CHAPTER W1

Overview of the Logic and Language of Psychology Research

Chapter Outline

This chapter provides an overview of the logic and language of the research process in psychology. Statistical methods are tools used in the research process. You will find the statistical procedures in *Statistics for Psychology* easier to understand if you appreciate the context in which they are used.

Usually, the purpose of a research study in psychology is to test a theory or the effectiveness of a practical application.[1] The strongest research procedures lead to unambiguous conclusions that apply to a wide variety of other situations and people. Weak research designs, even if their results are consistent with your predictions, leave open many alternative interpretations as to why those results were found or apply only to a narrow group of people or situations. (Sometimes, circumstances limit the sort of research procedure that is possible, yet the research may still seem worth pursuing, even if in a less than rigorous way. In fact, especially in the case of applied research, much of the most important work has been done by psychologists using, of necessity, less than perfect methods, but in very creative ways.)

[1] Research is sometimes done for other purposes, such as to explore relationships among measures, to determine the incidence of some characteristic in the population, or to develop a test or technique for use in other research. However, the basic logic of the usual kind of research (the focus of this min-chapter) shapes psychologists' approaches to almost all systematic research.

Most psychologists think about the logic of research in terms of a kind of ideal approach. A real-life study is evaluated in terms of the ways it does and does not come close to this ideal. In this chapter, we first discuss this ideal, the "true experiment." We then turn to four key areas in which studies do or do not come close to it: equivalence of participants across experimental groups, equivalence of circumstances across experimental groups, generalizability, and adequacy of measurement.

THE TRADITIONALLY IDEAL RESEARCH APPROACH

THE TRUE EXPERIMENT

The **true experiment** is the standard against which all other methods are compared. Consider the hypothesis "Changing the level of *X* causes a change in the score on *Y*." A true experiment systematically varies the level of *X*, keeping everything else the same, and looks at the effect on *Y*. For example, suppose you want to know whether flashing lights in a room affects people's scores on a math test. *X* is whether there are flashing lights in the room, and *Y* is math test scores. In a true experiment, each participant in a group of students might be tested in a room with the flashing lights. Participants in another, initially identical group of students, would each be tested under conditions that are completely identical, except there are no flashing lights in the room. Thus, the only difference between the two groups is the level of *X*, the presence or absence of flashing lights in the room. If the students in the room with flashing lights have lower scores on the math test (*Y*), it must be due to the lighting. (If they have higher scores, then *that* effect would also have to be due to the flashing lights.)

BASIC TERMINOLOGY OF THE EXPERIMENT

A group in which the level of *X* is changed is usually called the **experimental group**. The comparison group in which *X* is kept at normal levels is called the **control group**. The individuals studied in the research are called **participants**.[2] The variable that is systematically changed (*X*–for example, whether the lights flash or not) is called the **independent variable** (see also Chapter 17). The procedure of systematically changing the independent variable is sometimes called an **experimental manipulation** or *manipulating the independent variable*. The variable that is supposed to change as a result of the study (*Y*, if *X* causes *Y*–for example, score on the math test) is called the **dependent variable** (see also Chapter 17). Participants are taken from the **population**—all the people on earth of the type being studied. The particular participants selected to be studied from that population are called the **sample**.

Imagine you have two identical cans of a soft drink and you want to test the hypothesis that heating a can of soft drink will make it explode. (Don't try this at home!) To study this, you could put a match under one can (the experimental can) and not put a match under the other (the control can). If the experimental can explodes and the control can does not, the hypothesis is confirmed. Each can is a par-

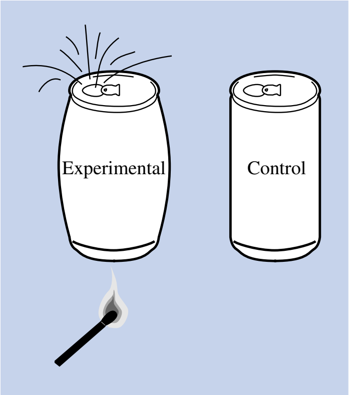

FIGURE W1-1 *An ideal experiment: One of two identical soft drink cans is heated, and the researcher observes to see if it will explode while the other does not explode.*

[2] Traditionally psychologists have used the term *subject*. However, in keeping with the recommendation of the *Publication Manual of the American Psychological Association*, we use the word *participant* here and throughout this book.

ticipant, heating or not is the independent variable, whether a can explodes is the dependent variable, and these two cans are samples, respectively, of the populations of all soft drink cans that are and are not heated (see Figure W1-1).

