Chapter 2

Doing Social Science Research

We said in the first chapter that social scientists conduct empirical research, or systematic observation, to generate, support, and modify theories and hypotheses about social behavior. In this chapter, we are more precise about the nature of scientific theories and hypotheses. The roots of this chapter lie in positivist/post-positivist perspectives. Constructivists focus less on generalizable theories and scientific methods shared with natural sciences. We also discuss the ways in which a particular piece

Research Methods in Social Relations, Eighth Edition. Geoffrey Maruyama and Carey S. Ryan.
© 2014 John Wiley & Sons, Inc. Published 2014 by John Wiley & Sons, Inc.
Companion Website: www.wiley.com/go/maruyama

of empirical research may be valid or invalid. In other words, we introduce a set of criteria that are used to evaluate the type of scientific research that focuses on theory and hypothesis testing.

The Nature of Social Science Theories and Hypotheses

In Chapter 1, we defined a **theory** as a set of interrelated hypotheses that is used to explain a phenomenon and make predictions about associations among constructs relevant to the phenomenon. Thus, a theory about social behavior has three features:

- It contains constructs of theoretical interest that it attempts to explicate or account for in some way.
- It describes associations among the constructs. These associations are frequently causal, specifying which constructs affect which others and under what conditions. Hypothesized associations are the heart of a theory.
- Finally, a theory incorporates hypothesized links between the theoretical constructs and observable variables that can be used to measure the constructs. These links specify the behaviors or other indicators of the constructs, which are measured and used to conduct empirical research. Nonscientific or naïve theories of social behavior also consist of constructs and causal relations among them. However, because the scientific study of social behavior relies on empirical research to support and modify theories, scientific theories also specify the observable (measurable) indicators that define the constructs of theoretical interest.

Two examples will clarify what constitutes a theory. A theory of political information processing (Lavine, Borgida, & Sullivan, 2000) holds that when people exhibit greater attitude involvement (i.e., they care more about an issue), they are more likely to engage in biased information-gathering strategies, which results in more extreme and unidirectional attitudes. These attitudes, in turn, result in less decision conflict and greater accessibility of the attitudes in memory. Notice in this example the chain of hypothesized causal associations among constructs: Attitude involvement leads to biased information gathering, which leads to extreme attitudes, which leads to greater attitude accessibility. A good theory also specifies how the constructs of interest could be measured, observed, or manipulated. For instance, attitude accessibility might be indicated by faster responses to items on an attitude scale.

Gaertner and Dovidio (2000, 2012) have developed a social categorization-based theory of intergroup bias known as the Common Ingroup Identity Model. According to this theory, members of different groups or social categories (e.g., Blacks and Whites) who view themselves as belonging to the same larger group (e.g., citizens of the United States), that is, as having a common group identity, have more positive attitudes and beliefs about the subgroup to which they do not belong. Encouraging people to think in more inclusive ways, for example, by increasing the salience of their common identity or introducing tasks that require them to cooperate, leads them to think more inclusively (i.e., to view themselves as belonging to the same larger group), which increases positive expectations, perspective-taking, empathy, and trust, and ultimately reduces prejudice and discrimination.

Gaertner and Dovidio (2012), as well as others, have used a variety of indicators of the constructs that are the focus of the theory. For example, in one experiment, Black interviewers asked White students from either the same or a different university to comply with a request (i.e., to be interviewed about their food preferences). The Black interviewers wore clothing that was associated with either the same university that the White students attended (so that they had a common identity) or a different university. In other words, a common ingroup identity was manipulated via the interviewers' clothing. Discrimination was indicated by lower compliance with the interviewers' requests. The results indicated that White student compliance with Black students' requests was much greater when they shared the same (59%) as opposed to a different (36%) identity. Note that there may actually be a great many specific indicators of particular constructs (e.g., common identity and discrimination) although all of them will have basic features in common.

Both of these examples show the basic structure of a theory. A theory is comprised of hypotheses, which in turn are comprised of statements about associations among constructs (e.g., attitude involvement leads to biased information gathering; creating an ingroup identity reduces prejudice) and associations between constructs and observable indicators (e.g., biased information gathering is indicated by selective attention to attitude-consistent information). The observable indicators are known as variables. A **variable** is any attribute that changes values across people or things being studied. Thus, hair color, IQ test scores, height, introversion, gender, and blood pressure are variables. In a given study, however, they would only be considered variables if at least two levels of them were included. For example, gender would not be considered a variable in a study that examined only men or only women.

A theory is thus made up of two types of **hypotheses**: (1) hypothesized associations among constructs and (2) hypothesized associations between constructs and observable indicators or variables or measures. Both types of hypotheses have characteristic forms. The first, concerning associations among constructs, typically takes the form:

> Construct A causes construct B for population X under condition Y.

Each of the examples of theories we discussed earlier contains hypotheses that conform to this model, although the word "causes" might be replaced with "leads to," "produces," or "is associated with." Note, however, that in any given hypothesis, much may remain implicit. For instance, the populations or conditions for which the causal association between construct A and construct B holds might not be explicitly mentioned.

A few further examples illustrate hypotheses about associations among constructs:

- Contact between the members of ethnic groups is more likely to reduce prejudice when the group members have equal status in the contact setting.
- Parents more involved in their children's education have children who perform better academically.

- Media portrayals of women that focus on youth, beauty, and sexuality contribute to gender inequality.
- People more concerned about others are more likely to support environmental protection policies.

The second type of hypothesis concerns associations between constructs and observable indicators. Because constructs are conceptual, they likely are difficult to measure directly and perfectly. Instead, researchers identify measurable yet likely imperfect variables that represent the constructs. The extent to which they are imperfect is called unreliability. For example, body mass index uses only height and weight information, ignoring muscle mass, etc., and is not a perfect measure of obesity or fitness. Associations between constructs and variables/measures usually are of this form:

Behavior X or response Y is a valid indicator of construct A.

Examples include the following:

- Higher scores on the Ambivalent Sexism Inventory indicate more sexist beliefs.
- The SAT or ACT is a valid measure of academic preparation.
- Poverty is indicated by school children's eligibility for free or reduced-price lunches.
- Obesity is indicated by a body mass index that is greater than 30.

What Makes a Theory Productive?

