaps what occurs under such circumstances is in some ways different from the real-life processes that we are trying to capture.

As with any threat to validity, it is an empirical question whether this potential problem actually applies, and microgenetic researchers cite evidence that at least in some instances it does not (Kuhn, 1995; Siegler & Crowley, 1992). Perhaps the major argument in support of the approach is a more conceptual one, however. Understanding how change occurs is both one of the most fundamental and one of the most challenging questions in developmental psychology, and it is clearly desirable to have as many methods as possible with which to attack it. Microgenetic techniques are one such method.

Condition Comparisons

Within-Subject Versus Between-Subject Designs

I turn now to the question of how to make comparisons between two or more tests or experimental conditions. I noted earlier that two general approaches are possible: administering all tasks or conditions to the same participants or assigning different participants to different experimental groups. The former is labeled a *within-subject design;* the latter, a *between-subject design.* Because the discussion of these two approaches will involve much back-and-forth comparison, it is simpler to consider them together rather than separately.

How does an investigator decide whether to make comparisons within or between subjects? Just as in the longitudinal versus cross-sectional decision, matters of convenience may often play a role. Usually (with a qualifier to be noted shortly), a within-subject approach means that fewer participants are needed. Suppose, for example, that we have three tasks whose difficulty we wish to compare, and we know that we will need at least 20 respondents on each task to determine whether any differences in difficulty are present. If we opt for a between-subject approach, we will need at least 60 people to complete the study; with a within-participant approach, however, a mere 20 may suffice. Whenever the pool of possible participants is limited, the economy of a within-subject design may be attractive.

Considerations of convenience do not always fall on the side of the within-subject approach, however. The smaller sample size in a within-subject study is bought at an obvious

price—namely, more time spent with each participant, either in longer experimental sessions or in a greater number of sessions. Especially in work with young children, lengthy or repeated sessions may tax the child's motivation or endurance. Even if the investigator is not concerned about such demands on the child, the parents or school authorities may be. In such situations, a between-subject design, which minimizes the demands on any one child, may be the most sensible approach.

Statistical considerations may also affect the within versus between decision. The statistical tests appropriate for within-subject comparisons are somewhat different from those appropriate for between-subject comparisons. Furthermore, within-subject tests are often more powerful than between-subject tests—that is, more likely to reveal a significant difference if a difference does in fact exist (I discuss the notions of significance and power more fully in chapter 8). This greater power stems from the reduction in unwanted variance afforded by the within-subject design. Recall the earlier discussion of primary variance compared with secondary or error variance. As I noted then, a goal of good experimental design is to maximize primary variance, or variance associated with the independent variable, and to minimize unwanted variance from other sources. I noted too that the inevitable differences that exist among different participants are one source of unwanted variance. Use of the same participants for all experimental conditions reduces such variance and hence enhances the power of any comparisons made. The result is a greater likelihood that a difference of a given magnitude will achieve statistical significance.

Both between-subject and within-subject designs are subject to their own particular forms of bias. The obvious threat in between-subject designs is selection bias. Because different people are assigned to different conditions, the possibility will always exist that any differences that are found between conditions reflect preexisting differences among the participants and not a true effect of the experimental manipulations. This possibility does not arise in a within-subject design, in which each participant responds under each condition. Note that this advantage of within-subject over between-subject designs parallels an advantage discussed earlier for longitudinal compared to cross-sectional approaches.

There are two ways to try to rule out possible selection biases in a between-subject design (recall Table 2.4). One is to match participants on variables of potential importance. I consider the pros and cons of matching shortly. The other is the approach discussed in chapter 2: random assignment of participants to different groups. If the sample size is sufficiently large and if the assignment to conditions is truly random, then preexisting differences among participants should be controlled and confounding of subject and condition avoided. As argued in chapter 2, the logic of the random-assignment approach is impeccable; the challenge is to ensure that the two "if" questions really do receive positive answers.

The most obvious threat to the validity of within-subject designs concerns the possible effects of extended testing. Consider a study in which the researchers wish to compare the relative difficulty of several cognitive tasks. They decide to use a within-subject design, in which every child receives every task. Because presenting several tasks takes time, the children may well become increasingly tired or bored as they move through the series of problems. If so, performance may be poorer on later tasks than on earlier ones. Alternatively, the children may be somewhat timid or confused at the start of the study but become increasingly relaxed and confident as the testing proceeds. In this case, performance may be better on later tasks than on earlier ones. In either case, the effects stemming from the repeated testing would cloud the intertask comparison that is the researcher's real interest.

"Warm-up" or "fatigue" effects of the sort just described fall under the general heading of order effects. The term **order effect** refers to any general tendency for response to change in a systematic fashion from early in a session to later in a session. Usually, the systematic change is either a general improvement or a general deterioration in performance.