FOUR CHARACTERISTICS OF THE IDEAL RESEARCH DESIGN

There are four key characteristics of an ideal research design:

1. The participants in the experimental and control groups are identical.
2. The experimental and control groups are exposed to identical situations (*except* for the manipulation of the independent variable).
3. The sample studied perfectly represents the intended population.
4. The measurement of the dependent variable is completely accurate and appropriate for what it is supposed to be measuring.

The rest of this chapter examines the various ways that real-life research tries to come close to each of these ideal conditions.

EQUIVALENCE OF PARTICIPANTS IN EXPERIMENTAL AND CONTROL GROUPS

Ordinarily, the main issue in deciding how well a study can give unambiguous conclusions is the equivalence of participants in the experimental and control groups. Suppose the basic math ability of the participants sent to the room with the flashing lights was different from those sent to the room without flashing lights. Whatever difference in math scores you found between the two groups at the end of the study, it would be hard to know what the difference means. The difference could be due either (a) to the manipulation of the independent variable (having the flashing lights or not) or (b) to their initial differences in ability. To avoid such ambiguous results, researchers aim for strict equivalence of the experimental and control groups. Five main strategies are employed: random assignment to groups, matched-group designs, repeated-measures designs, correlational research designs, and single-subject research.

RANDOM ASSIGNMENT TO GROUPS

The research procedure that comes closest to creating two identical groups of participants is called **random assignment to groups**. For example, if 100 people were available to be in the experiment, each person could be put in either the experimental group or the control group by flipping a coin. The two groups of 50 created in this way are not identical, but at least there will be no *systematic* difference between them.

It is important to emphasize that "random" means using a strictly random procedure, not just haphazardly picking people to go into the two groups. Any haphazard procedure is likely to create unintended systematic differences. For example, suppose you choose one group from students attending a morning class and the others from students attending an evening class. The two groups might differ because the kinds of people taking classes at these different times of day might differ. Or suppose one group are volunteers willing to participate in a self-esteem enhancement program and the control group is simply whoever is willing to take a self-esteem test. The kinds of people in the experimental and control groups might be quite different.

Random assignment rules out initial systematic differences between groups. Any actual difference that exists after random assignment will be entirely due to random processes. Thus, if there are differences on the dependent variable after the experiment, they only can be due to either the manipulation of the independent variable or to the random assignment. True random processes follow the laws of probability. Thus, the hypothesis-testing procedures that are covered starting in Chapter 6 are able to check the probability of whether the difference found in a study could have been due to the random assignment. If the statistical analysis indicates that this is unlikely, the only reasonable remaining explanation is that the manipulation of the independent variable caused the difference. This is the basic logic behind the analysis of results of experiments. And this is why random assignment and statistical methods are so important in psychology research.

MATCHED-GROUP DESIGNS

Sometimes random assignment to groups is not practical. For example, ethics would require that all students in a school district who need a certain reading program receive it; some cannot be randomly chosen to miss out. How can the program be shown to be the cause of improvements in the students? One widely used alternative research approach is the **matched-group research design**. For example, you might compare an experimental group of students who have been selected for the program in one school district to a control group of students in another district who also need the program but for whom it is not available. Every member of this control group could be matched to a member of the experimental group in terms of age, social class, sex, reading problem, and so forth.

Matched-group designs are much better than having no control group at all. In fact, if both groups are tested before and after, the matched-group design can lead to fairly unambiguous results. This situation, called a *matched-group pretest-posttest design*, is an example of a *quasi-experimental design*. A quasi-experimental design is any approach that reasonably approximates a true experiment but does not use random assignment.

However, no matter how well matched two groups are, and even when before-and-after testing is used, a researcher can never know for certain that there is no systematic initial difference between the groups. Indeed, in most cases, if you have not used random assignment, you know that there *is* a systematic initial difference-whatever it was that put people into one group or the other. (In the reading program example, the systematic difference might be that one group of students lives in a school district not progressive or well funded enough to offer the reading program.)