Some social science theories seem to attract a great deal of attention and lead to a great deal of research, whereas others seem to attract very little attention and produce very little research. What characteristics distinguish more and less productive or influential social science theories from each other? Table 2.1 lists the characteristics of a productive theory; our discussion here expands and explains these characteristics.

A good theory must first of all be falsifiable. When we say a hypothesis or theory is **falsifiable**, we mean that we could conceive of a pattern of findings that would contradict the theory. In other words, a falsifiable hypothesis is one for which a researcher can set up an empirical test and, if the findings turned out a given way, the researcher would conclude that the hypothesis had been disproven. The falsifiability criterion of a theory is often difficult for students to understand abstractly, so we borrow an analogy from Meehl's (1978) classic article to illustrate it. We could predict that the high temperature tomorrow will fall between −100° and +200° Fahrenheit. This is not a useful hypothesis because it is not falsifiable; barring cataclysm, the hypothesis will be supported. A hypothesis that allows for every possible imaginable outcome actually explains nothing. In contrast, a hypothesis predicting that tomorrow the high temperature will fall between 70° and 74° Fahrenheit is both eminently falsifiable and – if supported – potentially very useful.

This example makes two points: First, falsifiability is a necessary and minimum requirement for a theory. Second, the more specific a hypothesis is, the more useful

Table 2.1 Characteristics of a Productive Theory

A productive social science theory

- is falsifiable;
- states hypotheses as specifically as possible;
- is as parsimonious as possible;
- addresses an important social phenomenon;
- is internally consistent; i.e., the hypotheses do not contradict one another;
- is coherent and comprehensible;
- specifies its relevant constructs and how they are measured;
- agrees with what is already known about the topic;
- explains data better than existing theories on the same topic;
- agrees with existing theories about related topics;
- generates new insights about the topic.

the theory becomes. The second characteristic of a productive theory, then, is that it states specific hypotheses. Meehl (1978) argued that science progresses to the extent that theories have been subjected to and passed risky (i.e., difficult) tests and that the more such risky tests a theory has survived, the better corroborated it is. In Meehl's words, "a theory that makes precise predictions and correctly picks out *narrow intervals* or *point intervals* out of the range of experimental possibilities is a pretty strong theory" (p. 818; emphasis in original).

The third in our list of the characteristics of a good or productive theory is parsimony. Given equal explanatory power, theories that are simpler or more **parsimonious**, that is, those that specify fewer theoretical concepts and relationships, are preferred to those that are more complex. Parsimony may seem surprising given the complexity of human behavior. In an effort to account for this complexity, theories may become quite complex, sometimes involving substantial elaboration to integrate exceptions. The point is that for reasons of precision and clarity and to avoid the pursuit of research that is unlikely to be fruitful, information or ideas that are not necessary to account for a phenomenon ought to be excluded, only necessary elements should be retained, and simpler explanations should be chosen over more complex ones.

Ultimately, the criterion by which we evaluate theories is whether they provide compelling explanations and interpretations for the world around us. There are actually two components to this criterion. First, a productive or useful theory is one that addresses some important or significant phenomenon or social behavior that needs explication (e.g., gender inequality). One implication of this characteristic is that a theory's importance may change over time; a theory may be especially important or productive at one particular time and less so at other times. Phenomena that seem to demand attention change with time. For example, during the late 1980s environmental hazards were of major public concern; more recently, global warming has become a major concern and has thus stimulated a great deal of research. So theories are used and are productive in part if they address phenomena that are socially significant at a particular historical moment.

The second component of this criterion is that a useful theory provides a plausible and empirically defensible explanation for the phenomenon. A plausible explanation means that the theory must be internally consistent, coherent, and comprehensible. It must be accessible to those who use it, that is, to those who conduct the research in support of it, and it must not run entirely counter to common sense or ordinary explanations. It should be more parsimonious and/or do a better job of explaining research findings than existing theories on the topic. In addition, to stimulate research and to be empirically defensible, the theory must be relatively specific about its constructs and how they are to be measured. In other words, a productive theory includes hypotheses about the links between variables and constructs.

There are other criteria for defining a useful or productive theory. One is that the theory must be consistent with both existing research findings and existing theories for related phenomena. The need to be consistent with known research findings is obvious; there is little utility in proposing a theory that has already been empirically disproven. The need to be consistent with related theories is less obvious but also important. For instance, a theory about human memory must be relatively consistent with existing empirically supported theories about reading comprehension, judgment, and perception. Memory does not exist independently of these other cognitive phenomena and neither can an adequate theory of memory. Similarly, a theory about the origins and effects of poverty cannot afford to ignore theoretical approaches describing relations among ethnic groups, social stigma, and crime. The phenomena are linked, and so, too, must be the theories.

Finally, a productive theory is one that yields new insights or offers the possibility of unforeseen implications. That is, good theories grow and prosper as individual researchers examine their implications and extend them logically. Useful theories offer the possibility of growth, allowing researchers to think about connections that they would not have thought about otherwise. This aspect of theory development is exemplified by the use of computer simulations (Hastie & Stasser, 2000; Mosler, Schwarz, Ammann, & Gutscher, 2001). When a theory identifies a complex set of interrelated phenomena and is specific about the forms of the associations among them, one can often use a computer to examine the dynamic implications of the theory. Computer simulations of social phenomena often provide new and empirically testable hypotheses that derive from the theory's basic postulates and that may not have been seen otherwise (e.g., Anderson, 2007; Newell & Simon, 1961).

Exercise: Examining "productivity" of theories

Select a theory that you recently have read about or used in your research. Put it to the test, examining how well it meets the standards set forth: falsifiable, specific in its predictions, parsimonious, addresses an important phenomenon, plausibly explains phenomenon, consistent with prior research, yielding new insights. How does the theory you picked do? In what ways is it strong and in what ways is it not so strong?

The Functions of Research in Constructing Theories

A primary purpose of conducting empirical research is to test hypotheses. Although hypotheses are typically derived from theories, they may develop for other reasons as well. For example, hypotheses may be developed as a way to resolve conflicting research results; from case studies, systematic observation, or other types of qualitative research; or serendipitously as a result of research findings that were unexpected but seem potentially interesting to examine further. Hypotheses even develop because our personal experiences seem inconsistent with or are not explained by existing theories and hypotheses. For this and other reasons, diversity in the characteristics of scientists is critical to the advancement of science. Indeed, scientists generally have a great deal of latitude in their choices of questions to examine and in the types of methods they develop and employ to do so. And they differ in what answers they expect their science to provide (e.g., general laws vs. local solutions), creating healthy dialogues about the nature of scientific inquiry and discovery. Even within approaches, scientists often disagree in their interpretations of results, generating further research in an attempt to resolve the disagreement.