Another potential problem in within-subject designs is the possibility of carryover effects. A **carryover effect** occurs whenever response to one task or condition varies as a function of whether another task or condition precedes or follows it. Let us try a simple example to clarify this rather forbidding definition. Imagine that we wish to compare the relative difficulty of two tasks: A and B. We will suppose that either task, presented in isolation, elicits 50% correct responses from our sample. It turns out, however, that when task A is presented first, experience with A suggests a helpful means of attacking task B; correct responses to B consequently rise to 70%. In contrast, when task B is presented first, experience with B suggests a means of solution that is maladaptive for task A; correct responses to A consequently fall to 30%. Note that in this case there is no general improvement or decline across the experimental session; rather, the finding is that response to one task depends on whether that task is presented before or after the other task. Although the specific mechanism may differ, the general import of order effects and carryover effects is the same: complications in the interpretation of task or condition comparisons.

Problems created by order effects are most likely when the experimenter adopts a constant order of presentation for the different tasks or conditions. An obvious prescription follows: Whenever comparisons among tasks or conditions are of interest, a single order of presentation should be avoided. There are two alternatives to constant order. One alternative is to randomize the order of tasks or conditions. In certain cases, perhaps especially when the number of tasks is large, randomization may be the most sensible approach. Generally, however, a better alternative than randomization is **counterbalancing** of the order of presentation. Counterbalancing is conveyed more easily through example than through definition; a simple example is given in the upper left portion of Table 3.3. As can be seen, counterbalancing is a method for distributing a particular task or condition equivalently across the various possible ordinal positions. Thus, in the example, task A occurs equally often in the first, second, and third positions; furthermore, it precedes and follows tasks B and C equally often in each position. The counterbalancing in this case is complete—that is, all possible permutations of the three tasks are used. Clearly, with more tasks the number of possible permutations increases; with four tasks there are 24 permutations (these are shown in the upper right part of Table 3.3), and with five tasks there are 120 permutations. In such cases, complete counterbalancing may not be feasible; it is still possible, however, to select a subset of orders that will provide a reasonable degree of balancing. Examples of such orders for four-task and five-task studies are shown in the bottom part of Table 3.3.

Counterbalancing has two advantages over randomization. First, it ensures that there is no confounding of task and order of presentation, an outcome that cannot be ensured by randomization alone. Second, because confounding has been ruled out, it permits the researcher to compare the different orders of presentation and tease out any order effects or carryover effects that may be present in the data. Note, however, that such effects are likely to be identifiable only if the sample size is reasonably large and each order is represented sufficiently often. This point provides the qualifier for the earlier statement that within-subject designs require fewer participants than between-subject designs: Whenever possible effects of order are of interest, then the *N* necessary for a within-subject study may increase substantially.

Table 3.3 Examples of Complete and Partial Counterbalancing

Complete balancing	
Three tasks	*Four tasks*
ABC	ABCD BACD CABD DABC
ACB	ABDC BADC CADB DACB
BAC	ACBD BCAD CBAD DBAC
BCA	ACDB BCDA CBDA DBCA
CAB	ADBC BDAC CDAB DCAB
CBA	ADCB BDCA CDBA DCBA
Partial balancing	
Four tasks	*Five tasks*
ABCD	ABCDE
BDAC	BEDAC
CADB	CAEBD
DCBA	DCBEA
	EDACB

Thus far I have discussed a number of factors that a researcher can weigh in deciding between a within-subject and a between-subject design. In some cases, however, there is no decision to make; the nature of the research question dictates the design to be used. Specifically, whenever the interest is in within-subject patterning of performance, then a within-subject design is necessary. Whenever the interest is in definite and persistent change as a result of the experimental manipulation, then a between-subject design is necessary. I now elaborate on both of these points.

The argument with respect to within-subject patterning parallels an argument that was made earlier in support of longitudinal designs. There, we saw that questions concerning individual consistency or individual change over time require a longitudinal approach that studies the same people as they develop. Similarly, questions concerning the relation between two or more measures at any given time require a within-subject approach that studies the same people across the different measures. Suppose, for example, that we wish to know whether children's social skills relate to their popularity with peers (Cillessen & Bellmore, 2002). Clearly, we cannot assess social skills in one group of children and popularity in another group; rather, we must have both measures for all children. Or suppose (to return to an earlier example) that we wish to know whether children's IQs relate to their grades in school. We cannot assess IQ in one sample and grades in another sample; again, we must have both measures for all children. Examples such as these illustrate a prime rationale for within-subject study: to identify interrelations and patterning in development.

The argument with respect to manipulations that produce change is in some respects similar to points made earlier concerning testing effects in longitudinal designs and carryover effects in within-subject designs. The essential point is that administering one task or experimental condition may change participants in a way that makes them unusable for other tasks or conditions. Suppose that we wish to compare the effectiveness of several different methods of training conservation concepts (e.g., Smith, 1968). We select a group of nonconservers and

administer training condition A. We can hardly then take the same children and administer condition B, for if condition A is at all effective, many of the children will no longer be nonconservers! The same argument applies to any research whose goal is to bring about lasting change in its participants—intervention programs for so-called disadvantaged children, therapy programs for disturbed children, parent-education programs for new parents, and so forth. In each case, if we wish to compare the effectiveness of different programs we need a between-subject design that assigns different participants to the different approaches. Note too that the argument is not limited to attempts to produce sweeping changes à la intervention or therapy; the argument may apply to more focused, short-term changes as well. Suppose, for example, that we wish to know whether inducing children to use verbal rehearsal helps them on a short-term memory task (e.g., Ferguson & Bray, 1976). We cannot expect that children who have been taught such a strategy will necessarily abandon it once we remove the instruction to verbalize; rather, if we want a rehearsal–no rehearsal comparison we need to test separate groups of participants.

