Behavioral Science Research Methods

Stacey A. McCaffrey, Victoria A. Schlaudt, and Ryan A. Black

Research Design

In a grave crisis situation, such as a hostage negotiation, it may be foolish and reckless for an individual to engage in behavior that is not supported by research—behavior that would be defensible to a police captain, lawyer, or jury. Yet, every day people rely on their intuition to guide their decisions; some of these decisions are benign, while other choices have serious consequences. Decades of research have repeatedly shown how easily influenced people are by pseudo-science, attractive individuals, authority figures, and others who they view as similar to themselves (e.g., Milgram, 1963). Therefore, criminologists must rely on research in order to overcome the flaws inherent in intuition and reduce crime most effectively. For example, many readers may be familiar with the Drug Abuse Resistance Education (D.A.R.E.) program, a drug prevention program for kindergarten through 12th grade. First launched in 1983, the D.A.R.E. program is currently being used in all 50 states and 53 other nations (D.A.R.E. Board of Directors, 2014), with estimated annual costs between $1 and 1.3 billion (Shepard, 2001). However, the results of multiple empirical evaluations have concluded that the DARE program is ineffective and may even be counterproductive (e.g., Clayton, Catterello, & Johnstone, 1996; Ennett, Tobler, Ringwalt, & Flewelling, 1994). Other programs that have been found to be ineffective or even detrimental through research include boot camps to reduce recidivism (e.g., Bell, 2012; MacKenzie & Souryal, 1994; Wilson, MacKenzie, & Mitchell, 2005) and peer-group interventions to reduce delinquency in adolescents (Dishion, McCord, & Poulin, 1999).

These examples illustrate the importance of using research to guide decision-making and to critically evaluate existing programs. Examples used throughout this chapter will elucidate other uses for research and explicate the valuable information that can be obtained from methodologically sound research studies.

This chapter reviews foundational descriptive, quasi-experimental, and experimental research design strategies, while highlighting each of their strengths and weaknesses. Clinical applications of each design will be presented, and their unique vulnerabilities to validity threats will be discussed. A summary of the different designs is presented in Table 1. After

S.A. McCaffrey (✉)
Inflexxion, Inc., Waltham, MA 02451, USA
e-mail: smccaffrey@inflexxion.com

V.A. Schlaudt • R.A. Black
College of Psychology, Nova Southeastern University, Fort Lauderdale, FL 33314, USA
e-mail: vs517@mynsu.nova.edu; blackrya@nova.edu

V.B. Van Hasselt, M.L. Bourke (eds.), *Handbook of Behavioral Criminology*,
DOI 10.1007/978-3-319-61625-4_5

Table 1 Categories of research

Category	Key characteristics	Purpose	Types/designs
Descriptive research	• The systematic observation of behavior • There is no attempt to manipulate the variables • Data collection may occur at one point in time, across multiple time points, or may include analysis of previously collected data (archival)	• Gather information about a group or individual • Develop theories and hypotheses • Explore the statistical and temporal relationship between variables	• Naturalistic observation • Participatory observation • Case study • Survey/Interview • Case–control • Cohort design
Quasi-Experimental	• There is manipulation of the independent variable(s) • Random assignment is not feasible or appropriate • Limited ability to evaluate causality • Includes data collection at multiple time points	• Examine differences between groups on specific variables of interest	
Experimental	• Random assignment is used to assign participants to groups • There is manipulation of the independent variable(s) • Includes data collection at multiple time points	• Temporal precedence, statistical association, and nonspuriousness can be evaluated • More resistant to internal validity threats, increasing the researcher's ability to assess causality	• No treatment control • Waitlist control • Nonspecific treatment or attention placebo • Treatment as usual

reading this chapter, the reader should have a firm grasp of the various foundational research design strategies and should have the skills to develop a strong study that is resistant to internal validity threats. Perhaps more importantly, information presented in this chapter should alert the reader to various flaws that are inherent in research, boosting the reader's critical thinking skills that are necessary to function effectively in nearly any field of study.

As this chapter is intended for a more scientifically mature audience, it is assumed that the reader has an understanding of some basic principles of research, such as the definition of: a theory, a hypothesis, dependent and independent variables, random selection and random assignment, and an operational definition. For a more rudimentary discussion of research design in criminology, the reader is referred to Schmalleger (2006).

Descriptive Research

Purpose

A researcher may decide to conduct descriptive research in order to gather more information about an individual or a group of people that share some common characteristic. For example, the researcher may:

- Wonder how early exposure to physical abuse is related to delinquent behavior in adult life.
- Be interested in the effectiveness of current interventions to reduce depression in inmates.
- Want to design a new treatment to reduce the incidence of drunk driving among those who abuse alcohol.

- Need to understand the relationship between physical aggression and violent video games.
- Like to identify the factors that put an individual at risk for joining a gang.

Goals of descriptive research depend upon the research question to be answered. These goals may include generating theories and hypotheses about human behavior, or exploring the statistical or temporal relationships between variables. In many cases, the goal of the research study will influence the study design and data collection strategies that the researcher will choose. Data may need to be collected across multiple time points by employing a longitudinal design (e.g., gathering information from the same person or group of people once a month for a full year), at one point in time with a cross-sectional design, or by reviewing and analyzing previously collected data through archival research. Additionally, the researcher might choose to collect data from a group of participants over time (a *prospective* longitudinal study), or might choose to collect data from participants *retrospectively*, after an outcome has occurred. Although not always feasible, prospective strategies should be used whenever possible due to increased reliability and accuracy. Although studies have shown little agreement between "real-time" and "retrospective" data (e.g., Nagurney et al., 2005), retrospective instruments vary regarding their degree of reliability. For instance, the Time Line Follow Back (Sobell et al., 1996), a retrospective calendar-based measure of daily substance use, has demonstrated excellent reliability for assessing retrospective substance use for up to 1 year (Robinson, Sobell, Sobell, & Leo, 2014). Whether considering retrospective or prospective data collection methods, psychometric properties of the data collection strategy should always be explored to increase the likelihood that data is accurate and meaningful.