In any case, empirical research is conducted to examine hypotheses about the associations among constructs. In doing this sort of research, we usually make assumptions about the second sort of hypotheses, that is, those linking the constructs with variables or measures, the observable indicators of the unobservable constructs. For instance, we might conduct research designed to demonstrate that interracial contact decreases prejudice. In the process, we make assumptions about how both constructs, interracial contact and prejudice, are to be measured.

Although research that examines hypotheses of the first sort, causal associations among constructs, is more frequent, research on hypotheses of the second sort is also a primary activity of social scientists. Research designed to examine whether a given variable accurately or validly measures a given construct is called measurement research. **Measurement research**, sometimes referred to as psychometric or sociometric research, usually is conducted by examining whether two or more ways of measuring the same construct yield similar results. As will become apparent in later portions of this chapter and in Chapter 8, such research is vitally important to the success of research examining hypothesized causal associations among constructs. Only if we can successfully manipulate, observe, or measure the constructs of interest can we empirically examine hypotheses about the causal associations among them.

We have said that a primary purpose of conducting empirical research is to examine hypotheses. At this point we need to be more specific about what this means. There are four different functions or purposes of empirical research that, in total, constitute the process of examining social science hypotheses: (1) discovery, (2) demonstration, (3) refutation, and (4) replication.

Discovery

Researchers frequently gather information to attempt to discover what might be responsible for some phenomenon or behavior. For instance, in studying depressed clients, we might interview and observe the clients' families to see whether there are any patterns of interaction that might be responsible for the depression. In doing such systematic observation, we do not as yet have a well-defined hypothesis about the causes of depression. Rather, we are attempting to **discover** what might be plausible causes of constructs. Research as discovery is used primarily to develop or generate hypotheses. When conducting research for this purpose, the researcher is operating in an **inductive** manner, attempting to move from observation to the development of hypotheses, rather than the other way around, that is, from hypotheses to observation, which is known as **deductive** research.

Even in inductive research, research is rarely solely about discovery. Insofar as researchers have been thinking about and looking at prior research about the issue of interest, there is some ill-defined or implicit theoretical orientation that guides the research, even when the researchers have no explicit hypotheses they are examining. For instance, in the depression example, a researcher who interviews family members implicitly assumes that understanding the causes of depression might lie in the family and their interactions with the client. A researcher who believes that depression is a result of a genetic or neurochemical malfunction would not look for causes in patterns of family interaction and would likely not interview families. In other words, without some kind of underlying or implicit theory, researchers would not know where to begin looking for the causes of a phenomenon or behavior. Thus, the difference between inductive and deductive approaches is perhaps best thought of as a difference in degree. Even constructivists are not likely to conduct research as pure discovery or to proceed purely inductively, for choice of problem or setting involves presumptions about the problem and how to go about understanding it. Even when research is used primarily to generate hypotheses, researchers inevitably make theoretical assumptions in deciding what to observe or where a potential cause might lie.

Demonstration

If researchers have a hypothesis about the associations among constructs of interest, they are quite likely to gather data in an attempt to **demonstrate** or support it. Suppose, for instance, that researchers believe that living in an integrated neighborhood reduces prejudice. They might then try to generate information or make observations to demonstrate the validity of this hypothesis. For instance, they might interview residents of both integrated and segregated neighborhoods about their attitudes toward various ethnic groups. If the interviews showed that those who lived in integrated neighborhoods had more favorable attitudes, the findings from the research would be consistent with the hypothesis. Such consistency of observation with the hypothesis is the limit of what demonstration research can accomplish.

Research findings can only be consistent with or demonstrate a hypothesis. They can never prove the hypothesis. This point was made in the first chapter but bears repeating here. That residents of integrated neighborhoods express less hostility toward other ethnic groups than do residents of segregated neighborhoods does not mean that the hypothesis, which states that integration *causes* a reduction in prejudice, is correct. There are always alternative explanations that may be equally consistent with the research results. For instance, residents of integrated neighborhoods might express less hostility because they were initially less prejudiced before they moved into the neighborhood, and highly prejudiced people may choose not to live in integrated neighborhoods. Hence, although the research findings are consistent with the hypothesis or demonstrate that it might be correct, alternative explanations that may be equally consistent with the research results always remain.

Research designed to demonstrate a hypothesis is deductive rather than inductive. In other words, in demonstration research, the hypothesis generates the research, whereas in discovery, research is used to generate hypotheses. Scientists, when acting deductively, start with a hypothesis, which they then seek to support or demonstrate using information generated by empirical research.

As was true for inductive research, research is never pure deduction or pure demonstration. Although it could turn out that the research results are nearly perfectly consistent with the hypothesis, inevitably some inconsistencies or results emerge that cannot be entirely explained by the hypothesis. The researcher then proceeds inductively, examining the findings and hypothesis to determine how the hypothesis might be modified to account more precisely for the research findings. In this way, research never exclusively serves a discovery or a demonstration function, just as the researcher never reasons exclusively deductively or inductively.

Refutation

Although a hypothesis can never be proven to be true, it is possible to **refute** competing hypotheses. For instance, suppose we conduct research on the "integration reduces prejudice" hypothesis that we have been discussing. Suppose we find that residents of integrated neighborhoods express less hostility than do residents of segregated neighborhoods. We might then want to refute the competing or alternative hypothesis that residents of the two neighborhoods differed in prejudice initially, before they moved into the segregated or integrated neighborhoods. To do so, we would have to conduct further research, interviewing people when they first move into integrated neighborhoods and then following them over time. If we found that initially they expressed hostility equal to that of segregated residents but that over time they developed more positive attitudes, we would have generated evidence to refute the competing hypothesis.