There is a possible objection to this last example and the conclusion drawn from it. In the verbal-rehearsal case our interest is not in the relative effectiveness of several different treatments; the interest, rather, is in whether a single treatment will lead to improvement over a no-treatment baseline. It is true that we cannot apply the treatment and later expect to get a measure of performance in its absence. But why not proceed in the opposite order—that is, first measure the children's natural level of memory performance, apply the treatment, and then measure memory performance again? Doing so would give us an example of what Campbell and Stanley (1966) labeled a One-Group Pretest-Posttest Design. The rationale would be that any improvement in performance from the pretest to the posttest would reflect the effects of the intervening treatment. If this rationale is valid, then there is no need to set up separate groups of participants.

In certain simple situations this kind of One-Group design may be sufficient for the researcher's purposes. Generally, however, it is not. The weakness of such a design should be evident from the earlier discussion of experimental control: It permits a confounding of the experimental treatment with a number of other factors that might produce a pretest-to-posttest change.

Let us take intervention programs as the example to make this point. Imagine that we find a group of at-risk 4-year-olds, give them a test of "academic readiness," subject them to a 1-year intervention program designed to enhance academic skills, readminister the academic-readiness test at the end of the program, and find that scores have improved significantly. Evidence for the effectiveness of our program? Not necessarily. It may be that the improvement results from natural biological-maturational changes as the children age from 4 to 5—Campbell and Stanley's **maturation** variable. It may be that the improvement results from other events in the children's lives during the course of the program—Campbell and Stanley's **history** variable. It may be that the improvement results from practice effects gained from taking the initial pretest—Campbell and Stanley's *testing* variable. Or it may be that the improvement results from the natural upward movement of initially low scores upon retesting—Campbell and Stanley's *regression* variable. None of these rival hypotheses can be ruled out with a One-Group design; all could be ruled out if we included a separate, no-treatment control group.

This comparison of within-subject and between-subject approaches could use some summarizing. Table 3.4 lists the various pros and cons that have been discussed as relevant to the within-subject versus between-subject decision.

Both between-subject and within-subject approaches come in a variety of forms. I turn next to two of the most important variants:

Table 3.4 Relative Merits of Within-Subject and Between-Subject Designs

Factor	*Comparison of designs*
Convenience	Fewer participants with within; less time per participant with between
Statistical tests	Generally more powerful with within
Order or carryover effects	A problem with within; not a problem with between
Possible selection bias	A problem with between; not a problem with within
Focus on within-subject patterning	Must have within; impossible with between
Focus on procedures that produce lasting change	Must have between; impossible with within

matched-groups designs (a form of between-subject research) and time-series designs (a form of within-subject research).

Matched-Groups Designs

A clear comparison of different experimental conditions requires that the participants assigned to the different conditions be equivalent at the start of the study. I have discussed two methods for creating such equivalence: random assignment of different participants to different conditions, and repeated testing of the same participants across all conditions. I now add a third possibility: use of a **matched-groups design**, in which participants are matched prior to their assignment to conditions.

The notion of matching was introduced briefly in chapter 2 in the discussion of random assignment. The question was posed then: Why settle for random assignment; why not ensure equivalence by matching the participants assigned to different groups on all the characteristics that might be important? A little thought will suggest an answer: We can never identify all the characteristics that might be important variables, and even if we could do so, we could never get the necessary data and achieve the necessary matching. Matching is always necessarily partial matching. Still, partial matching is presumably better than none; why not utilize it? It turns out that doing so has both advantages and disadvantages.

Since the most often matched-for variable in research with children is probably IQ, I will use IQ as our example. If we wish to match children on IQ, we must first administer IQ tests to all of our potential participants (or, perhaps, go to the school files and obtain already-collected IQ data). We then group together children with identical or close-to-identical IQ scores. The number of children in a group will depend on the number of experimental conditions—pairs of children if there are two conditions, trios if there are three conditions, and so forth. Working within these same-IQ groups, we then randomly assign different children to the different experimental conditions. Note, therefore, that random assignment remains important even in a matched-group design. Note also, however, that the initial matching on IQ ensures what randomization alone cannot ensure: that the experimental groups end up with equal IQs.

The great strength of the matching approach is that it does provide such exact and certain control for variables that might otherwise bias results. If IQ really does relate to performance on our dependent variable, then it is critical

that there be no confounding of IQ and experimental condition. Matching also has certain statistical advantages. In much the same way as within-subject designs, a matched-group design reduces unwanted variance and hence increases the power of the statistical tests. Matching may be especially helpful, therefore, when the power to detect effects is known to be low—when the sample size is limited, for example, or when the expected differences between groups are small.