Gathering Information

Descriptive research strategies can provide rich sources of data, helping researchers better understand a phenomenon of interest, and assisting with the development of conceptual models, theories, and hypotheses. For example, if a researcher is wondering what variables put an individual at risk for joining a gang, he/she may decide to conduct unstructured interviews with family members of adolescents who joined gangs as well as family members of adolescents who did not join gangs. Examining differences between the two qualitative profiles may help the researcher develop theories about the risk factors for gang participation. This information could eventually be used to develop deterrent programs for youth who are considered to be at high risk.

Importantly, when the purpose of research is to gather descriptive information, causational conjectures cannot be made. That is, just because a team of researchers noticed that *Variable X* occurred more frequently in the profile of individuals who joined gangs than in the profile of individuals who it does not provide evidence that *Variable X* caused a person to join a gang. Nevertheless, they may develop a hypothesis from this type of research, which may fuel further exploration through an experiment to evaluate the causative relationship between variables.

While many different types of research designs might be employed when a researcher's goal is to gather information, the aforementioned example (i.e., exploring what variables put individuals at risk for joining gangs), would be considered a retrospective case–control design, a type of observational design where the researcher gathers information about two groups of adolescents (adolescents who joined gangs and adolescents who did not) to examine differences between the groups after the outcome (gang participation) already occurred. This type of study is not intended to provide causal information, but instead to create a theory or a hypothesis that can be tested in future research.

Statistical Association

Another use of descriptive research is to determine whether or not there is a statistical association between variables. For instance, a researcher may be interested in learning whether or not there is a relationship between education and delinquent behavior. To answer this question, the researcher would collect quantitative data about years of education (the independent variable, or IV) and frequency of delinquent behavior (the dependent variable, or DV), and then calculate a correlation coefficient between these variables. The correlation coefficient, often a *Pearson product moment correlation coefficient*, is a quantitatively derived description of the linear relationship between two variables, and ranges from −1 (a perfect inverse relationship between two variables, whereby if one increases the other simultaneously decreases at the same rate) to 1 (a perfect relationship between two variables, whereby as one variable increases or decreases so does the other in the same direction at the same rate). A correlation of 0 would indicate a lack of relationship between two variables. It should be noted that it is extremely rare to find a perfect correlation between variables because behaviors are multiply determined; that is, the DV is influenced by other factors aside from the IV. In fact, if two variables are perfectly correlated, the researcher should explore why a perfect correlation exists, and may determine that there is an artificial reason for the perfect correlation.

When examining the statistical relationship between variables, the researcher should also be mindful of the possibility that a confounding variable may wholly or partially explain the relationship between the IV and DV. For example, there is a strong positive correlation between ice cream sales and violent crimes, and there is also a strong positive correlation between the number of churches in a town and the frequency of violent crimes. In each of these situations, there is a third variable (warm temperatures, geographic areas of low socioeconomic status) that explains the relationship between the IV and DV. Ignoring the impact of confounding variables while also forgetting that correlation does not equal causation may incorrectly lead a researcher to hypothesize that eating ice cream causes a person to become violent.

Further, when two variables are found to be strongly correlated, the correlation coefficient does not provide any indication about which variable occurred first and may have influenced the other variable. For example, a city that has fewer guns may have fewer violent crimes (a strong positive correlation), but does the low rate of crime explain why so few people have guns, or does the low rate of guns explain why there are so few violent crimes? Archival data of registered guns and police reports of violent crimes in the city from the last 20 years may help the researcher determine which came first, a concept which is discussed next.

Temporal Precedence

Aside from understanding what variables are statistically related, a researcher may also be interested in knowing the order in which variables or behaviors occur. For example, the researcher may want to know whether: (a) exposure to violent video games led to physically aggressive behaviors, or (b) if the physically aggressive individual was simply drawn to these games, and consequently spent more time playing than those who were not physically aggressive. Understanding the order in which events occur is necessary to determine causation, and is called *temporal precedence*. Temporal precedence is best established with longitudinal prospective designs, as cross-sectional designs that ask individuals to recall the temporal order of events are generally less reliable.

It is important to note that just because two variables are statistically related and one occurred before another, does not mean that the first *caused* the second. In order to determine causation, a researcher would need to implement an *experiment*, which is described in detail below.

Descriptive Research Designs

While there are many types of descriptive research, six commonly used descriptive research strategies will be discussed in this chapter: naturalistic observations, participatory observation, case studies, surveys/interviews, case–control designs, and cohort designs. Depending on the research question, the behavior of interest, ethical/legal guidelines, and available resources, one or more of these descriptive research strategies may be most appropriate. Each of these descriptive research methods have strengths and weaknesses, and threats to validity must be carefully considered prior to initiating a study. Moreover, researchers must obtain appropriate institutional review board (IRB) approval to ensure that any research project that involves human subjects is being conducted in accordance with federal, institutional, and ethical guidelines (see Sect. "Ethical Issues" below).