REPEATED-MEASURES DESIGNS

Another research approach is to create two identical groups by testing the same people twice. This is called a **repeated-measures research design** (it is also called a *within-subjects research design*). The students in our example could be tested before and after the reading program.

The simplest repeated-measures design is a *single-group pretest-posttest design* in which, as the name implies, a single group of individuals is tested twice, once before and once after some experimental treatment. This kind of research design, however, is very weak in the sense that if you found a change, there are many possible alternative explanations for it. Merely being tested the first time can change a participant, so that when tested again, the person is different—different due to the initial testing, not to the experimental treatment. And time itself produces change.

More generally, in this kind of study, any change could be due to the reading program or to whatever else happened to the participants (besides the experimental treatment) during that time period. Or there could be preexisting trends for improvement, or the change could be due to general maturation and experience, or to people starting at a very low point that would naturally improve without the treatment, and so forth. (See also Chapter 9.)

Because it is such a weak research design, the single-group pretest-posttest design is considered a *preexperimental design*. Research of this kind is often extremely important as a first stage in exploring a research area, but any conclusions from a study of this type are very tentative and should be followed up by a stronger research design (such as a quasi-experimental or true experimental design).

In a laboratory setting, however, a repeated-measures design is often used in a way that makes it a true experiment. Consider again our interest in the effect of flashing lights in the room on performance on a math test. The researcher might test the same participant's performance with flashing lights (the experimental condition) and then test their performance without flashing lights (the control condition). A problem with this approach, however, is that the participants could be more familiar with the test the second time, creating a *practice effect* or *carry-over effect*, or they could be tired out by the time they get to the second task, creating a *fatigue effect*.

To deal with problems of this kind, researchers use a procedure called **counterbalancing**, in which half the participants are tested first in one condition and the other half first in the other condition. In this way, any practice, carry-over, fatigue, or similar effects are balanced out over the two conditions. Ideally, you use counterbalancing so that the condition a participant experiences first is determined by random assignment. In this situation, the study becomes a true experiment. Indeed, repeated-measures designs with counterbalancing and random assignment are among the most powerful research methods psychologists use because they make groups so very equivalent. (See Chapter 9 and Mini-Web Chapter W2.)

Correlational Research Designs

A **correlational research design** tests whether there is an association between two variables as they exist in a group of people, without any attempt at experimental manipulation. Thus, a correlational approach to studying self-esteem and job satisfaction might survey a group of workers on their self-esteem and their job satisfaction. Then you would see whether those scoring high on self-esteem tended also to score high on job satisfaction. (You figure the degree to which there is an association using a statistical technique called a "correlation coefficient," described in Chapter 3.)

The correlational approach is often the best that can be done under the circumstances and is widely used. But it is a fairly weak research design in that its results are open to many alternative explanations besides "*X* caused *Y*." For example, suppose you find that self-esteem and job satisfaction are associated in a correlational study. This could be due to high self-esteem causing high job satisfaction. But it could also be that high job satisfaction causes high self-esteem. The association between self-esteem and job satisfaction could even be due to other differences among the workers, such as age—perhaps being older causes workers to have both high self-esteem and high job satisfaction. (The various possible causal interpretations of the results of a correlational study are discussed in some detail in Chapter 3.) Thus, one advantage of a true experiment (when it is feasible) over a correlational study is that the true experiment manipulates the independent variable and then sees the effect on the dependent variable, making it quite clear what causes what.

Researchers are well aware of the limits of correlational designs. When possible, they try to rule out some alternative explanations, mainly by using sophisticated statistical procedures such as partial correlation (summarized in Chapter 17). Still, the correlational approach never produces results as clear as a true experiment or, in most cases, even as clear-cut as a quasi-experiment. However, often, the correlational approach is the strongest method that is practical—for example, you cannot randomly assign people to marry certain other types of people. Even when experiments are possible, they may be very costly to conduct. Thus, researchers may not be willing or able to test an untried idea experimentally. In these situations, correlational studies often provide a valuable first step in opening up a new area of research.