The process of supporting a hypothesis, and ultimately a theory that is made up of numerous hypotheses, is one of demonstration and repeated refutation of alternative hypotheses. Although in a formal sense there are always alternatives that have yet to be refuted, the remaining alternatives become more and more far-fetched, and gradually we develop confidence in a hypothesis through repeated demonstration and repeated refutations of alternatives to it. This brings us to the fourth purpose of research.

Replication

In Chapter 1, we argued that researcher biases inevitably affect how observations are gathered and interpreted. The only way to overcome these biases is to replicate the research. **Replication** means that other researchers in other settings with different samples attempt to reproduce the research. If the results of the replication are consistent with the original research, we have increased confidence in the hypothesis that the original study supported.

These then are the ways that research is used to develop, examine, support, and modify hypotheses. The functions or purposes of empirical research in examining hypotheses are not mutually exclusive. A given study is likely to serve a number of functions simultaneously. Research to demonstrate a hypothesis usually ends up as discovery as well. Likewise, replication inevitably involves discovery and refutation, as the conditions of replications change and hypotheses must be modified to account for those changes.

The purpose of empirical research is to inform hypotheses, to enable us to build better and more accurate hypotheses about how human beings behave. Of course, not all research is equally informative or useful in constructing and modifying hypotheses. It is to this issue that we now turn: What makes empirical research more or less useful in helping us to discover, demonstrate, revise, and ultimately support hypotheses?

Criteria for Evaluating Social Science Research

We discuss here four major criteria for evaluating social science research: construct validity, internal validity, external validity, and conclusion validity. To do so, we rely on one of the example hypotheses presented earlier:

> Parents who are more involved in their children's education have children who perform better academically.

Let us suppose that we want to examine whether this hypothesis is reasonable. To do so, we would want to gather information in such a way that our observations would be most informative about the merits of the hypothesis.

Construct Validity

To conduct research that will help determine whether our hypothesis is good or bad, that is, reasonable or not, and whether it should be modified in some way, we first need to measure successfully the theoretical constructs of interest. In this hypothesis, two constructs are involved: Parent involvement in children's education is the first theoretical variable. Researchers might assess it by asking parents to report how often they go to parent–teacher conferences and how often they talk to their children about what they are learning in school. Construct validity would then refer to the extent to which these parent reports assess parent involvement. Children's grade point averages or achievement test scores might be used to assess the second construct, that is, children's academic performance. And construct validity would refer to the extent to which these specific measures adequately assess children's academic performance.

Note that, in this example, a measure of the causal construct, parent involvement, is the **independent variable**. The measure of the affected construct, academic performance, is the **dependent variable**. And, again, the degree to which the specific variables accurately reflect or measure the constructs of interest is known as the **construct validity** of the research. A study has high construct validity to the extent that all constructs in the hypothesis are successfully measured or captured by the specific variables on which the researcher has gathered data.

Internal Validity

Assume we had met the first criterion for useful research, and we had good measures of both parent involvement in children's education and children's academic performance. Suppose we then gathered information on a number of parents and their children and found that, indeed, more involved parents' children performed better. Certainly this result is consistent with the hypothesis. What we do not know, however, is whether our research supports the notion that parent involvement *causes* better academic performance. The second criterion for useful or informative research, known as **internal validity**, concerns the extent to which conclusions can be drawn about the causal effects of one variable on another, that is, ensuring that only the independent variable can account for differences in the dependent variable. In research with high internal validity, we are better able to argue that associations are causal ones, whereas in studies with low internal validity, causality cannot be inferred as confidently because alternative explanations for the effects cannot be dismissed. In short, internal validity refers to the degree to which the research design allows causal conclusions to be drawn about the effect of the independent variable on the dependent variable.

External Validity

A third criterion for useful research is known as **external validity**, that is, the extent to which the results of the research can be generalized to the populations and settings of interest in the hypothesis. In the example we are considering, suppose the constructs were well measured (high construct validity). Suppose further that we found an association between parent involvement and children's academic performance and could reasonably claim that association to be a causal one (high internal validity). We then would want to know whether that causal association held in only the relatively few parents and children we observed in our research or whether we could generalize the causal associations to other parents and children whom we did not observe. According to the hypothesis, the effect appears among all parents and children. Clearly, it would not be possible to observe them all. But we might select for observation parents and children who are representative of a larger population, for example, parents and their children who are enrolled in elementary schools in the U.S., so that we would have greater confidence in generalizing the results of our research to others. Such a study would have relatively high external validity. A study from which generalization is difficult has relatively low external validity.

Conclusion Validity

Conclusion validity refers to the degree to which our data analyses – whether those analyses are quantitative or qualitative – allow us to draw appropriate conclusions about the presence or absence of relationships between our independent and dependent variables. It differs from internal validity in that we are not concerned about whether the relationship is a causal one; rather, we are concerned about whether the quality of the data and analyses provides a reasonable basis for concluding whether a relationship exists.

Although conclusion validity applies to qualitative analyses as well, it has most commonly referred to the statistical factors that affect the ability to reach a conclusion about the presence or absence of a relationship. Statistically speaking, it is possible to make one of two types of errors. A **Type I error** refers to incorrectly concluding that there is a relationship when in fact there is not; a **Type II error** refers to incorrectly concluding that there is not a relationship when in fact there is. The factors that affect the probability of making such errors concern the **statistical power** of the study, that is, the probability of finding the predicted relationship if the relationship truly exists. We do not consider issues of statistical conclusion validity in this book. Suffice it to say that drawing correct conclusions about the presence or absence of relationships requires (a) strong research designs in which (b) high-quality measures are used to (c) gather enough data (i.e., an adequate sample size) to see the effect (and thus one must consider the size of the effect that would be reasonable to expect) and (d) the use of data analyses that are appropriate to the research question and the nature of the data. This book focuses on two of these elements, namely, measurement and research design.

All four types of validity, summarized in Table 2.2, are important in evaluating research. However, their relative importance usually depends on the purposes the research is designed to serve. For instance, in the early stages of a research program, it might be sufficient to measure constructs that are associated with the behavior of interest, without worrying too much about whether the association is a causal one. In other words, in discovery research, construct validity might be relatively more important than internal validity. Or consider research in which the primary purpose

Table 2.2 Definitions of Research Validities

Construct validity:	To what extent are the constructs of theoretical interest successfully operationalized in the research?
Internal validity:	To what extent does the research design permit us to reach causal conclusions about the effect of the independent variable on the dependent variable?
External validity:	To what extent can we generalize from the research sample and setting to the populations and settings specified in the research hypothesis?
Conclusion validity:	To what extent do our analyses allow us to reach appropriate conclusions about the presence or absence of a relationship between the independent and dependent variable?

is replication. Such research is especially concerned with external validity because in replication we are concerned with whether a previously obtained result continues to be found in a new setting at a different time. Because the conditions of the original research and the replication are never identical, we are always examining issues of generalizability in replication research.