The main disadvantages of matching revolve around the question: Is it worth it? Matching typically requires a substantial investment of effort on the investigators' part, especially if they must pretest all of the potential participants (as opposed to relying on already-existing data). If the matched-for variable is not in fact related to performance on the dependent variable, then the matching will have added nothing. If the sample size is large and random assignment is used, the groups will probably end up equivalent anyway, and again matching will add nothing. The point here has to do with efficiency of effort. Any research project involves selection of a relatively few specific procedures from a much larger pool of potentially informative procedures. To devote a portion of one's limited time and effort to procedures that do not enhance the study is simply bad research practice.

In addition to the possible waste of effort, matching can sometimes create particular problems. In some cases, administering the matching pretest may bias the participants' responses to the later test of interest (Campbell and Stanley's reactivity variable). Perhaps, for example, being taken out of their classroom and given an IQ test is anxiety-arousing for some children and makes them suspicious of the friendly tester who later invites them to "come play a game." The tester's attempt to create a game-like atmosphere for the measures may therefore come to naught, and the validity of the study may be affected. Matching can also sometimes result in loss of participants. If participants are matched in the manner described earlier, then the unit becomes the matched group rather than the individual child—for example, the trios of matched-for-IQ children in a study with three experimental conditions. If any member of a trio is lost from the study for any reason, then the other two must be eliminated as well. Whenever dropout seems likely, matching may turn out to be a costly decision.

There is one situation in which matching is a tempting but usually unsound procedure. It is the case in which the investigator wishes to bring about equality in initially unequal groups of subjects. We saw an earlier example in the discussion of the differing educational levels of young adult and elderly adult samples. Consider another example, drawn from Neale and Liebert (1986). Imagine that you are interested in determining whether high school graduates achieve greater economic success in later life than do people who drop out of school. You are concerned, however, that the two groups differ in average IQ—perhaps a mean of 105 for the graduates and 90 for the dropouts. This IQ discrepancy provides an alternative explanation for any group differences you may find: Perhaps differences in economic success are simply reflections of differences in cognitive ability and have nothing to do with completing or not completing high school. You decide, therefore, to match the graduate and dropout groups in IQ. With IQ ruled out as a possible cause, you can more confidently attribute differences in economic success to the benefits of finishing school.

There are at least three problems with such a procedure, two of which I discuss here and one of which I defer for later consideration. First, the procedure guarantees some limit in external validity, since at least one of the two groups will not be completely representative of its parent population (i.e., either unusually high-IQ dropouts or unusually low-IQ graduates). Second, matching the groups on one dimension may systematically unmatch them on other dimensions that are related to finishing school.

Suppose, for example, that you decide to set the mean IQ for both groups at 90. In this case you will have a typical group of dropouts, but your group of graduates—precisely because they have succeeded despite mediocre IQs—is likely to be above average in other characteristics (e.g., motivation, family support) that contribute to school success. Conversely, setting the mean IQ at 105 will yield a typical group of graduates; now, however, the dropouts will be below average on the other determinants of performance in school. Matching the groups on one dimension may thus have the unintended effect of making them in general less similar rather than more similar.

The third problem with matching unequal groups is that it can lead to effects associated with statistical regression. In chapter 4 I will discuss how regression can be a problem in matched-groups designs in the context of a general consideration of statistical regression as a threat to validity.

Time-Series Designs

Time-series designs are easiest to introduce with an example. The goal of a project by Hall et al. (1971) was to reduce the disruptive "talking-out" behavior of a 10-year-old boy in special education classes. Their study, like all time-series research, proceeded in several phases.

The first phase was a **baseline** period: measurement of the initial frequency of the target behavior under normal classroom conditions. As Figure 3.4 shows, the behavior was indeed frequent—three to five outbursts across each of five 15-minute sessions. Following the baseline came the first application of the experimental treatment: The teacher ignored instances of talking-out behavior but paid increased attention when the child was behaving productively. The apparent result of this "contingent attention" regimen was a dramatic drop in talking out, as shown in the second part of Figure 3.4. The experimental treatment phase was followed by a reinstatement of the baseline conditions, during which the talking-out behavior shot back up in frequency. Finally, in a fourth and last phase the contingent attention treatment was restored, and talking out returned to a low level.

The Hall et al. study (1971) is an example of an *A-B-A-B* time-series design: an initial baseline phase (the first A), followed by an initial application of the experimental treatment (the first B), followed by a second baseline (the second A), followed by a second experimental treatment (the second B). Let us work through the rationale for each of these phases. The first baseline is clearly necessary—we need to know the initial, preintervention level of the target behavior to determine any effects of the experimental treatment. The first intervention phase is of course also necessary. But why not stop once we have shown that the experimental treatment reduces the behavior—that is, why go beyond an A-B design? The answer is that a simple A-B design would be subject to all the threats to validity (maturation, history, etc.) discussed earlier with respect to within-subject designs in general. These threats are especially difficult to rule out when we have only one participant for the research, as is the possibility that the change is simply random fluctuation that would have occurred even without the treatment. If we can show that the target behavior reemerges when we withdraw the treatment, we can be more certain that the treatment really was responsible for the decline. If we can show that a second administration of the treatment is associated with a second decline, then we can be even more certain that the treatment is the causal agent. And, of course, there are also pragmatic and ethical reasons for adding the final B phase in the A-B-A-B design. The goal, after all, is to reduce the undesirable behavior; thus, we do not want to end the study with the behavior still at its height.