Naturalistic Observation

In a *naturalistic observation*, or passive observational study, the researcher does not intend to manipulate the environment or the individuals that he/she is observing in any way; in fact, in order to obtain the most objective data and prevent observer reactivity (i.e., conscious or unconscious change in behavior as a consequence of knowing that one is being observed), the researcher may take precautions in order to ensure that the individuals being observed are unaware that they are being observed.

Before beginning an observation, the researcher should operationalize the behavior that he/she is going to observe, and decide: who will be observed, the setting in which the participant(s) will be observed, and when to observe. Depending on characteristics of the behavior being observed, the researcher must determine the frequency of the observation(s), the length of observation(s), the duration of the observation(s) (over 1 day, over several months, etc.), and how the data should be collected. Narrative, event-based, or time-based recording strategies may be considered (see Suen & Ary, 2014). For example, a researcher may utilize a longitudinal "high risk" cohort design (Kazdin, 2003), where a group of individuals who were exposed to a risk factor (e.g., children who have a history of being physically abused by a biological parent and were removed from the home) are observed over time. Specifically, the researcher may choose to observe the children for 30 min on the playground using an event-based coding scheme, collect academic test scores (considered "permanent product recording"), and have the children's teachers complete surveys once per month for 12 months following the last incident of abuse.[1]

Because data are being collected by humans, there is always error and subjectivity. However, there are steps that researchers can take when collecting observational data to increase reliability and accuracy. Strategies to increase the psychometric properties of the observational data should also be considered before data collection begins. For instance, two raters may be trained in data collection procedures and may be required to reach an interobserver agreement of 0.80 before data collection can commence. Of course, depending on the type of data being collected, different types of agreement (e.g., occurrence and nonoccurrence, correlation between two sets of scores) may be considered (Hintze, 2005).

According to generalizability theorists, an essential property of observational data is that data are *generalizable* across observers, time, setting, methods, and targets. That is, the incidents of physical aggression observed in the prison exercise yard collected during five 3-h observations this week using a specific event-based coding scheme are not useful for the researcher to understand the "usual frequency of physical aggression in prison exercise yards" unless this observational data is generalizable across observers, time, settings, method, and targets (i.e., prisons being observed). Although a discussion of generalizability theory is beyond the scope of this chapter, the reader is encouraged to consider the factors that may be limiting the generalizability and external validity of their observational data when making decisions. It is important to not make unfounded statements regarding the generalizability of findings.

Participatory Observation

Conversely, there are instances when a researcher may choose to interact with the individual or group that he/she is observing in order to gain a

[1] In this example, data was collected through multiple modalities and included multiple raters. This type of data collection is considered optimal because it provides the greatest amount of information and reduces error associated with each source of data.

more complete understanding of the group dynamic. For instance, policemen go undercover and join gangs or drug rings in order to learn more about the motives, roles, and inner workings of these groups. However, by joining these groups the police officers must take into account that their presence and behavior undoubtedly impacts group functioning, possibly biasing and distorting the data that they collect. Further, participating as a member of a gang or drug ring unconsciously impacts perceptions and beliefs, which may also threaten validity of observational findings. These biases must always be taken into account when interpreting participatory observation data.

Unique ethical issues arise in both naturalistic and participatory observation studies, because the individuals being observed are not consenting to be observed. Further, in the example above, there may be instances when an officer acts in unethical ways when he is engaging in activities with the gang members. The manner in which the officer handles these transgressions must be in line with ethical guidelines and the benefits must outweigh the costs. Thus, IRB approval is vital when considering an observational study, a topic which is discussed further in Sect. "Ethical Issues" below.

Case Study

One type of descriptive research, a *case study*, is particularly helpful when the subject of interest is extremely rare, such as studying aggressive behavior and lack of empathy in men with XYY syndrome. There are also instances where a researcher may be interested in studying human behavior that cannot (ethically or otherwise) be manipulated through an experiment. For example, researchers have interviewed serial killers in order to understand their backgrounds, the ways in which they view themselves and the world, criteria for victim selection, and their motivations and behaviors (Beasley, 2004). Although information obtained from interviews with these serial killers provides detailed and personal data, information acquired from a handful of self-selected male criminals may not generalize to other serial killers or other types of perpetrators (the limits of external validity are discussed later in this chapter). Further, the veracity of this type of data cannot always be verified (and serial killers may be particularly prone to lying as part of their pathology!). In case studies, the possible benefits of gathering rich qualitative data to help inform theory often outweigh the cost and obvious external validity limitations when other types of research are not feasible.

Surveys and Interviews

Surveys, behavioral rating scales, and unstructured, semistructured, and structured interviews are frequently incorporated into research studies as methods to gather information. Surveys that are delivered and completed over the internet may be particularly cost effective for researchers, but have lower response rates (33%; Nulty, 2008) than paper surveys handed out face-to-face, or interviews conducted face-to-face or over the phone. Examples of surveys and rating scales that may be used in the field of criminology include the Personality Assessment Inventory (PAI; Morey, 1991, 2007), the Detailed Assessment of Post-traumatic Stress (DAPS; Briere, 2001), the Trauma Symptom Inventory, Second Edition (TSI-2; Briere, 1995), the Minnesota Multiphasic Personality Inventory, Second Edition (MMPI-2; Butcher et al., 2001), and the Test of Memory Malingering (TOMM; Tombaugh, 1996).[2]

In addition to formal surveys and interviews, data can also be collected by asking the individual to self-monitor their behaviors through journaling or electronic data collection strategies such as cellular phone applications. Other monitoring devices, such as wearable health technology, may also be used to gather data, and the researcher should ensure that these technologies are compliant with healthcare laws and patient privacy, such as the Health Insurance Portability and Accountability Act (HIPAA).