The remainder of this chapter concerns the factors that determine whether a study has high or low construct, internal, and external validity; however, our discussion here serves only as an introduction to the subject of how valid and informative empirical research is designed and conducted. The major portion of this book is devoted to this topic as well. Hence, the remaining pages in this chapter serve as an introduction to many of the later chapters, in which some of the same issues are considered in greater detail.

Maximizing Construct Validity

Suppose we wanted to measure children's academic performance to test our hypothesis about the effects of parent involvement. There are a number of ways to measure academic performance. We could give the students achievement tests, look at their grades, ask teachers to evaluate their students verbally, and so forth. Each of these measures is called a variable. Earlier we defined a variable as any attribute that varies across the people or things that we are measuring. Another way to look at variables is to consider them simply as rules or ways of classifying people into different categories so that those who are in the same category are more similar in some way of interest than those who are in different categories. For instance, scores on an achievement test constitute a variable that is thought to measure academic performance. If we line up students according to their scores on the achievement test, we believe that students who are closer together in that rank order are more similar in academic performance than students who are farther apart.

Actually, however, variables never measure only the construct of interest. They measure other irrelevant characteristics as well. Think about an achievement test. To some extent it does measure academic performance; however, it also probably measures test-taking anxiety, motivation to do well, familiarity with English, and so forth. These other factors, in addition to pure academic ability, may influence whether students get relatively high or low scores on the test. To some extent, then, variables reflect not only the construct of interest but also **constructs of disinterest** – things we would rather not measure. Finally, variables contain random errors of measurement. For instance, scores on a test may be affected by recording errors or grading errors. Or students may guess the answers to some questions and these guesses will sometimes be correct and sometimes incorrect.

As shown in Figure 2.1, then, observed scores are made up of three components: (1) the construct of interest, (2) other things that we do not want to measure (constructs of disinterest), and (3) random errors. Thus, if we ordered students based on their scores on the test, that order would not be identical to the ordering that would result if somehow we could order them based on their true academic performance. The best we can do is to develop and administer measures in ways that minimize the

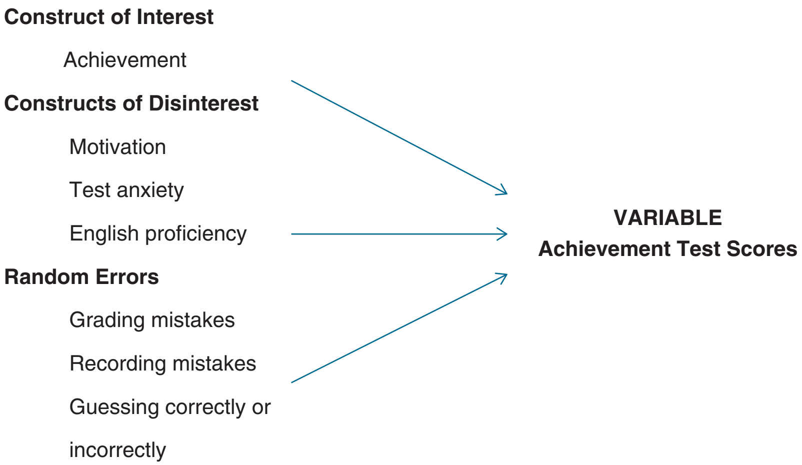


Figure 2.1 Three Components of a Variable.

influences of constructs of disinterest and random errors on the scores that we obtain; to the extent that we are successful, the variable is said to have construct validity.

But how do we know the degree to which a variable has construct validity? We cannot measure true academic performance directly and thus cannot know whether the rank ordering of students on the test is similar to the rank ordering of students on their true academic performance. The only solution is to measure other variables that we think are also measures of academic performance. For instance, school grades measure academic performance – the construct of interest, although they, too, probably measure other things, such as teachers' biases and preferences (e.g., leniency), students' extraversion, and so forth. Nevertheless, we can compare the ordering of students on what we think are our two measures of academic performance, test scores and grades. If the two orderings are similar, and the only thing these two variables measure in common is academic performance, then the similarity of their orderings is evidence for their construct validity.

Let us review the general point. All variables measure not only the construct of interest but other things as well, and we cannot know the true ordering of people on the construct. The best we can do is measure another variable that we think also assesses the construct and then compare orders on the two variables. If the two variables give us similar orderings of people, we have increased confidence that each of them is measuring, among other things, the construct we think they have in common. In short, we need to measure each construct in more than one way. Only if the different measures yield similar results can we have confidence that our variables capture the constructs of interest. Construct validity is thus best evaluated by employing **multiple operational definitions**, or multiple ways of measuring, and then comparing them to see whether they seem to be measuring the same things.

The need for multiple operational definitions and ways to evaluate the quality of variables are discussed in much greater detail in Chapter 8. Here we wish to stress the importance of construct validity. If empirical research is to be useful or informative, it must measure the constructs to which our hypotheses refer. If the observed variables do not have construct validity, there is no way the research can inform our theory. Even worse, poor construct validity may mislead researchers by yielding seemingly positive results that do not actually reflect the constructs of interest. Contributing to this problem is the fact that poor construct validity is harder to detect than problems with other types of validity. For example, there are widely known and accepted procedures for documenting adequate internal validity. If the criteria for establishing internal validity are not met in a given study, it is obvious to other researchers. Establishing construct validity is more challenging, and unless the authors of a research report are careful in describing their measures and validation procedures, it may not be clear whether a given construct was truly assessed.

Exercise: Evaluating construct validity

Construct validity is central to research. It establishes that you are measuring what you think you are measuring. If you are not sure, no one should believe what you find, for you might not be measuring what you think. Find an instrument (personality scale, survey, interview protocol, etc.) that you have used or have read about, and find out how its construct validity has been established. At this point in the book, it would be premature to go further. Anticipating future chapters, we will revisit this example and explore the following questions, so save the instrument that you picked. How was it validated? For what populations has it been validated? Has it been used in ways that go beyond how it was validated? Are there questions about its validity for other populations?