It should be clear from this description that a **time-series design** is a special form of within-subject research. It is within-subject in the sense that each participant receives every level of the independent variable and comparisons are made

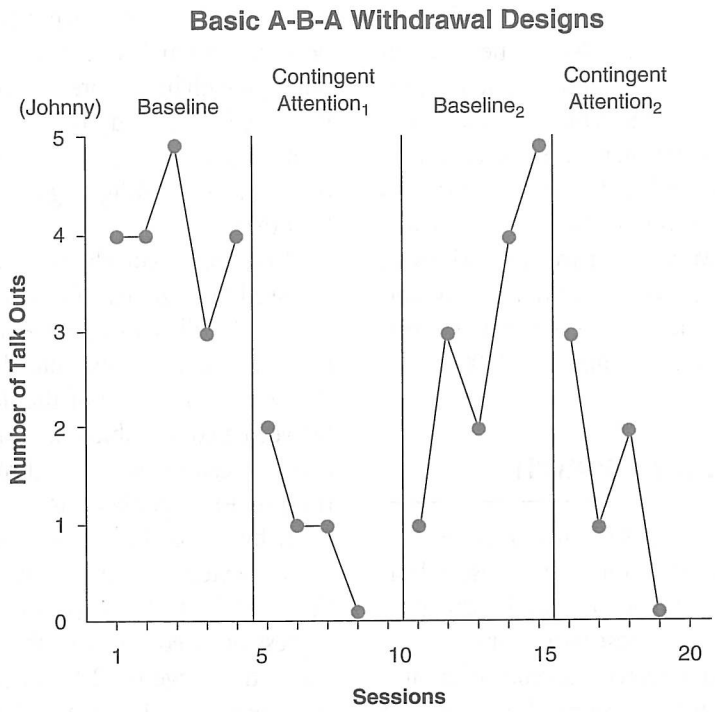

Figure 3.4 Example of a time-series design

NOTE: The level of the target behavior (talking out) varies as a function of the presence or absence of the experimental treatment.

SOURCE: From "The Teacher as Observer and Experimenter in the Modification of Disputing and Talking-Out Behaviors," by R. V. Hall, R. Fox, D. Willard, L. Goldsmith, M. Emerson, M. Owen, T. Davis, & E. Porcia, 1971, *Journal of Applied Behavior Analysis, 4*, p. 143. Copyright © 1971 by the Society for the Experimental Analysis of Behavior. Reprinted with permission.

within rather than between people. A time-series study also differs from the kinds of within-subject research discussed earlier in several respects, however. In most within-subject research, the levels of the independent variable represent different forms of some task or treatment (e.g., the easy-hard comparison in the Dufresne and Kobasigawa study); in a time-series study the levels are the presence or absence of the experimental treatment. In most within-subject studies, the comparisons are made within a single experimental session; in a time-series study, the comparisons are spread out over repeated sessions. Most within-subject studies sample and analyze groups of participants; many time-series studies (like the Hall et al., 1971, study) involve but a single participant. The time-series approach, in fact, constitutes the major source of designs for *single-subject research*—that is, research that attempts to identify effects of an experimental manipulation within a single participant. Finally—and related to these other points—time-series studies are often carried out for pragmatic purposes, the goal being to demonstrate the effectiveness of some intervention in ameliorating a problem behavior (as was true in the Hall et al. study). They are most often seen, therefore, in clinical or educational settings.

Time-series designs can involve complexities of both implementation and interpretation that I have not attempted to discuss. They can also

encompass naturally occurring time series (for example, variations in purchasing behavior in response to economic fluctuations) and not just, as in Hall et al. (1971), those that are experimentally created. And whether naturally occurring or experimentally induced, they can come in many forms in addition to the A-B-A-B design described here. More extensive discussions of time-series research can be found in Barlow and Hersen (1984), Cook and Campbell (1979), Kazdin (1998), and Velicer and Fava (2003).

Correlational Research

In chapter 1 I noted several contemporary, socially important issues for which research in developmental psychology can be informative. Let us return to one of these issues for an example of correlational research. McLeod, Atkin, and Chaffee (1972) were interested in possible effects of TV violence on aggression in children. They collected various measures of aggression in a sample of 6th-graders through 10th-graders. They also measured how much violent TV each child in the sample typically watched. Their interest was in whether there was a relation between watching violent TV and being aggressive—that is, did the children who watched the most violence on TV also tend to be the most aggressive? In their study (as in many other similar studies) there *was* such a relationship, an outcome compatible with the hypothesis that watching violent television promotes aggression.