[2] It should be noted that some of these instruments must be administered, scored, and interpreted by a licensed psychologist, an individual working under the supervision of a licensed psychologist, or another qualified mental health professional.

Case–Control Design

Case–control designs refer to strategies where the researcher studies groups of individuals who vary on the outcome, or criterion, of interest. The most basic case–control design includes two groups, one group (the "case") which has the criterion of interest (e.g., a history of physical abuse as a child), and one group (the "control") which does not. By evaluating differences between the two groups, the researcher hopes to better understand the impact of the criterion on the groups of individuals who were exposed to it (Kazdin, 2003). Case–control designs may be cross-sectional, where case and control groups are selected and assessed in relation to current characteristics. They may also be retrospective, where the goal is to draw inferences about some antecedent condition that has resulted in, or is associated with, the outcome (Kazdin, 2003). For example, the researcher may be interested in studying two groups of adolescents who were both exposed to physical abuse as a child; the first group (case) has a history of delinquent behavior, while the second (control) has no history of delinquent behavior. By studying these two groups, the researcher hopes to identify variables that are associated with the outcomes. As previously discussed, when participants are asked to retrospectively recall information, this information is subject to biases and distortions, including selective recall, inaccurate recall, and recall biased by the outcome (Kazdin, 2003). Therefore, prospective designs are generally preferred to retrospective designs.

Cohort Design

Also known as a prospective longitudinal study, the cohort design involves following groups of people over time to identify variables that lead to an outcome of interest. For example, the researcher may be interested in following groups of young children from neighborhoods with high gang involvement in order to identify the variables that lead to future gang participation. By tracking these children over time before they are involved in gangs (and assuming not all of the children end up joining gangs), the research will have the ability to quantitatively determine which variables predict gang participation and which variables may constitute protective factors, and predict resistance to gang involvement. This example is considered a single-group cohort design, because one group of children is studied over time to evaluate the emergence of an outcome (gang involvement). A researcher may employ a *multi-group cohort design* by following two or more groups who vary on a criterion (e.g., exposure to a risk factor) over time, or an *accelerated multi-cohort longitudinal design*. This modified longitudinal design is able to reduce the amount of time need to complete the full longitudinal study by evaluating multiple groups of individuals who are different ages across shorter intervals of time. For example, three groups of children (5, 8, and 11 years) may each be tracked for 2 years to understand the developmental path to gang participation from ages 5 to 13 without having to study the same group of children for 8 years. The accelerated multi-cohort longitudinal study may also reduce cost and attrition associated with more traditional lengthy longitudinal designs.

Experimental Research

Whether a researcher chooses to use a descriptive research strategy or an experimental method depends on the research question, the population of interest, ethical/legal matters, and available resources. The key difference between descriptive research and experimental research has to do with whether or not the researcher is manipulating the IV. Examples of IVs that cannot or should not be manipulated include: exposure to physical or sexual abuse, gang participation, the number of times that a person has committed a violent crime, frequency of engaging in drunk driving, etc. In these cases, a descriptive strategy should be used in lieu of an experimental design. However, when feasible and appropriate, an experiment is preferable over descriptive research because an experiment is able to provide information that is not attainable through other designs—namely, it can determine causality.

There are many different types of experimental research designs, several of which will be explicated in this chapter. Frequently, researchers will incorporate a *control group* into the design that does not receive the experimental manipulation (IV). The purpose of the control group is to evaluate the effects of the IV on the DV *above and beyond* changes observed in the DV that are not due to the IV. For instance, if a researcher implements a new depression intervention for inmates, reduction in depressive symptoms at the end of the intervention could be explained by numerous factors aside from the actual intervention—the prison may have increased the time that inmates spend exercising or the time that they spend outside in the sunlight, a new friendlier warden may have been transferred to the prison, depression may have decreased with the coming of the spring months, or depression symptoms may have simply remitted over time. Similarly, the researcher might find that depressive symptoms actually increased following exposure to the intervention. While it is possible that the intervention was actually counterproductive and increased depressive symptoms, this change could also be due to other environmental factors. Therefore, the researcher may include a control group in the study as a way to measure the effect of these other variables on inmates' depressive symptoms. Without a control group, the researcher would not have a "bar" to compare results from the intervention group, making it impossible to know whether or not the intervention was effective.

Through various design strategies, the researcher attempts to control for variables aside from the IV that may influence the DV, which are considered threats to internal validity. In general, random selection and random assignment of participants to experimental and control conditions can help control for many threats to internal validity. Just as there are instances when the experimenter is unable to randomly select participants, there are also occasions when the experimenter cannot (for feasibility or ethical reasons) randomly assign participants to the experimental or control groups. For example, a researcher may implement Intervention A (radio messages warning against drunk driving) in City A, and Intervention B (road signs warning against drunk driving) in City B, to evaluate their impact on DUI rates. Because the researcher cannot randomly assign people to live in one of the two cities during the course of the study, a quasi-experimental research design may be necessary.

Quasi-Experimental Designs

The key differentiating factor between experimental and quasi-experimental designs is that in the latter, the researcher does not randomly assign participants to groups. One of the serious threats to internal validity inherent in a quasi-experimental design is *selection*, as the lack of assignment makes systematic group biases more likely (Kazdin, 2003). Although researchers can identify potential confounding variables and attempt to control for them statistically through regression-based models, there is no guarantee that these corrections will equalize groups on important variables. For example, if the group of children who were abused largely came from single-parent homes while the children who were not abused came from two-parent households, it is not possible to truly "equalize" these groups of children through statistics.