Maximizing Internal Validity

Certain characteristics of the research design affect the internal validity of a study – the extent to which we can infer causal connections from an association between two variables. These characteristics are discussed in more detail in Chapter 10. Our discussion here is intended to provide an intuitive understanding of how to maximize our ability to argue for causal connections.

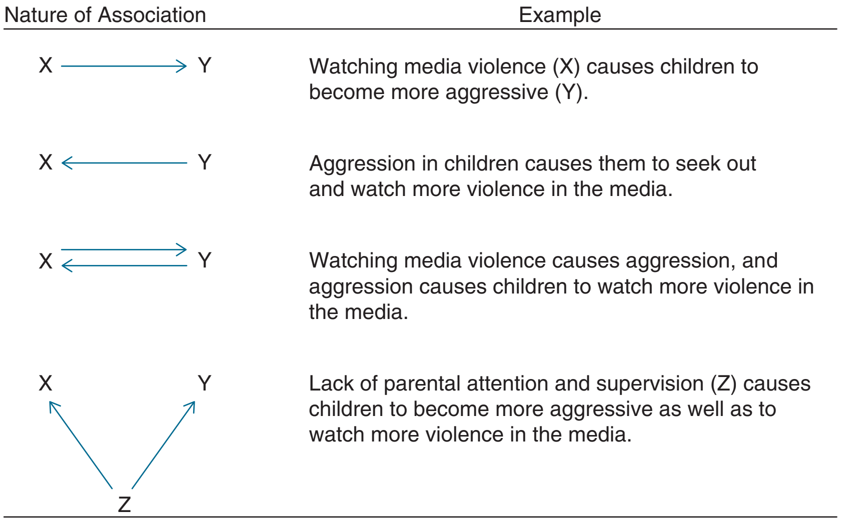
In the parent involvement example, suppose we were able to assess parent involvement and the academic performance of their children. Suppose further that our research had perfect construct validity: Our measures of both parent involvement and children's academic performance measured those constructs and nothing else. Finally, suppose we found that the children of more involved parents tended to have academic performance scores that were higher than those of children whose parents were less involved. In other words, we found that the two variables, parent involvement and academic performance, were related in the predicted way. Could we argue from this association that we have evidence for a causal effect of parent involvement on children's academic performance? We could not.

Simply showing that people or groups (e.g., parent–child dyads) that have high scores on one variable (i.e., degree of parent involvement) have higher scores on a second (i.e., academic performance) does not necessarily mean that one variable causes the other. Although an empirical association or **correlation** between two variables is necessary, it is not sufficient for reaching causal conclusions. Inappropriately inferring causality from a simple association between two variables is called the **correlational fallacy**. This concept is simple yet utterly important: Correlation does not imply causation.

Consider a couple of classic examples in which two variables are associated, but there is no causal effect of one on the other. Elementary school children who have larger feet tend to be better readers; foot size and reading ability are related. Obviously, however, this relationship does not mean that foot size affects reading ability. Rather, age affects both foot size and reading ability, which is why the two are related. Another example: In some European countries following World War II it was noticed that more babies were born where more storks were roosting. What accounts for this association is population density. Where there are lots of people, there are lots of chimneys, where storks are fond of roosting. Likewise, where there are lots of people, there are lots of babies. Hence, storks and babies are found together. The inability to draw causal conclusions may be obvious in these examples, but when one has strong notions about the specific cause of a given effect, it can be difficult to identify alternative explanations and tempting to draw a causal conclusion.

In sum, whenever two variables are associated with each other, there are four possible explanations for their association: (1) variable X causes variable Y; (2) variable Y causes variable X; (3) variable X causes variable Y and Y causes X, in which case we could talk about reciprocal or bidirectional causation; or (4) some third variable, *Z*, causes both X and Y. The latter possibility is often called the "hidden third variable" problem, hidden because the researcher might have only measured X and Y and so cannot examine *Z*, and may not even know what *Z* is. Figure 2.2 illustrates these possibilities with a research question of social importance: What is the nature of the association between media violence and aggression in children? The research findings are very clear on the existence of an association: As the amount of media violence watched by children increases, so does their aggression.

As Figure 2.2 shows, there are four possible explanations for the association between exposure to media violence and aggression. One possibility is that watching violence on television does cause children to be more aggressive. A second possibility is that aggressive children seek out and watch more violent television programming. In other words, the causal path runs in the opposite direction. To make things even more complicated, it is possible that both of these hypotheses are true and that media violence and aggression cause each other in a complicated pattern over time called **reciprocal causation**. Lastly, it is possible that a third variable causes both aggression and the watching of violent television programming; inadequate parental attention and supervision, for example, is one such plausible third variable. It is the job of researchers to consider and examine all of these possibilities and indeed a great deal of research has examined and continues to examine the association between media



Watching media violence (X) causes children to become more aggressive (Y).

Aggression in children causes them to seek out and watch more violence in the media.

Watching media violence causes aggression, and aggression causes children to watch more violence in the media.

Lack of parental attention and supervision (Z) causes children to become more aggressive as well as to watch more violence in the media.

Figure 2.2 Possible Causal Pathways to Explain the Association Between Exposure to Media Violence and Aggression.

violence and aggression. We leave it to readers to consider the strength of the research evidence supporting a causal relationship (see, for example, Bushman & Huesmann, 2012; Gentile & Bushman, 2012).

We use this example to highlight two additional points. First, broad conclusions about the effects of one construct on another require a great deal of research conducted by different researchers who have different perspectives, using different research designs and different variables or indicators of the broader constructs. To the extent that the research methodology varies (and thus so too do the strengths and weaknesses of the studies) and yet the research results converge, we are in a much better position to draw broader causal conclusions (as opposed to a causal conclusion about the effect of a specific independent variable on a specific dependent variable in a single study). Second, the ability to draw causal conclusions does not mean that the effect holds for every single individual. In other words, exposure to media violence may indeed cause aggression, but that does not mean that every child who watches violent shows or plays violent video games will necessarily become aggressive – just as not every individual who smokes will necessarily develop lung cancer. Further, most phenomena are multiply determined so that there may be many factors in addition to media violence that cause aggression and many factors in addition to smoking that cause lung cancer.