The study by McLeod et al. is an example of **correlational research**. It is correlational because there was no manipulation of an independent variable. McLeod et al. did not experimentally control the type of TV that their sample watched, nor did they control the level of aggression that the children showed. Instead, both TV viewing and aggression were *measured* as they naturally occurred, the intent being to see whether scores on one index covaried with scores on the other. Such a relation might be positive, with high scores on one measure tending to go with high scores on the other. This was the case in the study by McLeod et al. Or the relation might be negative, with high scores on one measure tending to go with low scores on the other.

Outcomes in correlational research are often assessed through use of a **correlational statistic**, which will be discussed more fully in chapter 8. For now, let us note that a correlation statistic is a measure of the degree of relation between two variables; it ranges from −1 (a perfect negative relation) through 0 (no relation) to +1 (a perfect positive relation). In the study by McLeod et al., the correlations varied to some extent, depending on the age and sex of the sample and the particular measure of aggression used; most of the values, however, fell in the range of .2 to .3. Such correlations indicate a modest positive relation between TV violence and aggression.

Although correlation statistics are typically associated with correlational research designs, it is important to note that the statistic and the design are separable. Statistics other than correlations can be used to examine the results of correlational research. McLeod et al., for example, might have divided their sample into high, medium, and low TV watchers and then used *t* tests or analysis of variance to compare levels of aggression across the three groups. In this case the statistic would be different, but the design would remain correlational. Because of this independence of design and statistic, some researchers prefer the term *nonexperimental* for the kind of research at issue here. Whatever the label, the defining aspect of such research is that variables are simply measured, not experimentally controlled.

Correlation and Causation

One of the truisms of research is that correlation does not imply causation. That is, simply

from knowing that two variables are correlated, we cannot establish what causal relation, if any, holds between them. Thus, the results of the McLeod et al. study are compatible with the hypothesis that TV violence causes aggression, but the results cannot prove that this hypothesis is true.

Before discussing why correlation does not imply causation, it is worth noting that the reverse direction does hold true: Causation does imply correlation. That is, if two variables are causally related, we should expect (except in unusual circumstances) to find a correlation between them. Thus, correlation is a necessary basis for inferring causality, but it is not sufficient.

This basic limitation in correlational research stems from the absence of experimental control. As has been stressed repeatedly, it is control—control over the nature of the independent variable, control over the assignment of participants to conditions, control over other potentially important variables—that makes internally valid conclusions about cause and effect possible. Because correlational research lacks all these forms of control, the best that such studies can do is to demonstrate that two or more measures covary. They cannot tell us why.

Consider the McLeod et al. study. There are in this study, as in most correlational studies, three possible explanations for the correlation. One possibility is that watching violent TV causes children to be more aggressive. Had McLeod et al. experimentally manipulated TV viewing, they might have established this conclusion with some confidence. But because there was no experimental manipulation, there is a second possibility: Perhaps children who are already aggressive seek out violent TV. In this case it is the aggressive tendency that causes the TV viewing, not the reverse. Finally, there is still a third possibility: Perhaps TV viewing and aggression are both caused by some third factor but are not themselves causally related. It may be, for example, that certain parents' child-rearing practices promote both aggressive behavior and a liking for violent TV; the two measures thus covary, but neither one has any causal effect on the other.

This argument can be put in more general terms. Whenever there is a correlation between variable A and variable B, three possible explanations must be considered: A causes B, B causes A, or some third factor C causes both A and B.

The inability to establish causal relations is obviously a critical limitation to correlational designs. Why, then, are such designs used? The basic reason is that such designs are often the best that we can do. Many variables cannot be experimentally manipulated for ethical or practical reasons—parental child-rearing practices, for example, or exposure to drugs during the prenatal period. In such cases the only approach possible is a correlational one. In other cases experimental manipulation is possible but difficult, especially if the goal is to combine experimental control with a natural setting. The topic of TV violence and aggression provides an example of this point. It *is* possible to manipulate TV viewing experimentally and to measure subsequent aggression, and a large number of studies have done so; to varying degrees, however, all such studies are subject to criticisms of artificiality and lack of external validity. In a study like that of McLeod et al., however, the focus is squarely on the two variables of interest: naturally occurring TV viewing and naturally occurring aggression. A final virtue is that the correlational approach may allow us to sample a wider range of variation than is possible with an experimental design. In an experimental study of TV and aggression, we would probably have to limit ourselves to presenting two or three different types of TV experience. With a correlational approach, however, we can encompass the whole range of naturally occurring experiences, from 2 or 3 hours per week viewing on the one end to perhaps 40 or 50 hours on the other.

Ways to Strengthen Causal Inferences

Causality cannot be established with certainty from a correlational design. There are techniques, however, for heightening the plausibility of any causal inferences that might be drawn. In this section we will consider several such techniques.

A first strategy is quite commonsensical but still worth noting. In some cases one of the A-B causal directions is ruled out by the nature of the variables. Suppose that we find a positive correlation between body size and level of aggression. It is plausible that body size in some way affects aggression (although we would still need to specify exactly how). It is not plausible, however, that level of aggression has any causal effect on body size. In cases such as this, we need to entertain just two hypotheses: A causes B, or C causes A and B. The B to A link, however, is not a concern.