Moreover, just because the groups may be equal on measured characteristics (e.g., family income, number of children in the household) that are believed to influence the DV, does not mean that they are equivalent on *unmeasured* characteristics. If these unmeasured characteristics are related to the DV, they can threaten the internal validity of the study. For example, adolescents who were physically abused as children likely differ from same-age peers who were not abused in many important ways. That is, teens with a history of abuse were likely exposed to countless other risk factors, such as inadequate housing, limited education, exposure to violent behavior, poor diet, etc. Consequently, without random assignment it is not always possible to statistically equalize groups of people on important variables that may influence the DV. Researchers must consider this possibility when drawing conclusions from a quasi-experiment.

Features of Experimental Designs

In a true experimental design, participants are randomly assigned to study groups, and are strategically exposed to certain conditions or variables for a certain period of time, as determined by the experimenter. Before introducing the various experimental designs, several important features of experiments are discussed, including the selection of participants, ethical issues, and random assignment. Next, popular research designs are presented, including the no treatment control, waitlist control, nonspecific treatment or attention placebo, treatment as usual, and multiple group designs.

Selection of Participants

In experimental studies, there are several ways to select participants. First, they can be *randomly selected* from the population of interest so that each individual from this population has an equal chance of being chosen to participate. Use of random selection is intended to increase the representativeness of the sample that is collected, decreasing potential bias in the sample.[3] Although random selection is highly preferable, it is not always possible. For example, if membership in a particular population is unknown (e.g., serial killers, drug dealers, sexual abuse victims), random sampling is impossible.

When considering potential participants for a sample, the researchers should consider the population to which they want to generalize the results. That is, studying drug use behaviors in adolescents from private schools and middle to upper socioeconomic backgrounds would not generalize, or accurately represent, drug use among adolescents from public schools and low socioeconomic backgrounds—an issue pertaining to external validity, discussed later in this chapter. Alternatively, a researcher may be interested in studying inmate behavior X. For practicality reasons, it would be impossible to collect a random sample from the population of all inmates in the world. However, random sampling would become more feasible if the population is defined as "inmates from maximum security prisons in the Midwest United States." However, generalizability of results obtained from this study to other inmate populations is unknown—an empirical question that would need to be answered through follow-up studies.

Another way to select participants is through *convenience sampling*. This sampling method is "convenient," because the participants may have volunteered to participate or are easily accessible. Beyond accessibility, there is no scientific reason for the particular individuals to participate. While convenience sampling is generally economically advantageous when compared to random sampling strategies, the sample produced by convenience sampling cannot be considered representative, seriously threatening external validity of the study. For example, if researchers are interested in the effectiveness of different types of therapeutic preventative interventions in a prison setting to reduce recidivism, those who volunteer to participate may differ in important ways from those who declined participation. Inmates who volunteered for the study may be highly motivated, or may have different attitudes toward therapy than those who did not volunteer. Therefore, the results may not extend to other groups of inmates. Alternative sampling approaches, such as cluster or stratified sampling may also be considered, and the reader is referred to Daniel (2012) for more information about sampling.

Ethical Issues

The Nuremberg Code, created in 1948, was the first document to advocate for voluntary informed consent from human subjects. This international document was developed following criminal proceedings against German physicians and administrators in 1946, which included charges for conducting medical experiments on thousands of concentration camp prisoners without consent. Most subjects died or were permanently crippled as a result of the experiments. The continued violation of human rights in research studies

[3] While random selection strategies are intended to provide a representative sample of the population, random selection strategies sometimes fail, producing a sample that is not similar on important characteristics (e.g., race, age, gender) to the population. For information about alternative sampling strategies, see Daniel (2012).

(e.g., the Tuskegee Syphilis Study (Center for Disease Control, 2015), Milgram's Obedience Study (Milgram, 1963), Zimbardo's Stanford Prison Experiment (Haney, Banks, & Zimbardo, 1973)) led to the National Research Act (1974) and development of the Belmont Report. This report, considered a foundational document for the ethics of human subjects in the United States, outlines the principles of respect for persons, beneficence, and justice in research (United States, 1978).

In 1981, the United States government developed Institutional Review Boards (IRBs), and mandated that all research involving human subjects cannot initiate until it has been reviewed and approved by an IRB. The explicit purpose of the IRB is to ensure that all research studies are ethical, safe, and in line with federal, state, and international guidelines. In order to gain approval from the IRB, a study must take steps to reduce risk and to maximize benefit, and benefits from the study must outweigh potential risks for participants.

The IRB also classifies certain populations as "vulnerable populations," providing additional protection to ensure that their rights are protected. Vulnerable populations, as defined by the IRB, include: pregnant women, human fetuses and neonates, prisoners, children, cognitively impaired persons, students and employees, minorities, economically and/or educationally disadvantaged, and individuals who are AIDS/HIV positive. Therefore, when criminologists intend to work with one of these vulnerable populations, they must follow specific guidelines when developing research studies to safeguard potential participants' rights. Additionally, in 1996 the Health Insurance Portability and Accountability Act (HIPAA) was passed by Congress. This act provides additional protection of health care information, including confidentiality and regulations surrounding data collection, storage, and management. More information about IRB and HIPAA can be found at the U.S. Department of Health and Human Services website (http://www.hhs.gov/).