The issue of establishing causality matters a great deal in science. Scientists are in the business not only of describing and predicting events but also of controlling them. In the social sciences, especially, we wish ultimately to put our findings to the use and benefit of humankind; to design effective interventions for social problems, we must be able to identify the *causes* of those problems and those causes must be manipulable. Returning to the media violence example, if watching media violence causes aggression, decreasing the amount of gratuitous violence in television shows and movies or otherwise limiting exposure to violent media may be an effective way of reducing aggression. However, aggression may also be caused by other factors, for example, inadequate parenting and access to weapons, in which case interventions that target these factors may also be useful. Identifying causes is crucial. So, too, is the evaluation of interventions designed to "treat" the problem – a topic that is discussed in Chapter 15.

We stress this point because this is an area in which the public is sometimes misled by the news media or other information outlets, such as the Internet. Frequently, findings of research studies are presented in a distorted manner because reporters or news anchors draw an inappropriate or unwarranted causal conclusion from a correlational finding. This may occur even when the researchers who conducted the study were very careful to draw appropriate conclusions. For example, a newspaper headline might state "Exercise Makes Pregnancy Easier" and the accompanying article claim that pregnant women who exercise regularly have shorter and less painful deliveries and healthier babies. The causal claim is explicit, but a careful reading reveals that the scientists had only correlational data available; that is, pregnant women were asked how often they exercised, and frequency of exercise was then related to pregnancy outcomes.

We are not trying to argue that exercise is not good for pregnant women. Exercise may indeed cause better deliveries and healthier babies. Our point is merely that one should not draw that conclusion based on the research finding presented in the hypothetical newspaper article. Other explanations exist for the association between exercise and pregnancy outcomes. Women who are concerned with their health in general probably engage in a variety of activities in addition to exercising that are good for their unborn babies. For instance, they might avoid drinking alcohol or smoking during pregnancy; their diets might be healthier; they might take prenatal vitamins; or they might obtain better prenatal care overall. And it could be one or more of these other factors, not exercise, that accounts for the positive outcomes. As detailed in Chapter 10, the only way we can conclude that exercise is a causal factor is to conduct research, especially randomized experiments, that allows us to rule out these alternative explanations.

Misrepresentations of scientific research abound. Because they usually sound plausible, readers may fall into the trap of agreeing with the causal claim. Our hope is that, even if you do not become a practicing social scientist, you will become a discriminating consumer of findings from empirical research on social behavior. We encourage readers to scrutinize newspapers, magazines, and Internet reports with a critical eye, looking for inappropriately drawn causal inferences. Their frequency will be both surprising and disheartening.

To illustrate other threats to internal validity, we return to our example of parent involvement and children's academic performance. Assume that parents were assigned to be either highly involved or not involved in their children's education. Assume further that children in the two groups subsequently differed in their academic performance. We cannot argue for causality in this case because it is possible that the parents and/or children in the two groups differed in other ways that would account for the difference in performance. In particular, the children's academic performance may have differed initially prior to their parents being involved or not. Indeed, participants in the two groups may have differed in many ways that could account for the subsequent difference in children's performance. These types of initial differences between groups of research participants that may affect the dependent variable represent a **selection threat** to internal validity.

How might we get around this selection threat? One way might be to place the parents and their children in the involved and uninvolved groups so that there were no initial differences in academic performance. If we could do so, we could be more confident that differences in performance later on were not due to initial differences in academic performance. As we have suggested, however, the parents and children in the two groups may differ in other ways, for example, motivation, that could influence subsequent academic performance. Indeed, we might observe a difference simply because children in the two groups were changing or learning on their own at different rates. Hence, if we find a difference in academic performance at a later time, we still cannot infer that the difference is caused by the difference in parent involvement. The difficulty of reaching causal conclusions because the individuals in the two groups might be growing or maturing at different rates is known as the **selection by maturation threat** to internal validity.

What is needed is a way to equate the parents and children in the two groups not only now but also in the future. There is really only one way to accomplish the goal of establishing equivalent groups. Suppose for each parent we flipped a coin. Certainly, there is no reason to expect that heads or tails would be related to the child's academic performance now or in the future, nor would we expect the result of the coin toss to be related to hair color, height, or any other person characteristic. By definition, a variable whose values are randomly determined, like the flip of a coin or the throw of a die, is unrelated on average to all other variables now and in the future. Hence, if we decided who was to be in the high involvement group and who is to be in the low involvement group by a flip of a coin, we would expect no differences in academic performance later if parent involvement made no difference.

The lesson is that we can infer causality from the association between two variables only if people have been randomly assigned to the levels of the independent variables. Parent involvement in children's education is the independent variable in our example and it has two levels: highly involved and uninvolved. If it were related to academic performance later and if children were assigned to its levels on a random basis, we would be able to argue that it had a causal effect on academic performance, the dependent variable. Research studies carried out in this manner, with random assignment to the independent variable, are called **randomized experiments**. They are discussed in considerable detail in Chapter 10.

Although randomized experiments are the best choice if causal conclusions are to be drawn from the research, they require researchers to have a great deal of control. Researchers must be able to determine who is in which group, for example. Frequently, such control over the independent variable is impossible. It would be difficult and unethical to ask parents to be uninvolved in their children's education. When such control is not possible, some type of **quasi-experimental research** may be used instead. Quasi-experimental designs are discussed in Chapter 11. Briefly, quasi-experiments are those in which research participants are not randomly assigned to levels of the independent variables. Although they do not permit causal inferences with the same degree of confidence as randomized experiments do, they are essential tools for social scientists. Although some internal validity is sacrificed, quasi-experiments can still yield exceedingly rich and useful information. Randomized experiments are useful and particularly valuable for causation, but they are not the only tools in the researcher's bag.

Maximizing External Validity

In Chapter 9, we consider procedures designed to increase the external validity of research, that is, procedures that increase our ability to generalize the research results to the populations and settings of interest. We introduce these procedures here, again using our example about the effects of parent involvement on children's academic performance.