The logical-argument approach is relevant to the issue of the directionality of a causal relation between A and B. A second method is especially appropriate for eliminating third-factor C explanations. It makes use of a statistical procedure called the **partial correlation technique**. Partial correlation is a procedure for statistically removing, or "partialing out," the contribution of one variable from a correlation between two other variables. Essentially, what the partial-correlation technique does is to hold the potentially troublesome third variable constant while examining the relation between the two variables of interest. It is equivalent to asking how A and B relate in a sample in which everyone has the same score on variable C. The issue, of course, is whether the A-B correlation remains significant even when we control for C.

Suppose that we find a positive correlation between TV viewing and aggression but suspect that some third factor, such as methods of child rearing, actually produces the correlation. Assuming that we can obtain acceptable measures of child rearing, we could then use the partial-correlation technique to eliminate the contribution of child rearing from the TV-aggression correlation. If we find that the correlation remains as large or about as large as it was originally, we could conclude that child rearing was not an important confounding factor. Conversely, a substantial drop in the size of the correlation would indicate that child rearing does make an important contribution to the TV-aggression correlation.

Although the specific procedures differ, the goal behind the partial-correlation technique is the same as that for the matching technique discussed earlier in the chapter. In both cases the researcher seeks to eliminate confounding factors by equalizing them across the groups being compared. With matching, the equalization comes before the fact, in the assignment of participants to groups; with partial correlation, it comes after the fact, in the statistical removal of the confounding factors. Partial correlation also shares the same basic limitation that we saw in the case of matching: It is impossible through such techniques ever to remove *all* possible confounding factors. There are, in other words, lots of variable Cs, and no researcher is ever able to measure and control for them all.

A third approach to extracting causality from correlational data is concerned with the directionality issue—that is, does A cause B or does B cause A? The starting point is a basic fact about causality: Causes must come before their effects. Thus, if we can chart variations in the relation between A and B over time, we can come closer to determining whether it is A that leads to B or the reverse.

Charting across-time relations requires longitudinal study. Imagine that we conduct a study with 5-year-olds in which we measure both how much violent TV each child watches and how aggressive each child is. Imagine also that we study the same children 3 years later at age 8, again measuring both TV viewing and aggression. Clearly, we would then have two

standard correlational studies, one with 5-year-olds and one with 8-year-olds. But we would also have an across-time correlational study, in that we could examine the correlations between the measures at age 5 and the measures at age 8. Suppose (to posit a simple case) we find that TV viewing at age 5 correlates with aggression at age 8 but that aggression at age 5 does not correlate with TV viewing at 8. In other words, early variations in TV experience are predictive of later variations in aggression; early variations in aggression, however, do not predict later TV viewing. Such an outcome would be compatible with the hypothesis that TV is the causal agent in the TV-aggression relation. It would not prove the hypothesis, even in the relatively simple case in which the results come out as clearly as in our example. But it would add support to the argument.

Discussion of a final method of strengthening causal inferences will serve also to introduce a basic point about research methods. Sometimes it is possible to complement the correlational approach with an experimental examination of the same issue. What we can do, in other words, is manipulate the variable that we believe to be causal and measure the effects on the other variable, thus creating a true independent variable–dependent variable relationship. As I noted earlier, the literature on TV violence includes a number of such experimental studies, in which TV viewing has been manipulated and subsequent aggression measured. Such studies provide exactly the forms of control that are lacking in correlational research. Because we have experimentally manipulated variable A, there can be no uncertainty about the causal direction between A and B—variations in B must follow from variations in A rather than the reverse. And because we can control factors other than our independent variable, there can be no third factor C to confound any A-B relation. We can be much more certain, therefore, about any cause-and-effect conclusions that we might draw.

The general point that this example illustrates concerns the value of converging operations when investigating complex, hard-to-study topics. The term **converging operations** (also labeled the *multimethod approach*—e.g., Brewer & Hunter, 1989) refers to the use, either within or across studies, of a variety of distinct methods of studying a particular topic (the converse—the exclusive use of one method—results in the threat to validity that Cook and Campbell, 1979, labeled *mono-method bias*). The basic idea is that the strengths of one method can, to at least some extent, compensate for the weaknesses of another method, and that conclusions based upon a convergence of evidence from different methods can be held with a greater certainty than can conclusions based on one approach alone.

This argument certainly applies to the issue of TV violence and aggression. Experimental approaches to this issue are uniquely suited for the identification of causality; at the same time, such studies may suffer from a variety of problems (artificiality, reactivity, etc.) that make their external validity doubtful. Correlational designs avoid many of the pitfalls of manipulative studies; as we have seen, however, the correlational approach is intrinsically limited in what it can tell us about cause and effect. It is precisely because of these limitations of any one method that we need a convergence of evidence from different methods. Thus, the correlational studies of TV viewing give us more confidence that experimental demonstrations of the impact of TV violence really do have some real-life generalizability. Correspondingly, the fact that experimental manipulations of TV violence affect aggression gives us a basis for arguing that TV viewing really is the causal factor in the TV-aggression correlation.[2]

Summary

This chapter addresses three issues that fall under the heading of experimental design:

comparison of different age groups, comparison of different experimental conditions, and contrasts between experimental and correlational designs.