Random Assignment

After participants are selected, they must be assigned to one of the study's groups. If possible, participants should be *randomly assigned* to a group to increase the likelihood that groups are equivalent on all important characteristics. Because random assignment is based on probability, similar to flipping a coin, it does not ensure equivalency. Therefore, before initiating a study, the researcher must verify that the groups are indeed similar on important variables so that any differences found at the end of the study can be attributed to the IV instead of another variable. If the groups are not equal, the researcher may need to reassign participants to groups. The researcher may also implement other techniques (e.g., matched-pairs design, randomized block design; Kazdin, 2003) to ensure equivalency.

To illustrate, consider the depression intervention scenario discussed above. If inmates self-selected or were assigned to a treatment group based upon when they entered the study, there may be meaningful differences between the groups at baseline. For example, one group may have more severe depressive symptoms and a history of recurrent depression. This example represents a threat to validity, called *selection bias*. When this threat is not accounted for, there is doubt as to whether preexisting differences between the groups are actually responsible for between-group differences in depression rates after the intervention, rather than the intervention itself. Although it is impossible to identify or measure all of the variables that may be related to depressive symptoms, researchers must decide a priori which variables are most important to equate between groups. For instance, they may decide that years in prison, gender, medication, or possibility for parole are important, as they are likely to affect depressive symptoms and response to intervention. Previous research can provide information about which variables may be most relevant. By equalizing groups on these variables, the researcher is strengthening the internal validity of the study.

Popular Experimental Designs

Now that the reader is familiar with features of experimental designs, several popular designs will be introduced. This is not an exhaustive list of experimental designs, and interested readers are referred to Hagan (2014) for a more comprehensive review.

No-Treatment Control

One common design in an experimental study is a no-treatment control design. In this type of design, one group receives an active intervention (the experimental group) and another does not (the control group). The two groups are assessed on the DV before the intervention takes place, called *baseline*, and again after the intervention. Importantly, the control group is only contacted twice—once for baseline measurement and once when the treatment group is finished with the intervention. In this manner, participants in the control group are receiving limited amounts of special treatment from researchers, such as attention or special privileges, factors which could increase *reactivity* and function as an intervention itself (Kazdin, 2003).

Take, for instance, a researcher who wants to evaluate the impact of a depression intervention for inmates. Inmates who meet a minimum level of depressive symptoms qualify for the study, and may subsequently be selected from a pool of other qualified inmates and assigned to either the experimental or control group. Both groups then complete a measure of depressive symptoms prior to initiation of the 8-week intervention, and complete the same measure at the end of the 8 weeks. If the two groups are relatively similar on levels of depression and other important characteristics at baseline, and if the intervention group showed significantly reduced rates of depression at the end of the study, the researcher may conclude that the intervention was effective.

This design controls for many types of threats to the internal validity of the study, namely maturation and history. By incorporating a control group into the study, the researcher is able to determine if changes in the experimental group's depressive symptoms are due to the intervention, or if they simply decreased with the passage of time. This threat to validity is called *maturation*. If the experimental group's depressive symptoms decreased while the control group's symptoms persisted, the most probable and parsimonious explanation is that the intervention was effective. Use of a control group also reduces the threat of *history*, or the possibility that an event other than the intervention accounted for changes in the DV.

Like any design, a no-treatment control group has limitations. One main limitation is the ethical issue of withholding treatment from people who may need it. A second limitation is related to the construct validity of the study, which speaks to the causal attribution for the intervention's effectiveness (i.e., why did depressive symptoms decrease in the experimental group?). The experimental group received not only the intervention, but also attention and special treatment that was not afforded to the control group. Therefore, it is possible that the reduction in depressive symptoms in the experimental group is due to the differential attention that this group received.

Last, in a setting such as a jail, the intervention may be diffused to the control group. For example, if an inmate from the experimental group and an inmate from the control group share a cell, the inmate receiving the intervention might share components of the intervention with the cellmate. This diffusion could attenuate actual differences between the groups.

Waitlist Control

A similar, yet distinct type of experimental study is one that utilizes a waitlist control group. In this design, the control group receives the intervention after the experimental group received the intervention. This design addresses the ethical issue of withholding treatment from a vulnerable population, while simultaneously addressing validity threats of history and maturation.

A limitation of this design when compared to the no-treatment control is that a waitlist control group does not allow for the inclusion of follow-up studies. That is, in a no-treatment control design, researchers could feasibly assess differences between groups *X* months after the intervention to evaluate rates of relapse. However, in a waitlist control design, these comparisons are impossible to conduct. Further, differential attention and diffusion of treatment remain possible threats to validity with this design.

Nonspecific Treatment or Attention Placebo

In order to disentangle the effects of attention and contact with participants from the true intervention effect, a nonspecific treatment (also called an "attention placebo") can be incorporated into the study design. In the attention placebo group, participants experience some nonspecific aspects of an intervention that are potentially "active" (i.e., they have the potential to influence the DV), such as attention from a researcher and active listening from a therapist.

While intended to account for issues created by no-treatment or waitlist control groups, the use of an attention placebo has its own unique limitations. First, developing a "nonspecific" treatment is conceptually problematic. Second, it is difficult to create a control group that specifically provides attention without exposing participants to other potentially active components of the intervention. For example, in an attention placebo group with a therapist, the therapist may utilize jargon or therapeutic language that may be considered part of the active intervention. Third, if the attention placebo is expected not to produce an effect, then it may be unethical to administer an ineffective intervention to a vulnerable population. Finally, depending on the setting of the study and the contact of participants across intervention and control groups, diffusion of treatment may be a potential limitation of this design.