Single-Subject Research

Finally, some research studies involve an intense examination of a single group, organization, or individual, using the *case study* or *participant observation* approach. Such **single-subject research** is not considered experimental or even correlational. However, in clinical psychology and in some other social science fields, such as sociology and anthropology (and the sociological and anthropological approaches to organizational behavior, education, criminology, communication, and so forth), this kind of research is valuable because it gives a rich understanding of all the complexities of what is being studied rather than forcing attention on a few variables that may or may not be the most critical. In all areas of psychology, as well as the other social sciences, single-subject research is also considered useful as a precursor to other, more rigorous research approaches. (These topics are discussed in Chapter 2.)

Single-subject research is also used in a highly systematic way by researchers in the behaviorist tradition developed by B. F. Skinner. A single participant—whether an animal, like a rat or a pigeon, or a client in a behavior therapy program—is studied over time, with the researcher systematically manipulating the conditions that affect the participant and observing the changes that result. Statistics are not usually used; the pattern of results should be so clear that you don't need statistics.

Summary of Research Designs

Table W1-1 summarizes the various research designs we have considered, noting their advantages and disadvantages as compared to the ideal of identical experimental and control groups.

EQUIVALENCE OF CIRCUMSTANCES FOR EXPERIMENTAL AND CONTROL GROUPS

The ideal study involves not only identical groups but also testing them under identical circumstances.

In practice, it is quite difficult to test two groups under circumstances where the sole difference is the manipulation of the independent variable. In a physics laboratory such equivalence may be possible. But when conducting research with human beings, circumstances are never equivalent. One strategy designed to maximize equivalence is to use an isolated location, such as a cubicle in a psychology building.

TABLE W1-1 Major Research Designs and Their Advantages and Disadvantages

Design	Advantages	Disadvantages
True experiment (random assignment to conditions)	Ensures no systematic difference between conditions.	Can be impractical or unethical
Matched-group (without random assignment)	Controls for obvious differences between conditions; may be most practical with intact groups.	Groups may differ systematically on variables on which they were not matched.
Matched-group pretest-posttest	Controls fairly strongly for initial differences among participants; is often practical where random assignment is not.	Systematic differences between groups may influence impact; pretest measuring procedure can confound results.
Repeated-measures true experiment (random assignment)	Ensures no systematic difference; minimizes random differences by making participants their own controls.	Practice or carry-over effects; procedure may be difficult to implement.
Single-group pretest-posttest	Provides some control; is often the only practical approach.	Impossible to know if change would have occurred without the experimental treatment.
Correlational	Is relatively easy to do with existing groups.	Difficult to determine direction of causality.
Single-subject	Permits deep understanding of processes.	Difficult to generalize results.

This minimizes external influences and interruptions that might make one session of the experiment different from another. A related approach is to standardize the situation as much as possible. For example, the instructions to participants may be tape recorded.

There are, however, two special problems that plague much psychology research-particularly applied research—with regard to equivalence of circumstances: placebo or Hawthorne effects and experimenter effects.

Placebo and Hawthorne Effects

Placebo effects are the influence of a participant's expectation or motivation to do well. **Hawthorne effects** are the influence of the attention the participant receives and of the participant's reaction to being a participant. For example, if one wing of a factory is trained in a new program and one wing is not, there are several differences between the situations the two groups are in. One wing uses the new way of operating resulting from the program, and the other wing does not—this is the manipulation of the independent variable. But another difference is that those in the wing getting the new program know they are getting a new program and may thus expect to get benefits (creating a placebo effect). Yet another is that one wing has received special attention and the other wing has not (creating a Hawthorne effect—

the term comes from a 1927 study done at the Hawthorne Works plant of the Western Electric Company in Cicero, Illinois). These additional differences between groups greatly complicate the interpretation of the results.

How can researchers deal with these undesired differences in circumstances? The best solution is to conduct a study in which both groups receive some treatment that they believe should be helpful, however only one group actually receives a treatment consisting of more than mere attention and raised expectations. For example, in medical research, both groups would receive pills that look and taste identical, but one group's pills contain the active ingredient and the other group's do not. No one in the experiment knows who is receiving the real drug. A drug that looks and tastes like the real thing but is actually inactive is called a *placebo* (Latin for "I shall please").

In psychology, it is often impossible or unethical to set up a control group condition where a person receives a treatment that is believed to be effective but in fact is not. A situation where it is feasible to use a true placebo control group and also the research personnel are unaware of which participants are in which group is called a *double-blind procedure*.