Two designs have been most common in examinations of different age groups: the longitudinal and the cross-sectional. In a *longitudinal study* the same participants are studied across some span of time. Such an approach provides the only direct measure of age changes as opposed to age differences; it also provides the only way to study individual stability or individual change over time. On the negative side, longitudinal research is costly and time-consuming, factors that undoubtedly contribute to its relative infrequency. Longitudinal research is also subject to a number of biases. These biases include *selective drop-out* of participants, *testing* effects stemming from repeated exposure to the same measures, and the inevitable confounding between the age of the participant and the time of measurement.

In a *cross-sectional study* different participants are studied at different ages. The cross-sectional approach is generally more economical than the longitudinal approach, it avoids many of the problems of longitudinal study, and for many research questions it is perfectly adequate. Cross-sectional designs also have their limitations, however. Because each participant is studied just once, a cross-sectional study cannot provide direct evidence of changes with age. *Selection bias* in the formation of the different age groups may hamper the age comparisons. A further problem, which applies to both cross-sectional and longitudinal designs, is that of *measurement equivalence:* selecting measuring instruments that are equally appropriate for the age groups being compared. Finally, cross-sectional designs also contain an inevitable confounding: between the age of the participants and the generation or cohort to which they belong.

Limitations of the classic longitudinal and cross-sectional approaches have led in recent years to the development of alternative designs. In a *time-lag design,* age is held constant while generation and time of measurement are varied. Such designs provide an estimate of the importance of factors that are confounded with age in the traditional designs. More ambitious are the various *sequential designs,* which involve combinations of the simpler longitudinal, cross-sectional, and time-lag approaches. Sequential designs are unquestionably more informative than the simpler approaches; they are also more costly, however, and they still do not remove all possible sources of confounding.

The second section of the chapter is devoted to designs for comparing different tasks or experimental conditions. Two main approaches exist: *within-subject designs,* in which every participant responds to every task or condition, and *between-subject designs,* in which different participants are assigned to the different tasks or conditions. The within-subject approach is sometimes more economical, often affords greater statistical power, and is free of some of the problems (such as selection bias) that can affect between-subject designs. A within-subject approach is also essential when the interest is in within-subject patterning of performance. A between-subject approach, in turn, avoids many of the problems of within-subject testing—in particular, *order* or *carryover effects* stemming from the repeated testing. A between-subject approach is also essential when the experimental manipulation is intended to produce definite and lasting change.

The discussion turns next to specific variants of the between-subject and within-subject approaches. With a *matched-groups design,* participants are matched prior to assignment to experimental conditions. The advantage of matching is that it ensures that groups are equivalent on variables that might affect performance. Possible disadvantages include the increased time and effort, the potentially biasing effects of taking a matching pretest, the increased subject attrition if any matched-for participant is lost from the study, and the

possibility that the groups will be systematically unmatched on variables other than the matched-for variable. In a *time-series design,* an experimental treatment is repeatedly administered and withdrawn, and changes in behavior are charted as a function of the treatment's presence or absence. Such designs are used most frequently in clinical or educational settings, often in the form of single-subject research.

The chapter concludes with a discussion of *correlational research.* In a correlational study there is no control of an independent variable; rather, two or more variables are measured, and the interest is in whether scores on the different measures covary. Correlational designs may be the only research option available for variables whose experimental manipulation is either impossible or very difficult. Furthermore, correlational research has the advantage of encompassing more levels of a variable than is possible in a controlled experimental study. On the negative side, the absence of experimental control means that correlational designs are intrinsically limited in what they can tell us about cause and effect. Methods useful for reducing the uncertainty and moving closer to the determination of causality include logical analysis of which causal directions are possible; *partial correlation,* in which the contribution of third-factor variables is statistically removed; longitudinal study to trace the pattern of correlations over time; and experimental manipulation of one of the variables.

Exercises

1. One theme of the chapter concerns the difficulty of distinguishing age effects from cohort or generational effects. Consider your own cohort. Are there experiences your generation has had that are at least somewhat different from those of other generations? What sorts of effects might these generational differences have in cross-sectional comparisons?

2. One way to think through the complexities of sequential designs is to imagine specific outcomes and what they would mean. Consider the cohort-sequential design schematized in the table below. The dependent variable is IQ, and we will assume that the means for the various groups range from 90 to 110. For each of the following outcomes, generate means that would be consistent with the specified result: (a) an effect of age alone, (b) an effect of cohort alone, (c) effects of both age and time of measurement.

		Time		
		1990	*2000*	*2010*
		Age M	Age M	Age M
Cohort	1930	60	70	80
	1940	50	60	70
	1950	40	50	60

3. This chapter stresses both the value and the difficulty of longitudinal research. One alternative approach to the study of across-time stability or change is the *retrospective method.* The retrospective method goes backward in time, typically beginning with some