Treatment as Usual Group

A researcher may need to know whether a new intervention is more effective than the standard of care that is currently being utilized (referred to as "treatment as usual," or TAU). For example, a new educational prevention program to reduce rates of drunk driving may be compared to the current educational materials that are provided during standard Driver's Education courses. In this scenario, new drivers may be randomly assigned to receive the new educational materials (experimental group) or the standard educational materials (TAU), and their driving records may be tracked for DUIs over the next 12 months.

Treatment as usual (TAU) comparison group studies eliminate the ethical dilemma of withholding treatment inherent in the no-treatment control, waitlist control, and attention placebo designs by providing standard treatment. Unfortunately, the TAU design is unable to account for maturation, or the possibility that the behavior decrease on its own without any intervention. Consequently, this limitation has led some researchers to use multiple groups designs, discussed next.

Multiple Group Design

In a multiple group design, more than two groups are incorporated into the study. The choice of which type of groups to include depends on what the researcher is studying and the question that he/she is trying to answer. Further, in some cases, no standard of care may exist, or withholding intervention through a no-treatment control or waitlist design may be considered unethical or denied by the IRB.

In general, incorporating different types of control groups into a study increases the researcher's ability to rule out threats to validity. However, the benefits of having multiple groups needs to be weighed against the costs; having several different groups may tax limited resources and make it more challenging for the researcher to obtain an appropriate sample size for adequate power.

Threats to Validity

Although threats to validity have been mentioned throughout the chapter, this section reviews these threats and highlights additional threats that may be encountered in research. When designing a study and considering various threats, the researcher should prioritize internal validity. Simply, internal validity is the likelihood or plausibility that the IV accounts for changes in the DV. To protect internal validity, the research study must render alterative explanations of the results implausible through careful study design (Kazdin, 2003). Without internal validity, the results of the study are uninterpretable and meaningless. External validity, or the extent to which results generalize to the population of interest, should be prioritized next, along with construct validity, or the presumed reason for *why* the intervention produced the outcome. Finally, statistical conclusion validity, or the extent to which the statistical analyses used in the study support the conclusions drawn about the intervention and the outcome, should be considered. The following review focuses on threats to internal validity. Further information about validity threats can be found in Kazdin (2003).

History

The internal validity threat of history is plausible when events common to all subjects within the experiment or outside of the experiment may be responsible for the results. Consider the following example: during the course of an intervention intended to reduce drunk driving, several teens from a high school within the community where the study is taking place pass away from an automobile accident related to drinking and driving. This tragic event may easily explain reduced rates of drinking and driving over the next few months. If the research study did not include a control group that was administered during the same time as the intervention, the impact of the intervention would be unclear.

Maturation

Maturation, or the process of changing over time, is often associated with change in a variety of outcomes. For example, psychological symptoms may change as individuals get older (e.g., only about half of children diagnosed with oppositional defiant disorder will meet criteria for the disorder 3 years after diagnosis (Barkley, 1997); symptoms of attention-deficit/hyperactivity disorder typically decline over time (Biederman, Mick, & Faraone, 2000)). This threat, which often co-occurs with history, can be evaluated with the use of control groups.

Selection Biases

Selection biases can seriously threaten validity when differences between groups at baseline are related to the DV, or outcome of interest. Often, selection biases can be prevented through the assignment of subjects to groups, and verifying that the groups are equal on important characteristics prior to intervention.

Attrition

Another serious threat to internal validity is *attrition*, or participant drop out. Attrition becomes problematic when there is differential attribution between groups, which disrupts random assignment. For example, when considering the treatment study for depression described above, the researcher may initially conclude that the intervention demonstrated a significant decrease in depression symptoms and that this decrease was significantly greater than the change in the control group's depression symptoms. However, this finding might be explained by attrition if, for example, the individuals with the most severe levels of depression were not benefitting from the intervention and dropped out of the experimental group. Thus, researchers should attempt to understand the reasons for attrition and should attempt to quantify attrition across groups and, in some cases, attempt to statistically control for the attrition.

Diffusion of Treatment

Diffusion of treatment occurs when participants in the control group receive the intervention. This may be more likely when participants live in close proximity to one another and have frequent contact, such as in a prison. If the intervention is effective, diffusion of treatment attenuates the impact of the treatment, and enhances the impact of the control condition (e.g., TAU, attention placebo). In this case, the effect of the intervention may be found not significantly different from the control group, causing the effective intervention to be deemed "ineffective."

Reaction of Controls

When individuals participate in a control group, they may receive additional attention, support, special privileges, or monetary compensation. Each of these factors can significantly impacting the participant's behavior and mindset, potentially influencing the study's outcome. Additionally, if an individual knows that they are participating in the control group, this knowledge could also influence their behavior in unpredictable ways (e.g., compensatory performance or demoralization; Kazdin, 2003). In fact, if a participant is aware that they are participating in an experiment, even if do not know whether they are in the experimental or control group, this could led to attenuation or exaggeration of the outcome. Therefore, when appropriate researchers should strive to utilize "double-blind" (the researchers delivering the intervention and the participants receiving the intervention are unaware of group status) procedures, these procedures may not fully protect against reactivity.

Conclusion

The purpose of this chapter was to provide the reader with an understanding of the importance of research design and its impact on the type of information gathered through empirical studies. Important issues such as ethics within research, threats to validity, and types of experimental and nonexperimental design were presented. Types of participant selection and recruitment, and data collection were also outlined to help readers appreciate the ways in which study design methodologies can be employed to counter validity threats. The ultimate goal of this chapter was to increase readers' acumen as consumers of research and research scientists.