Placebo and Hawthorne effects are the most common problems in drawing unambiguous conclusions from results of applied research in areas such as clinical, educational, and organizational psychology.

Experimenter Effects

Experimenter effects, including *experimenter bias*, are the unintended influences of the researcher on the study. For example, in a study of the effects of psychotherapy, suppose that the researcher is a therapist evaluating the mental health of the participants. In this case, it is quite possible that the therapist's desire to see the experiment work creates a predisposition to see participants in the experimental group as having improved more. Even if an independent observer rates the two groups but knows who is in which group, a desire for the experiment to come out a particular way may unintentionally influence the observer's evaluations.

The preferred solution to this problem is called *blind conditions of testing*. This means that the experimenter, at the time of interacting with the participant, is not aware of whether the participant is in the experimental or control group. (We already considered above what is called double-blind testing, where neither researcher nor participant knows what condition they are in. There we were emphasizing the importance of the participant not being aware of who was in what condition; here we are emphasizing the importance of the experimenter not knowing who is in which condition at the time of testing.)

REPRESENTATIVENESS OF THE SAMPLE

The third requirement for an ideal study is that the sample of participants studied accurately represent the population to which the study is supposed to apply. This representativeness is called **generalizability** or *external validity*. (*Internal validity* refers to the equivalence of the experimental and control groups and equivalence of circumstances.)

Participants in psychology research are often college students, and it is assumed that what is discovered about them applies to the larger population of people in general. In a study of the effect of flashing lights on performance, the general pattern of results with college students probably applies to most other human beings. However, in many other types of research, who the participants are is very important. For example, college students would probably not be suitable participants for studies of attitudes toward children—their experience does not commonly include parenthood. You cannot study reading skills in suburban schools and generalize to all students in all schools or study job satisfaction in the computer industry and generalize to all industries.

Another problem involves how a study's participants are recruited. For example, in a mail survey of knowledge about an issue, some individuals will return the questionnaire and some will not. Presumably there are systematic differences between those who do and do not—it is likely that those who do may know more about the issue being studied. Using only the questionnaires that are returned, the researcher may conclude that people are more knowledgeable about an issue than if the researcher had been able to study the entire population. Similarly, people who volunteer to participate in an experiment may differ from those who do not. For example, volunteers may have personalities that are more responsive to the needs of others.

Random sampling is considered the optimal method for ensuring that a sample is representative of its population. Random sampling means that researchers begin with a list of everyone in the population about which they want to generalize their results, such as a list of all psychotherapists in the nation, then use a random procedure (such as a random number table) to select a sample from this population. This produces what is called a *probability sample* because every member of the population being studied has an equal probability of being included in the study's sample. (See Chapter 5.)

Do not confuse random sampling with random assignment to groups, which we discussed earlier. Both procedures use true random procedures, but random sampling is a method of selecting the sample to study; random assignment to groups is a method of deciding which members of the sample will be in the experimental group and which in the control group.

MEASUREMENT

The fourth condition we noted for an ideal study is that the measures should be accurate and appropriate. There are three main kinds of measures used in psychology research: **self-report measures**, such as questionnaires or interviews; *observational* or **behavioral measures**, such as rating scales of children's play behavior, number of customers who go through a turnstile, number of milliseconds to respond in a reaction time experiment, or number of times a rat presses a bar; and **physiological measures**, such as hormone levels, heart rate, or blood flow in a particular brain area. All three kinds of measurement are evaluated mainly in terms of their reliability and validity.

RELIABILITY

The **reliability** of a measure is its accuracy or consistency. That is, when you apply the same measure to the same thing, under identical circumstances, how similar are the results? In psychology, the results are not necessarily similar at all-for example, the same person taking the same questionnaires on different days may get a quite different score. Sometimes, questions are worded poorly, so that a person may answer in one way at one time and in another way at another time. Or, people may simply mark some or all of their answers in the wrong place on one or more occa-

TABLE W1–2 Types of Reliablity

Test-retest reliability	Correlation of tests administered to the same people on different occasions
Internal consistency	Correlation among the items
Interrater reliability	Correlation among different raters' scores when rating the same group of people or objects

sions. Self-report measures are not the only ones that can be unreliable. Observational measures may be unreliable because observers may disagree. Physiological measures are often highly erratic from moment to moment.