Suppose we had measured well the two constructs, parent involvement and children's academic performance, and had done what we could to ensure internal validity. How would we ensure that our research results were generalizable to the extent we desired? First, rather than remaining implicit in the hypothesis, the population and setting to which generalization is sought should be made explicit before the research is conducted. We need to define as precisely as possible the group of people and the settings for which we think our hypothesis holds. For instance, we could be a bit more precise by saying that we expect parent involvement to positively affect the academic performance of children enrolled in public schools located in large U.S. cities. If we gathered data from the entire population and found support for our hypothesis, generalization to the desired population would not be a problem.

However, it is neither efficient nor necessary to measure every person in the population or every setting of interest. Rather, we can gather data from a sample of the population. To enhance generalization, we want to select a sample so that it is representative of the population. But how would we do so? The only way we can be confident about generalizing from a sample to a population is to draw a probability sample, for example, a **random sample**. Obtaining a random sample involves doing something like flipping a coin to determine whether each member of the population is to be included in the sample.

Note that random sampling is not the same as random assignment. Using a random process to select a sample from a population is done to enhance our ability to generalize, that is, external validity. Using a random process to assign participants to levels of the independent variable is done to increase internal validity, the ability to reach a causal conclusion about the effect of the independent variable on the dependent variable.

Frequently in the social sciences it is not practical to draw a random sample. We might wish to generalize to parents and children across the country, but it may not be possible to obtain such a sample or to conduct a study that involves people who are spread across such a large geographic area. Generalization must then be done on a theoretical or conceptual basis. We must speculate about how parents and children whom we have not observed might differ from those we have, and then we must decide whether those differences would be expected to influence whether parent involvement affects children's academic performance. Such speculation ultimately gives rise to further research. Indeed, replicating research in other settings and with other samples is an important part of maximizing external validity.

Basic and Applied Research

Throughout this chapter, we have focused on use of research to develop, test, and refine theories. Such research is referred to as **basic research**. But not all research is intended to test theories. Some, for example, is intended to *apply* theories to real-world settings to see whether the theories can help improve outcomes. Other research is intended to answer practical questions about how well different methods or approaches work in particular settings and with specific populations. Some of the latter types of research may not apply any disciplinary knowledge, but examine effectiveness of intuition of practitioners or practices developed atheoretically. Not surprisingly, such research is called **applied research**, or, more recently, **translational research**. If basic and applied research are viewed as ends of a continuum, most research falls somewhere between the end points, for basic research simultaneously can be used to develop practices that work, and applied research can help develop and refine theories. Lewin (1946), for example, argued that practical settings are great places to develop theory, for theories should be relevant to the world and help explain everyday human behavior. In his words, "there is nothing so practical as a good theory." If one agrees with Lewin, then research becomes *both* basic *and* applied, for it develops and applies theories to address practical issues, and we should not view basic and applied research as ends of a continuum but as two dimensions (basic/not and applied/not) of research types.

Summary

There are two major foci of this chapter. The first concerns the purposes of empirical research for the scientific study of social behavior. We argued that research is used fundamentally to examine hypotheses and develop theories. As such, research can be used for discovery, demonstration, refutation, and replication.

Discovery is the inductive process of gathering data to formulate hypotheses. Demonstration is predominantly a deductive process, gathering data that we hope are consistent with a hypothesis. Although such demonstrations can be used to support a hypothesis, the hypothesis can never in fact be proven because there always remain alternative ways to account for a research finding. Research as refutation involves the attempt to refute competing hypotheses, that is, to show that alternative explanations for previous results are not valid. Finally, research as replication involves repeating research with different samples or in different settings to increase confidence in a previous demonstration. In all four cases, discovery, demonstration, refutation, and replication, the ultimate reason for gathering empirical data is to develop, support, evaluate, and refine our hypotheses so that they do a better job of describing and explaining social behavior.

In the second half of the chapter, we defined four criteria that determine the extent to which research is useful in examining hypotheses: construct validity, internal validity, external validity, and conclusion validity. Research has high construct validity if the variables that are in fact measured correspond closely to the constructs that the hypotheses implicate. Research that is internally valid permits us to reach causal conclusions about the association between the independent and dependent variables. Research that is high in external validity enables us to generalize the results from the sample studied to the population and settings of interest. Finally, conclusion validity concerns whether the quality of our data and analyses is adequate for drawing conclusions about the presence or absence of a relationship between our independent and dependent variables. In addition to defining these validities, we discussed the basic conditions for achieving each one. Finally, we introduced the distinction between basic and applied research. This discussion serves to introduce the more complete presentations in subsequent chapters.

Go online Visit the book's companion website for this chapter's test bank and other resources at: www.wiley.com/go/maruyama

Key Concepts

Applied research
Basic research
Conclusion validity
Construct validity
Constructs of disinterest
Correlation
Correlational fallacy
Deductive
Demonstration
Dependent variable
Discovery
External validity
Falsifiability
Hypotheses
Independent variable
Inductive
Internal validity
Measurement research
Multiple operational definitions
Parsimony
Quasi-experimental research
Random sample
Randomized experiment
Reciprocal causation
Refutation
Replication
Selection by maturation threat
Selection threat
Statistical power
Theory
Translational research
Type I error
Type II error
Variable

On the Web

http://www.burns.com/wcbspurcorl.htm Very clearly written discussion of the correlational fallacy, with lots of examples and quotes from relevant readings.

http://faculty.washington.edu/chudler/stat3.html Great web page entitled "How to Lie and Cheat with Statistics." Reviews how graphs and charts can be misleadingly created and reported in the media. Nicely done and interactive.

http://chem.tufts.edu/science/FrankSteiger/theory.htm Good discussion of the distinction between facts and theory, in the context of the evolutionary debate.

Further Reading

Blastland, M., & Dilnot, A. (2010). *The numbers game: The commonsense guide to understanding numbers in the news, in politics, and in life*. New York, NY: Gotham Books.

Cross, C. (1996). *The tainted truth: The manipulation of facts in America*. New York, NY: Simon & Schuster.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243*, 1668–1674.

Huff, D. (1993). *How to lie with statistics*. New York, NY: Norton.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*, 806–834.

Paulos, J. A. (2001). *Innumeracy: Mathematical illiteracy and its consequences*. New York, NY: Hill & Wang.

Stanovich, K. E. (2013). *How to think straight about psychology* (10th ed.). Boston, MA: Pearson.