References

Barkley, R. A. (1997). *Understanding the defiant child program manual*. New York, NY: The Guilford Press.

Beasley, J. O. (2004). Serial murder in America: Case studies of seven offenders. *Behavioral Sciences and the Law*, *22*, 395–414. doi:10.1002/bsl.595

Bell, S. J. (2012). *Youth offenders and youth justice: A century after the fact* (4th ed.). Toronto: Nelson Education.

Biederman, J., Mick, E., & Faraone, S. V. (2000). Age-dependent decline of symptoms of Attention-Deficit/Hyperactivity Disorder: Impact of remission definition and symptom type. *The American Journal of Psychiatry*, *157*, 816–818. doi:10.1176/appi.ajp.157.5.816

Briere, J. (1995). *Trauma Symptom Inventory professional manual*. Odessa, FL: Psychological Assessment Resources.

Briere, J. (2001). *Detailed Assessment of Posttraumatic Stress (DAPS)*. Odessa, FL: Psychological Assessment Resources.

Butcher, J. N., Graham, J. R., Ben-Porath, Y. S., Tellegen, A., Dahlstrom, W. G., & Kaemmer, B. (2001). *MMPI-2: Manual for administration, scoring, and interpretation* (Revised ed.). Minneapolis: University of Minnesota Press.

Center for Disease Control. (2015). *The Tuskegee timeline*. Retrieved from http://www.cdc.gov/tuskegee/timeline.htm

Clayton, R. R., Catterello, A. M., & Johnstone, B. M. (1996). The effectiveness of Drug Abuse Resistance Education (project DARE): 5-Year follow-up results. *Preventative Medicine*, *25*, 307–318.

D.A.R.E. Board of Directors. (2014). *2014 Annual Report*. Retrieved from http://www.dare.org/d-a-r-e-annual-report-for-2014/

Daniel, J. (2012). *Sampling essentials: Practical guidelines for making sampling choices*. Thousand Oaks, CA: SAGE Publications. doi:10.4135/9781452272047.n5

Dishion, T. J., McCord, J., & Poulin, F. (1999). When intervention harm. Peer groups and problem behavior. *American Psychologist*, *54*, 755–764. doi:10.1037/0003-066X.54.9.755

Ennett, S. T., Tobler, N. S., Ringwalt, C. L., & Flewelling, R. L. (1994). How effective is drug abuse resistance education: A meta-analysis of Project DARE outcome evaluations. *American Journal of Public Health*, *84*, 1394–1401. doi:10.2105/AJPH.84.9.1394

Hagan, F. E. (2014). *Research methods in criminal justice and criminology* (9th ed.). Upper Saddle River, NJ: Pearson Education.

Haney, C., Banks, W. C., & Zimbardo, P. G. (1973). A study of prisoners and guards in a simulated prison. *Naval Research Review*, *30*, 4–17.

Hintze, J. M. (2005). Psychometrics of direct observation. *School Psychology Review*, *34*, 507–519.

Kazdin, A. E. (2003). *Research design in clinical psychology* (4th ed.). Needham Heights, MA: Allyn & Bacon.

MacKenzie, D. L., & Souryal, C. (1994). *Multi-site evaluation of shock incarceration: Executive summary*. Washington, DC: US Department of Justice/ NIJ.

Milgram, S. (1963). Behavioral study of obedience. *The Journal of Abnormal and Social Psychology*, *67*, 371–378. doi:10.1037/h0040525

Morey, L. C. (1991). *The Personality Assessment Inventory professional manual*. Odessa, FL: Psychological Assessment Resources.

Morey, L. C. (2007). *Personality Assessment Inventory professional manual* (2nd ed.). Lutz, FL: Psychological Assessment Resources.

Nagurney, J. T., Brown, D. F., Sane, S., Weiner, J. B., Wang, A. C., & Chang, Y. (2005). The accuracy and completeness of data collected by prospective and retrospective methods. *Academic Emergency Medicine*, *12*, 884–895.

Nulty, D. D. (2008). The adequacy of response rates to online and paper surveys: What can be done? *Assessment*, *33*, 301–314. doi:10.1080/02602930701293231

Robinson, S. M., Sobell, L. C., Sobell, M. B., & Leo, G. I. (2014). Reliability of the Timeline Followback for cocaine, cannabis, and cigarette use. *Psychology of Addictive Behaviors*, *28*, 154–162. doi:10.1037/a0030992

Schmalleger, F. (2006). *Criminology today* (4th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Shepard, E. M. (2001). *The economic costs of D.A.R.E.* Institute of Industrial Relations, Research Paper Number 22. Retrieved from www.drugpolicy.org/docUploads/DAREfinalRP.pd

Sobell, L. C., Sobell, M. B., Buchan, G., Cleland, P.A., Fedoroff, I., & Leo, G.I. (1996, November). *The reliability of the Timeline Followback method applied to drug, cigarette, and cannabis use.* Paper presented at the 30th Annual Meeting of the Association for Advancement of Behavior Therapy, New York, NY.

Suen, H. K., & Ary, D. (2014). *Analyzing quantitative behavioral observation data*. New York, NY: Psychology Press.

Tombaugh, T. N. (1996). *Test of Memory Malingering (TOMM)*. New York, NY: Multi-Health Systems.

United States. (1978). *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research*. Bethesda, MD: The Commission.

Wilson, D. B., MacKenzie, D. L., & Mitchell, F. N. (2005). Effects of correctional boot camps on offending. *Campbell Systematic Reviews*, *1*. doi:10.4073/csr.2003.1