There are three types of indicators of degree of reliability: (a) **test-retest reliability**, in which the same group is tested twice; (b) **internal consistency**, in which, for example, scores on half the questions are compared to scores on the other (*Cronbach's alpha*, the most common approach to internal consistency, is described briefly in Chapter 17); and (c) **interrater reliability**, used for observational measures, which is the degree of agreement between observers. These kinds of reliability are summarized in Table W1-2.

VALIDITY

The **validity** of a measure refers to whether it actually measures what it claims to measure. (The word validity is also applied to entire studies, as in internal validity and external validity, when it refers to the appropriateness and breadth of the conclusions that can be drawn from the results.)

A measure that is not reliable cannot be valid. An unreliable measure does not measure anything. But even if a measure is reliable (accurate and repeatable), it is not necessarily valid for measuring what it is intended to measure. For example, consider a marital satisfaction questionnaire with many items such as "How likely are you to stay with your spouse over the next several years?" The questionnaire may turn out to be highly reliable (for example, people may answer all the questions on it quite consistently). But instead of measuring marital satisfaction, it might really be measuring commitment to the marriage. And respondents might be committed not because they are satisfied but because they have no alternative to married life or because they feel they are very unattractive and could only do worse if they left their partner.

Another reason that a test may not be valid, even if it is reliable, is that rather than measuring the intended variable, it is actually measuring a tendency for the respondents to try to make a good impression or to say yes or to answer with some other **response bias**. One way to address the problem of trying to make a good impression is to include a *social desirability scale*, sometimes called a *lie scale*. When a participant's score on such a scale is high, the researcher may simply throw out that participant's test. Alternatively, scores on a social desirability scale may be used in a statistical procedure, such as partial correlation or an analysis of covariance (both briefly described in Chapter 17), to adjust the person's score on the regular part of the measure.

Validity of a measure is more difficult to assess than reliability. Several methods are used. **Content validity** results when the content of the measure appears to get at all the different aspects of the things being measured. Usually, this is determined by the judgment of the researcher or other experts.

TABLE W1-3 Types of Validity of a Measure

Content validity	The content of the test appears to experts to encompass the full range of what the test claims to measure.
Criterion-related validity	Scores on the test correlate with some other indicator of what the test is supposed to measure.
Predictive validity	The test score predicts scores on another variable that ought to be predicted by the test, given what it is claimed to measure; a type of criterion-related validity.
Concurrent validity	The test score correlates with another variable measured at the same time that is already known to be related to what the test is claimed to measure; a type of criterion-related validity.

There are, however, more systematic means of evaluating the validity of a measure. Determining **criterion-related validity** involves doing a special study in which the researcher compares scores on the measure to some other likely indicator of the same variable. For example, a researcher might test the validity of a measure of mental health by comparing scores of people in a mental hospital to people from the general population. One type of criterion-related validity is a measure's *predictive validity*—for example, whether scores on a job skills test taken when applying for a job predict effective performance on the job. Predictive validity is used especially where a measure is designed for predictive purposes, such as job or educational placement. Another type of criterion-related validity is *concurrent validity*. This refers to the procedure of comparing scores on one measure to those on another that directly measures the same thing—for example, a new, short intelligence test compared to an existing, longer intelligence test. All three ways of assessing validity are summarized in Table W1-3.

You may also see the term *construct validity*, which is used in a variety of ways. Even textbooks on psychological measurement disagree about it. Sometimes, it includes criterion-related validity and sometimes content validity. Often, it refers to the measure's being used in a study in which there was a predicted result borne out by the study. Because the measure used was successful in producing the predicted result, it shows that the idea (or "construct") behind that measure proved itself under the theory.

KEY TERMS

behavioral measures
content validity
control group
correlational research design
counterbalancing
criterion-related validity
dependent variable
experimental group
experimental manipulation
experimenter effects

generalizability
Hawthorne effects
independent variable
interrater reliability
internal consistency
matched-group research design
participants
physiological measures
placebo effects
population
random assignment to groups
random sampling
reliability
repeated-measures research design
response bias
sample
self-report measures
single-subject research
test-retest reliability
true experiment
validity